FAST OFFLINE EQUILIBRIUM LEARNING IN α-POTENTIAL GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

An α-potential game is a multi-player non-cooperative interaction in which a global potential function tracks individual player incentives up to a structural bias α. While identifying a Nash Equilibrium (NE) in generic general-sum games is known to be computationally intractable, the potential game structure provides a scalar objective whose approximate maximizers yield approximate equilibria. In this paper, we study the offline learning of NE in α-potential games using KL regularization. To analyze this process, we propose a novel *Reference-Anchored* offline data coverage framework that anchors data requirements to a known reference policy rather than an unknown optimum. Building on this, we propose Offline Potential Mirror Descent (OPMD), a decentralized algorithm that achieves an accelerated $\widetilde{\mathcal{O}}(1/n)$ statistical rate for the regularized game, compared with the $\widetilde{\mathcal{O}}(1/\sqrt{n})$ rate typical of unregularized offline multi-agent learning. To our knowledge, this work characterizes the first fast-rate offline learning approach for α-potential games.

1 INTRODUCTION

Finding a Nash equilibrium in general-sum games is PPAD-complete (Daskalakis et al., 2009; Chen et al., 2009), establishing that equilibrium learning is computationally intractable in generic general-sum games. This fundamental barrier poses a significant challenge for multi-agent reinforcement learning (MARL), where the goal is to derive equilibrium policies in high-dimensional environments. Fortunately, recent theoretical advancements show that the α-potential game framework provides a general form for dynamic N-player non-cooperative games (Guo et al., 2025; Li, 2025; Guo et al., 2026). This structure provides a sufficient route to approximate equilibrium: any ϵ-approximate maximizer of the global potential function is an $(\epsilon + \alpha)$-Nash equilibrium (Guo et al., 2026). This connection provides a unified optimization framework for analyzing equilibria in general-sum games.

The α-potential framework includes exact potential games and perturbations whose unilateral incentives remain approximately aligned with a common potential (Li, 2025; Guo et al., 2026). This framework encompasses practically significant scenarios such as traffic congestion, autonomous driving, smart grid management, and perturbed team games (Guo et al., 2026). This structural alignment provides a route to approximate Nash equilibria through a single-objective optimization task (Guo et al., 2025; Li, 2025).

In the offline setting, the difficulty of learning stable strategies is compounded by distribution shift, where the learned policies explore regions of the state-action space poorly represented in the historical logs (Levine et al., 2020). To establish sample-efficient recovery, standard methodologies typically rely on the principle of explicit pessimism (Jin et al., 2021; Zhang et al., 2023b). These approaches utilize manually designed lower confidence bounds (LCB) or uncertainty bonuses to penalize values in regions with low data support (Cui & Du, 2022b; Zhong et al., 2022). While theoretically robust, these pessimistic mechanisms suffer from two primary drawbacks. First, they often require the construction of complex uncertainty quantifiers over joint action spaces, leading to significant computational overhead. Second, existing coverage conditions for these methods are typically anchored to an unknown optimal equilibrium, introducing a conceptual circularity—known as the "Curse of the Unknown Optimum"—that makes the assumptions difficult to verify in practice.

In this work, we establish that KL regularization relative to a fixed and known reference policy provides a standalone, pessimism-free mechanism for equilibrium learning in α-potential contextual

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.

bandits. To this end, we propose the *Reference-Anchored* coverage framework, a novel formulation of unilateral concentrability that resolves the conceptual circularity of prior work by anchoring data requirements to a reference policy rather than an unknown optimum. We prove that this regularized framework not only stabilizes learning but also leverages the geometry of KL regularization and the potential structure to achieve fast statistical rates. Specifically, we show that decentralized regularized learning can achieve fast $\tilde{\mathcal{O}}(1/n)$ statistical rates for the regularized game, compared with the $\tilde{\mathcal{O}}(1/\sqrt{n})$ rates typical of unregularized offline multi-agent learning (Cui & Du, 2022b; Zhang et al., 2023b).

A key identity in our analysis is that the KL-regularized Mirror Descent update forms an exact log-linear interpolation between the current iterate and the empirical regularized best response. This structural identity allows us to recast the evaluation mismatch between best responses and the learning iterates in terms of optimization progress, enabling acceleration beyond the standard $\tilde{\mathcal{O}}(1/\sqrt{n})$ statistical rate (Cui & Du, 2022b; Zhong et al., 2022). Consequently, we establish cancellation of first-order estimation-error terms at empirical regularized equilibria and control the evaluation mismatch during decentralized learning, enabling the accelerated $\tilde{\mathcal{O}}(1/n)$ statistical rate. Our key contributions are summarized as follows:

- **Reference-Anchored Coverage**: We introduce a novel unilateral concentrability condition that anchors data requirements to a known reference policy rather than an unknown optimum. This formulation provides a tractable framework for offline MARL, resolving the conceptual circularity of prior work.
- **Theoretical Fast-Rate Framework**: We propose the Regularized Offline Potential Equilibrium (ROPE) algorithm to establish that a fast $\tilde{\mathcal{O}}(1/n)$ statistical rate is achievable in regularized α-potential games. Our analysis demonstrates that the geometry of KL regularization enables the cancellation of first-order errors, going beyond the $\tilde{\mathcal{O}}(1/\sqrt{n})$ statistical rate typical of unregularized general-sum games.
- **Decentralized Algorithmic Recovery**: We develop Offline Potential Mirror Descent (OPMD), a decentralized algorithm that implements the regularized equilibrium learning process through iterative self-play. We prove that OPMD preserves the $\tilde{\mathcal{O}}(1/n)$ rate (up to the structural bias α) by exploiting the log-linear geometry of the Mirror Descent updates to resolve linear evaluation mismatch as a vanishing optimization error.

2 Related Work

Potential Games. The study of dynamic potential games generalizes the static framework to state-dependent environments, providing existence and convergence results for Nash Equilibria under the assumptions of the respective Markov potential-game models (Leonardos et al., 2021; Yan & Liu, 2025). To accommodate the diverse structures of real-world strategic interactions, recent theoretical advancements have introduced the α-potential framework, which characterizes systems where unilateral deviations track potential changes up to a structural bias $\alpha \geq 0$ (Guo et al., 2025; Li, 2025). Within this landscape, Guo et al. (2025) established that the α-potential structure is universal, proving that bounded finite games can be characterized as an α-potential game. The resulting equilibrium guarantee is informative when α is small. This characterization has been extended to stochastic differential games and constrained settings, with explicit characterizations of α in terms of player heterogeneity and interaction intensity (Das & Eksin, 2024; Guo et al., 2025). Furthermore, Yan & Liu (2025) identified sufficient conditions on reward design and transition dynamics that allow for the systematic construction of potential games, with specific applications to autonomous driving and traffic management.

While the α-potential framework connects potential maximization to approximate equilibrium, existing literature primarily focuses on the *planning* regime—where the optimizer has access to the potential landscape or best-response oracles (Das & Eksin, 2024; Guo et al., 2026)—or the *online learning* regime. In the latter, Ding et al. (2022) proposed independent policy gradient (IPG) algorithms for large-scale games, establishing an exact-gradient iteration bound with no explicit state-space-size dependence and a sample-based guarantee with function approximation. Similarly, Li (2025) provided a comprehensive analysis of RL algorithms in α-potential games across both discrete and continuous time. However, these results are derived for the online setting, where agents benefit

from active exploration and immediate environmental feedback. Despite the computational tractability afforded by the potential structure under the assumptions of these works, the statistical properties of equilibrium learning in the *offline* regime—specifically under general function approximation and distribution shift—remain largely unexplored. Our work bridges this gap by demonstrating that the structural alignment of α-potential games can be exploited to achieve accelerated statistical rates in the offline setting without the need for explicit pessimistic mechanisms.

General-Sum Games. The transition from single-agent reinforcement learning to multi-agent environments necessitates a shift from identifying value-optimal policies to recovering stable equilibria (Roughgarden, 2016; Cui & Du, 2022a). While single-agent objectives seek to maximize a unified return, multi-agent dynamics are characterized by individual strategic incentives, typically modeled through Coarse Correlated Equilibria (CCE) or Nash Equilibria (NE) (Daskalakis et al., 2009; Cui & Du, 2022b). Recovering a Nash Equilibrium is fundamentally challenging due to its PPAD-complete complexity in general-sum settings, which has historically led many methodologies to focus on the more computationally tractable CCE, where stability is established via no-regret learning or correlated strategies (Chen et al., 2009; Yan et al., 2024). However, the pursuit of Nash stability remains essential for decentralized applications requiring robust independent decision-making, where coordinated or correlated mechanisms are infeasible or strategically undesirable (Song et al., 2021; Cui & Du, 2022b).

In the offline regime, the difficulty of equilibrium learning is compounded by the lack of interactive exploration and the presence of severe distribution shift (Levine et al., 2020; Jin et al., 2021). Standard methodologies typically address these gaps by adapting the principle of pessimism to the game-theoretic setting, utilizing uncertainty bonuses or lower confidence bounds (LCB) to penalize values in poorly explored joint-action regions (Zhong et al., 2022; Cui & Du, 2022b; Zhang et al., 2023b). Despite their theoretical robustness, these pessimistic frameworks suffer from two primary bottlenecks. First, many existing approaches rely on complex uncertainty quantifiers over joint action spaces, potentially adding computational overhead (Cui & Du, 2022b; Yan et al., 2024). Second, existing algorithms that target Nash Equilibria often require centralized equilibrium computation or iterative value updates (Zhong et al., 2022; Zhang et al., 2023b). Our work addresses these challenges by demonstrating that regularized decentralized learning can successfully reduce regularized Nash exploitability at fast statistical rates without the need for explicit pessimistic bonuses, exploiting the potential structure to support decentralized equilibrium learning.

3 Preliminaries

We consider an m-player contextual bandit game possessing an α-potential structure. In this section, we formalize the environment, the regularized objectives, and the potential-based equilibrium concepts analyzed in this work.

3.1 m-Player Contextual Bandits

The game is defined by the tuple $\mathcal{M} = (\mathcal{N}, \mathcal{X}, \{\mathcal{A}_i\}_{i=1}^m, \{r_i\}_{i=1}^m, \rho)$, where $\mathcal{N} = [m]$ is 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. Players simultaneously select actions $a_i \in \mathcal{A}_i$, forming a joint action profile $\boldsymbol{a} = (a_1, \dots, a_m) \in \mathcal{A} := \prod_{i=1}^m \mathcal{A}_i$. Each player i receives a reward $r_i(x, \boldsymbol{a}) \in [0, 1]$. A joint policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ maps contexts to distributions over joint actions. In this work, we primarily focus on product policies, where $\pi(\boldsymbol{a} \mid x) = \prod_{i=1}^m \pi_i(a_i \mid x)$.

Here $\Delta(\mathcal{A}_i)$ denotes the probability simplex on $\mathcal{A}_i$, $[m] = \{1, \dots, m\}$, and $\pi_{-i} = \prod_{j \neq i} \pi_j$ is the opponents' product policy. We take $\mathcal{X}$ to be standard Borel, all action sets to be finite and nonempty, and all reward functions and policy kernels to be measurable.

3.2 KL-Regularized Objectives

KL regularization relative to a fixed reference policy has emerged as a fundamental mechanism for enforcing behavioral constraints and incorporating prior knowledge in multi-agent systems. In many practical scenarios, such as the alignment of large language models through Reinforcement Learning from Human Feedback, KL regularization is used to penalize policy drift and encourage proximity to

a trusted reference distribution (Munos et al., 2024). Here, the KL penalty is part of each player's objective, and our goal is to learn a Nash equilibrium of the resulting regularized game.

For a fixed reference joint policy $\pi^{\text{ref}} = (\pi_1^{\text{ref}}, \dots, \pi_m^{\text{ref}})$ and a regularization coefficient $\eta > 0$, the regularized value function V_i^π for player i is defined as:

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

We denote the global regularized return as $J_i(\pi) := \mathbb{E}_{x\sim\rho}[V_i^\pi(x)]$. Each player i seeks to maximize their own regularized return given the strategies of their opponents π_{-i}. The corresponding regularized best-response value is defined as $V_i^{\dagger,\pi_{-i}}(x) := \max_{\pi_i'} V_i^{\pi_i',\pi_{-i}}(x)$. The reference policy is fixed independently of the offline observations. Π_i denotes the measurable policies with finite expected KL divergence from π_i^{ref}, and $\Pi = \prod_i \Pi_i$.

3.3 The α-Potential Structure

We leverage the framework of α-potential games to characterize the strategic interactions (Li, 2025; Guo et al., 2026).

Assumption 3.1 (True policy-level approximate potential)**.** There exists $\Phi : \Pi \to \mathbb{R}$ such that for any player i, any joint policy $\pi \in \Pi$, and any unilateral deviation $\pi_i' \in \Pi_i$,

$$|(J_i(\pi_i', \pi_{-i}) - J_i(\pi_i, \pi_{-i})) - (\Phi(\pi_i', \pi_{-i}) - \Phi(\pi_i, \pi_{-i}))| \le \alpha. \tag{1}$$

In this work, we exploit this structure to use potential maximization as a sufficient route to approximate equilibria.

3.4 Approximate Nash Equilibrium and Total Exploitability

Our goal is to recover a joint policy $\pi^* \in \Pi$ that constitutes an ϵ-Nash Equilibrium (ϵ-NE). Conceptually, an ϵ-NE represents a state of near-stability where no player can unilaterally improve their regularized return by more than the tolerance ϵ:

$$J_i(\pi_i^*, \pi_{-i}^*) \ge \sup_{\pi_i' \in \Pi_i} J_i(\pi_i', \pi_{-i}^*) - \epsilon, \qquad \forall i \in [m].$$

Here $\epsilon \ge 0$ is the equilibrium tolerance, while α denotes the structural bias.

For a learned policy $\hat{\pi}$, we evaluate its performance using the NE Total Exploitability, or Nash Gap, which aggregates individual suboptimality to provide a scalar measure of the policy's distance from equilibrium:

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

In this formulation, $V_i^{\dagger,\hat{\pi}_{-i}}(x) - V_i^{\hat{\pi}}(x)$ represents the maximum unilateral improvement available to player i given the fixed strategies of their opponents.

The α-potential structure facilitates a reduction of this equilibrium search to a single-objective optimization problem (Guo et al., 2026). Formally, for some optimization error $\epsilon \ge 0$, let $\hat{\pi}$ denote an ϵ-approximate maximizer of the global potential function Φ, satisfying $\Phi(\hat{\pi}) \ge \sup_{\pi\in\Pi} \Phi(\pi) - \epsilon$.

A core result in the characterization of α-potential games is that any such policy $\hat{\pi}$ is guaranteed to be an $(\epsilon + \alpha)$-Nash Equilibrium of the underlying regularized game. This structural alignment establishes potential optimization as a sufficient route for equilibrium learning.

3.5 Offline Learning Model

The learner has access to an offline dataset $\mathcal{D} = \{(x_\tau, \boldsymbol{a}_\tau, \mathbf{r}_\tau)\}_{\tau=1}^n$ generated by an unknown behavioral distribution $\mu(x, \boldsymbol{a})$ with $\mathbf{r}_\tau = (r_{\tau,1}, \dots, r_{\tau,m})$. We assume access to function classes $\{\mathcal{Q}_i\}_{i=1}^m$ for reward estimation, where each class $\mathcal{Q}_i$ consists of functions mapping $\mathcal{X} \times \mathcal{A} \to [0, 1]$. Following standard literature (Xie & Jiang, 2021; Zhang et al., 2023b), we assume realizability, i.e., $r_i \in \mathcal{Q}_i$. For any estimate $\hat{Q}_i \in \mathcal{Q}_i$, the pointwise regression error is defined as $\mathcal{Z}_i(x, \boldsymbol{a}) := \hat{Q}_i(x, \boldsymbol{a}) - r_i(x, \boldsymbol{a})$.

The rows $(x_\tau, \boldsymbol{a}_\tau, \mathbf{r}_\tau)$ are iid, $(x_\tau, \boldsymbol{a}_\tau) \sim \mu$, $r_{\tau,i} \in [0,1]$, and $\mathbb{E}[r_{\tau,i} \mid x_\tau, \boldsymbol{a}_\tau] = r_i(x_\tau, \boldsymbol{a}_\tau)$. Each class $\mathcal{Q}_i$ is finite and consists of measurable functions. The estimator $\widehat{Q}_i$ minimizes empirical squared loss over $\mathcal{Q}_i$, using fixed measurable tie breaking. The function classes are fixed independently of the offline dataset. For a finite vector v, $\|v\|_\infty = \max_a |v(a)|$; for density ratios, $\|\cdot\|_\infty$ denotes the μ-essential supremum.

4 Data Coverage and Concentrability

A fundamental challenge in offline reinforcement learning is the distribution shift between the logged dataset $\mathcal{D}$ and the target policy (Levine et al., 2020; Cui & Du, 2022b; Zhang et al., 2023b). To establish sample-efficient recovery, one must assume the offline data provides sufficient coverage of the state-action regions relevant to the target objective. In this section, we review the landscape of concentrability assumptions in multi-agent systems and formalize our novel reference-anchored approach.

4.1 From Single-Agent to Joint Concentrability

In single-agent offline RL, sample efficiency is typically established under the assumption that the data distribution μ covers the optimal policy π^* (Levine et al., 2020).

Assumption 4.1 (Single-Agent Concentrability)**.** There exists a constant $C^* \geq 1$ such that

$$\left\| \frac{\rho(x)\pi^*(a \mid x)}{\mu(x,a)} \right\|_\infty \leq C^*.$$

However, in multi-agent games, simply covering a single optimal profile is insufficient for learning a stable equilibrium (Cui & Du, 2022a). The most straightforward extension is **joint concentrability**, which demands that the dataset support all possible simultaneous deviations by the agents.

Assumption 4.2 (Joint Multi-Agent Concentrability)**.** There exists a constant $C_{\text{joint}} \geq 1$ such that for any joint product policy $\pi = \prod_{i=1}^m \pi_i$:

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

Here, the vector $\boldsymbol{a}$ denotes the joint action profile of all m players. Under the joint concentrability condition above, the required coverage coefficient is at least the *product* of action spaces ($\prod_{i=1}^m |\mathcal{A}_i|$), resulting in an *exponential* growth in complexity known as the "curse of multiagents" (Cui & Du, 2022b). This scaling makes the assumption practically prohibitive for environments with many participants or a large action space. In contrast, Cui & Du (2022b) demonstrate that a strategy-wise approach can achieve linear scaling ($\sum_{i=1}^m |\mathcal{A}_i|$), effectively breaking this curse by reducing the complexity to the sum of individual action spaces.

4.2 Unilateral Concentrability and the Circularity Problem

To bypass the exponential scaling of joint action spaces, recent works have adopted **unilateral concentrability** (Cui & Du, 2022b;a; Zhang et al., 2023b). This principle requires that the dataset only needs to support scenarios where a single agent deviates from a target equilibrium strategy π^* while all others remain stationary.

Assumption 4.3 (Equilibrium-Anchored 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 $\pi' \in \Pi_{\text{uni}}(\pi^*)$:

$$\left\| \frac{\rho(x)\pi'(\boldsymbol{a} \mid x)}{\mu(x,\boldsymbol{a})} \right\|_\infty \leq C_{\text{uni}}^*. \tag{2}$$

In two-player zero-sum Markov games, Cui & Du (2022a) demonstrated that unilateral concentration is a necessary assumption for learning a Nash Equilibrium (NE) from offline data. Crucially, Cui & Du (2022b) demonstrate that this strategy-wise approach allows the sample complexity to scale *linearly* with the *sum* of action spaces ($\sum_{i=1}^m |\mathcal{A}_i|$), effectively breaking the curse of multiagents.

The Curse of the Unknown Optimum. Despite its improved scaling, classical unilateral concentrability is anchored to an unknown optimal equilibrium policy π^* (Cui & Du, 2022a; Zhang et al., 2023b). This introduces a fundamental conceptual circularity: the data sufficiency required to find an equilibrium is expressed in terms of that equilibrium itself. In general-sum settings, where equilibria are potentially non-unique and finding one is PPAD-complete (Daskalakis et al., 2009), this ambiguity makes the assumption difficult to verify before learning.

4.3 Proposed: Reference-Anchored Unilateral Concentrability

By utilizing KL regularization, we solve the circularity problem by anchoring the coverage requirement to a *fixed and known* reference policy π^{ref}. This shifts the theoretical burden from an elusive optimal strategy to a concrete anchor.

Assumption 4.4 (Reference-Anchored Unilateral Concentrability)**.** Let $\Pi = \prod_{i=1}^{m} \Pi_i$ be the class of joint policies, and let $\pi^{\text{ref}} = (\pi_1^{\text{ref}}, \ldots, \pi_m^{\text{ref}})$ be the fixed reference policy. We define the set of *reference-anchored unilateral deviation policies* $\Pi_{\text{ref}-\text{uni}} \subset \Pi$ as the union of policies where exactly one player deviates to an arbitrary policy while all other players adhere to the reference policy:

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

Let μ denote the underlying joint context-action distribution of the offline dataset $\mathcal{D}$. We assume there exists a finite constant $C_{\text{uni}} \geq 1$ such that for any $\pi' \in \Pi_{\text{ref}-\text{uni}}$, the density ratio is uniformly bounded:

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

Advantages of Reference-Anchored Coverage. Assumption 4.4 offers two critical advantages over prior formulations:

1. **Verifiability**: Unlike π^*, the reference policy π^{ref} is known to the learner. It can be chosen as the behavior policy that generated the data or a previously validated iteration, allowing the coverage coefficient C_{uni} to be defined without knowing π^*.
2. **Reduced Coverage Burden**: Earlier model-based analyses sample every joint state–action pair using a generative model (Zhang et al., 2023a; Subramanian et al., 2023). By focusing on unilateral deviations from the reference, we obtain $C_{\text{uni}} = \max_i |\mathcal{A}_i|$ under uniform reference logging, compared with $C_{\text{joint}} = \prod_i |\mathcal{A}_i|$ for joint concentrability.

We further denote $C_{\text{shift}} \coloneqq \exp(\eta(m-1))$ to quantify the density ratio shift between a learned policy and the reference policy across $m-1$ opponents. Combined with C_{uni}, it characterizes the total statistical overhead for reference-anchored offline equilibrium learning.

5 Regularized Offline Potential Equilibrium

In this section, we establish the Regularized Offline Potential Equilibrium (ROPE) framework, which approximates a Nash equilibrium in α-potential contextual bandits by solving a regularized empirical game. By incorporating KL regularization relative to a known reference policy, ROPE provides a mechanism for stable equilibrium recovery from offline data without relying on the manually tuned, explicit pessimistic bonuses often required in standard methodologies (Jin et al., 2021; Cui & Du, 2022b).

The connection between potential maximization and approximate equilibrium is given in Proposition B.1.

5.1 The ROPE Framework

The ROPE algorithm (Algorithm 1) operates through a two-stage process: statistical estimation followed by strategic computation. First, the learner utilizes the offline dataset $\mathcal{D}$ to construct

empirical estimates of the reward functions for each player via least-squares regression. These estimates serve as the foundation for the empirical game. Second, the algorithm computes a joint policy that constitutes an exact regularized Nash Equilibrium for this estimated environment.

Algorithm 1 Regularized Offline Potential Equilibrium (ROPE)

Input: Offline dataset $\mathcal{D}$, reference policies π^{ref}, regularization parameter $\eta > 0$, function classes $\{\mathcal{Q}_i\}_{i=1}^m$
for player $i = 1, \dots, m$ **do**
  Empirical Estimation: 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
Equilibrium Computation: For each context $x \in \mathcal{X}$, compute a joint product policy $\hat{\pi}(\cdot|x) = \prod_{i=1}^m \hat{\pi}_i(\cdot|x)$ satisfying the exact regularized best-response condition for all $i \in [m]$:
  $\hat{\pi}_i(\cdot|x) \in \operatorname{argmax}_{\pi_i \in \Delta(\mathcal{A}_i)} \left\{ \mathbb{E}_{a_i \sim \pi_i, \boldsymbol{a}_{-i} \sim \hat{\pi}_{-i}}[\hat{Q}_i(x, \boldsymbol{a})] - \eta^{-1} \mathrm{KL}(\pi_i \| \pi_i^{\text{ref}}) \right\}$
for player $i = 1, \dots, m$ **do**
  Value Update: Update the empirical value function for Player i:
    $\hat{V}_i(x) \leftarrow \mathbb{E}_{\boldsymbol{a} \sim \hat{\pi}}[\hat{Q}_i(x, \boldsymbol{a})] - \eta^{-1} \mathrm{KL}(\hat{\pi}_i \| \pi_i^{\text{ref}})$
end for
Output: Learned joint policy $\hat{\pi}$

5.2 Statistical Guarantees

By leveraging the Reference-Anchored Unilateral Concentrability framework (Assumption 4.4), we establish that the equilibrium identified by ROPE achieves a fast statistical rate. The following theorem demonstrates that anchoring to a known reference policy allows the learner to attain a $\widetilde{\mathcal{O}}(1/n)$ statistical rate for the regularized game.

Theorem 5.1 (Statistical transfer for an empirical equilibrium)**.** *Under the offline learning model in Section 3.5 and Assumption 4.4, ROPE satisfies, with probability at least $1 - \delta$ over the offline data,*

$$\mathrm{Gap}_{\mathrm{NE}}(\widehat{\pi}) \leq \widetilde{\mathcal{O}}\left(\frac{m\eta C_{\mathrm{shift}} C_{\mathrm{uni}} \log(2m|\mathcal{Q}|/\delta)}{n}\right).$$

where $|\mathcal{Q}| := \max_i |\mathcal{Q}_i|$.

The proof of Theorem 5.1 is provided in Appendix B.3.5. The accelerated $\widetilde{\mathcal{O}}(1/n)$ statistical rate is a consequence of the structural alignment between the learned equilibrium $\hat{\pi}$ and the empirical regularized best responses, yielding cancellation of first-order estimation errors.

6 Offline Potential Mirror Descent

While the ROPE framework (Algorithm 1) establishes that solving the empirical game yields a fast statistical rate, it assumes access to an exact equilibrium oracle. In high-dimensional environments, explicitly constructing and optimizing the global potential Φ is often computationally prohibitive. To bridge this gap, we leverage the fact that in α-potential games, individual improvements track changes in the global potential function Φ (Guo et al., 2025; 2026). The proposed Algorithm 2 (OPMD) exploits this structure; it allows players to maximize their local utilities independently, serving as a practical, decentralized approach for learning an approximate equilibrium.

6.1 The OPMD Algorithm

The OPMD algorithm utilizes decentralized self-play, with all players updating simultaneously from the current joint policy.

Because the game possesses an α-potential structure, unilateral changes in individual returns track finite differences of the global potential function Φ. Consequently, when players independently update their strategies via Mirror Descent, the joint policy trajectory effectively performs decentralized ascent

on the potential landscape up to estimation and structural errors. OPMD achieves the accelerated $\widetilde{\mathcal{O}}(1/n)$ statistical rate characterized in Theorem 6.1.

Algorithm 2 Offline Potential Mirror Descent (OPMD)

Input: Offline dataset $\mathcal{D}$, reference policies π^{ref}, regularization $\eta > 0$, function classes $\{\mathcal{Q}_i\}_{i=1}^m$, step size $0 < \gamma \le 1/\max\{1, m-1\}$, iterations $T \ge 1$
for player $i = 1, \dots, m$ **do**
  Empirical Estimation: Estimate Q-functions via least-squares:
    $\hat{Q}_i \leftarrow \mathrm{argmin}_{f \in \mathcal{Q}_i} \sum_{\tau=1}^n \left(f(x_\tau, \boldsymbol{a}_\tau) - r_{\tau,i}\right)^2$
  Initialize base policy: $\pi_i^{(1)}(\cdot|x) = \pi_i^{\mathrm{ref}}(\cdot|x)$ for all $x \in \mathcal{X}$
end for
Decentralized Self-Play:
for $t = 1, \dots, T$ **do**
  for player $i = 1, \dots, m$ **do**
    Marginalized Value: Evaluate expected Q-value against 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: Perform local KL-regularized policy improvement:
      $\pi_i^{(t+1)}(a_i|x) \propto \left(\pi_i^{\mathrm{ref}}(a_i|x)\right)^{\frac{\gamma}{\eta+\gamma}} \left(\pi_i^{(t)}(a_i|x)\right)^{\frac{\eta}{\eta+\gamma}} \exp\left(\frac{\eta\gamma}{\eta+\gamma} \bar{Q}_i^{(t)}(x, a_i)\right)$
  end for
end for
Output Selection: Sample one index shared by all players, $t^* \sim \mathrm{Uniform}(1, \dots, T)$
Output: The product joint policy $\tilde{\pi} = \pi^{(t^*)}$

6.2 Statistical Guarantees

By exploiting the inherent geometry of the Mirror Descent iterates, OPMD (Algorithm 2) achieves the fast statistical rate established for the ROPE framework, subject to the structural bias of the game. For the analysis, write $\beta := \gamma/(\eta + \gamma)$.

Theorem 6.1 (Simultaneous OPMD under a true approximate potential)**.** *Under the offline learning model in Section 3.5 and Assumptions 3.1 and 4.4, the update above satisfies, on the same dataset event of probability at least* $1 - \delta$*,*

$$\mathbb{E}_{t^*}[\mathrm{Gap}_{\mathrm{NE}}(\tilde{\pi})] \le \underbrace{\widetilde{\mathcal{O}}\left(\frac{m^2}{\beta T}\right)}_{\textit{optimization}} + \underbrace{\widetilde{\mathcal{O}}\left(\frac{m^2\alpha}{\beta}\right)}_{\textit{bias}} + \underbrace{\widetilde{\mathcal{O}}\left(\frac{m^2 C_{\mathrm{shift}} C_{\mathrm{uni}} \log(2m|\mathcal{Q}|/\delta)}{n}\right)}_{\textit{statistical}}.$$

This theorem establishes the $\widetilde{\mathcal{O}}(1/n)$ scaling of the statistical error. The full proof of Theorem 6.1 is provided in Appendix B.3.6.

7 Numerical Results

We conduct numerical experiments in three-player, two-context binary-action games. We use coordination and congestion families to study statistical scaling. The coordination family rewards strategic alignment among players, whereas the congestion family captures competition for resources whose rewards decrease as more players choose the same resource. For a joint policy, the population regularized Nash gap is the sum, over players, of the largest increase in each player's KL-regularized return from changing only its own policy while the other players' policies remain fixed. Lower is better, and zero corresponds to a Nash equilibrium of the KL-regularized game. The statistical experiment contains 1248 independent offline datasets. Appendix C gives the games, uncertainty calculation, numerical certification, and reproduction steps.

Offline statistical scaling. We compare three-player near-potential coordination and congestion games. For each family, we vary the offline sample size over 13 approximately log-spaced values from $n = 128$ to 8192, using 32 independent datasets at each size while keeping the game, reference

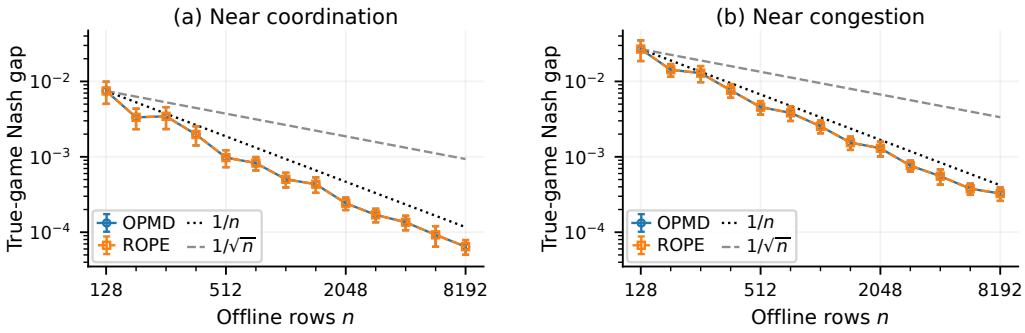


Figure 1: Three-player near-potential coordination and congestion games. Mean population regularized Nash gap over 32 datasets per sample size, with 95% confidence intervals. Dotted $1/n$ and dashed $1/\sqrt{n}$ reference lines both pass through the OPMD mean at $n = 128$. OPMD and ROPE nearly overlap. Numerical values are provided with the code.

and logging policies, regularization, and optimization horizon fixed. The horizon is chosen before observing the offline data. We report the time-averaged OPMD Nash gap.

The population regularized Nash gap decreases from 7.47×10^{-3} to 6.45×10^{-5} for coordination and from 2.68×10^{-2} to 3.26×10^{-4} for congestion. Congestion has larger observed gaps, while both families exhibit approximately $1/n$ decay, with fitted log–log slopes -1.113 and -1.071, respectively. These slopes summarize the observed scaling over the tested sample sizes. OPMD and ROPE nearly overlap across all sample sizes (Figure 1), demonstrating comparable statistical accuracy in these experiments. Together, these numerical results show that both algorithms achieve approximately $1/n$ statistical scaling in the tested coordination and congestion games, supporting our fast-rate analysis.

Additional ablations in the coordination family examine optimization with known rewards (Appendix C.5) and robustness to larger potential deviations (Appendix C.6).

8 Conclusion

This paper established that α-potential games allow for accelerated statistical efficiency in offline equilibrium learning through KL regularization. We introduced the ROPE framework (Algorithm 1) and the decentralized OPMD algorithm (Algorithm 2), demonstrating that regularized Mirror Descent achieves a fast $\widetilde{\mathcal{O}}(1/n)$ statistical rate for approximate Nash equilibria (Theorem 5.1 and Theorem 6.1). Our proposed algorithms bypass the standard $\widetilde{\mathcal{O}}(1/\sqrt{n})$ statistical bottleneck characteristic of generic offline learning without requiring explicit pessimistic bonuses.

Future research may extend these fast-rate guarantees to multi-step environments and explore the adversarial robustness of recovered equilibria under perturbations to the transition dynamics. Additionally, investigating alternative divergences beyond the KL penalty may reveal broader classes of geometric identities for resolving distribution shift in complex strategic interactions. By establishing a pessimism-free foundation for fast-rate recovery, our work provides a statistically efficient path toward large-scale strategic decision-making in complex multi-agent environments.

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

Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Rémi Munos, Nicolas Heess, and Martin A. Riedmiller. Maximum a posteriori policy optimisation. In *Proceedings of the International Conference on Learning Representations*, 2018. 13

Shicong Cen, Fan Chen, and Yuejie Chi. Independent natural policy gradient methods for potential games: Finite-time global convergence with entropy regularization. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 2833–2838. IEEE, 2022. doi: 10.1109/CDC51059.2022.9993175. URL https://arxiv.org/abs/2204.05466v2. Extended proof version: arXiv:2204.05466v2. 18

Claire Chen and Yuheng Zhang. Pessimism-free offline learning in general-sum games via KL regularization. *arXiv preprint arXiv:2605.00264*, 2026. 15

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. 13

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. doi: 10.1145/1516512.1516516. Article 14. 1, 3, 13

Qiwen Cui and Simon S Du. When are offline two-player zero-sum markov games solvable? *Advances in Neural Information Processing Systems*, 35:25779–25791, 2022a. 3, 5, 6, 13, 30

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. 1, 2, 3, 5, 6, 13, 30

Soham Das and Ceyhun Eksin. Learning nash in constrained markov games with an α-potential. *IEEE Control Systems Letters*, 8:808–813, 2024. 2

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. 1, 3, 6, 13

Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo Jovanovic. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*, pp. 5166–5220. PMLR, 2022. 2

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the International Conference on Machine Learning*, 2019. 13

Xin Guo, Xinyu Li, and Yufei Zhang. An α-potential game framework for N-player dynamic games. *SIAM Journal on Control and Optimization*, 63(4):2964–3005, 2025. doi: 10.1137/24M1707316. 1, 2, 7

Xin Guo, Xinyu Li, Chinmay Maheshwari, Shankar Sastry, and Manxi Wu. Markov α-potential games. *IEEE Transactions on Automatic Control*, 71(1):275–290, 2026. doi: 10.1109/TAC.2025.3589416. 1, 2, 4, 7

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, 3, 6, 13

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy Q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, 2019. 13

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. *Reinforcement Learning: State-Of-The-Art*, 2012. 13

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. *arXiv preprint arXiv:2106.01969*, 2021. 2

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *ArXiv Preprint*, 2020. 1, 3, 5, 13

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

Xinyu Li. *An α-Potential Game Framework for Non-Cooperative Dynamic Games: Theory and Algorithms*. University of California, Berkeley, 2025. 1, 2, 4

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

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. 4, 13

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. 13

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. 13

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. 13

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. 13

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. 3

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *ArXiv Preprint*, 2017. 13

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. 13

Ziang Song, Song Mei, and Yu Bai. When can we learn general-sum markov games with a large number of players sample-efficiently? *arXiv preprint arXiv:2110.04184*, 2021. 3

Jayakumar Subramanian, Amit Sinha, and Aditya Mahajan. Robustness and sample complexity of model-based marl for general-sum markov games. *Dynamic Games and Applications*, 13(1):56–88, 2023. 6

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. 13

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. 13

Huiwen Yan and Mushuang Liu. Markov potential game construction and multi-agent reinforcement learning with applications to autonomous driving. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pp. 6838–6844. IEEE, 2025. 2

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. 3, 13

Chenlu Ye, Wei Xiong, Yuheng Zhang, Nan Jiang, and Tong Zhang. Online iterative reinforcement learning from human feedback with general preference model, 2024. 13

Kaiqing Zhang, Sham M Kakade, Tamer Basar, and Lin F Yang. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *Journal of Machine Learning Research*, 24(175):1–53, 2023a. 6

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, 2023b. 1, 2, 3, 4, 5, 6, 13

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*, 2025. 13

Yuheng Zhang, Claire Chen, and Nan Jiang. Beyond pessimism: Offline learning in KL-regularized games. *arXiv preprint arXiv:2604.06738*, 2026. 13

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. 13

Han Zhong, Wei Xiong, Jiyuan Tan, Liwei Wang, Tong Zhang, Zhaoran Wang, and Zhuoran Yang. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. In *International Conference on Machine Learning*, pp. 27117–27142. PMLR, 2022. 1, 2, 3, 13

A Additional Related Work

Foundations of Offline Reinforcement Learning. Offline reinforcement learning (RL) addresses the challenge of deriving optimal policies from static datasets without environment interaction (Lange et al., 2012; Levine et al., 2020). The central difficulty in this regime is the presence of distribution shift, where the target policy's visitation frequency diverges from the logged behavior, which can make value estimates unreliable outside the support of the logged data (Fujimoto et al., 2019; Kumar et al., 2019; Liu et al., 2024; Liu & Zhang, 2024; Chen et al., 2025; Liu et al., 2025). To establish sample-efficient guarantees, the literature has converged on the principle of *pessimism*, which penalizes value estimates in state-action regions lacking sufficient data support (Liu et al., 2020; Xie et al., 2021; Uehara & Sun, 2021). These conservative frameworks, often implemented via lower confidence bounds (LCB) or conservative value iteration, have been shown to achieve minimax optimality across a variety of linear and general function approximation settings (Rashidinejad et al., 2021; Jin et al., 2021; Li et al., 2024).

Complexity and Learning in Multi-Agent Games. The transition from single-agent RL to multi-agent strategic environments significantly expands the computational and statistical complexity of the learning task. Computing a Nash Equilibrium (NE) in generic finite general-sum games is PPAD-complete (Papadimitriou, 1994; Chen et al., 2009; Daskalakis et al., 2009). In the offline setting, joint-action coverage requirements can scale exponentially with the number of players (Cui & Du, 2022b;a). Recent advancements in offline Markov games have mitigated this by utilizing *strategy-wise* uncertainty bonuses and *unilateral concentrability* to recover stable equilibria such as Coarse Correlated Equilibria (CCE) and Nash Equilibria (Zhong et al., 2022; Zhang et al., 2023b; Yan et al., 2024). Our work demonstrates that the α-potential structure provides a decentralized alternative for learning approximate equilibria.

Regularized Learning and LLM Alignment. KL regularization relative to a fixed reference policy has become a cornerstone of behavioral anchoring in reinforcement learning (Schulman et al., 2017; Xiong et al., 2023). By penalizing deviations from an initial distribution, this mechanism effectively stabilizes policy updates and prevents the "forgetting" or drift often seen in high-dimensional optimization (Abdolmaleki et al., 2018; Zhang et al., 2025). This approach is most prominently utilized in the alignment of Large Language Models (LLMs) via Reinforcement Learning from Human Feedback (RLHF), where a KL penalty ensures the model remains anchored to a trusted reference distribution (Ouyang et al., 2022; Rafailov et al., 2023; Munos et al., 2024). While regularized dynamics have been analyzed for single-agent tasks and multi-agent games (Xie et al., 2024; Ye et al., 2024; Zhao et al., 2025; Zhang et al., 2026; Chen et al., 2026), our work characterizes the efficacy of KL regularization for equilibrium learning in general-sum potential games, bypassing the need for explicit pessimistic bonuses.

Statistical Efficiency and Fast Rates. Many offline-learning guarantees exhibit $\widetilde{\mathcal{O}}(1/\sqrt{n})$ statistical error rates (Jin et al., 2021; Shi et al., 2022). While recent investigations have identified accelerated $\widetilde{\mathcal{O}}(1/n)$ "fast rates" in specific settings, these results include single-agent problems or two-player zero-sum games that leverage skew-symmetry and minimax dynamics (Rashidinejad et al., 2021; Nayak et al., 2025; Zhang et al., 2026). By exploiting the geometric properties of the KL-regularized potential function and the log-linear interpolation of Mirror Descent iterates, we establish that the fast $\widetilde{\mathcal{O}}(1/n)$ rate is achievable for approximate Nash equilibria of the regularized game, providing a new benchmark for statistical efficiency in multi-agent strategic decision-making.

B Proof

The appendix is organized around three structural ingredients of our analysis. Section B.1 isolates how the α-potential landscape turns decentralized mirror descent into approximate ascent on the global potential. Section B.2 develops the log-linear interpolation identity that controls the accumulated score-tracking error. Section B.3 combines these ingredients with a Bregman bound on the regularized best-response gap, a fitted-to-true reward transfer, and a single concentrability-shift argument to close both Theorem 5.1 and Theorem 6.1.

NOTATION, DOMAINS, AND THE PER-ITERATE DECOMPOSITION

Throughout the analysis, the reference policy has full support for ρ-almost every context.

Return and gap notation used in the proofs. We use $\mathrm{KL}(p\|q) = \sum_a p(a)\log(p(a)/q(a))$, with natural logarithms and $0\log 0 = 0$. For the calculations below, write

$$U_i(\pi) = \mathbb{E}_{x\sim\rho,\boldsymbol{a}\sim\pi(\cdot|x)} r_i(x,\boldsymbol{a}),$$
$$K_i(\pi_i) = \mathbb{E}_\rho \mathrm{KL}(\pi_i(\cdot\mid x)\|\pi_i^{\mathrm{ref}}(\cdot\mid x)), \qquad J_i(\pi) = \mathbb{E}_\rho V_i^\pi = U_i(\pi) - \eta^{-1}K_i(\pi_i).$$

We write $J_i(\pi) = \mathbb{E}_\rho V_i^\pi$. The finite-expected-KL domain for these return differences is $\Pi = \prod_i \Pi_i$; the reference-anchored coverage bound extends to all unilateral rules by the approximation in (25).

For pointwise and player-wise gaps, write

$$g_i(x;\pi) = V_i^{\dagger,\pi_{-i}}(x) - V_i^\pi(x), \qquad g_i(\pi) = \mathbb{E}_\rho g_i(x;\pi),$$
$$\mathrm{Gap}_{\mathrm{NE}}(\pi) = \sum_i g_i(\pi), \qquad \widehat{\mathrm{Gap}}_{\mathrm{NE}}(\pi) = \sum_i \mathbb{E}_\rho \varepsilon_{\mathrm{pot},i}(x;\pi).$$

Here $\varepsilon_{\mathrm{pot},i}(x;\pi) = \widehat{V}_i^{\dagger,\pi_{-i}}(x) - \widehat{V}_i^\pi(x)$ is the empirical unilateral gap. Nonnegativity gives $\max_i g_i(\pi) \le \mathrm{Gap}_{\mathrm{NE}}(\pi) \le m\max_i g_i(\pi)$. Thus a bound on the total gap also bounds each player's gap; a total bound $m\epsilon$ alone does not imply an ϵ-NE.

All logarithms are natural. For a finite action vector v, $\|v\|_1 = \sum_a |v(a)|$, $\|v\|_\infty = \max_a |v(a)|$, and $\langle v, w\rangle = \sum_a v(a)w(a)$. For two probability vectors, $\mathrm{TV}(p,q) = \frac{1}{2}\|p-q\|_1$. The notation $\mathbb{E}_\rho$ integrates contexts and $\mathbb{E}_\mu$ integrates logged context–action pairs. Conditional expectations over actions always hold the displayed context fixed.

Symbol	Meaning
n, T, δ	Number of offline rows, number of update sweeps, and dataset failure probability.
η, γ, β	Inverse regularization strength, proximal step parameter, and interpolation weight $\beta := \gamma/(\eta+\gamma) \in (0,1)$.
α, Φ	Value-level tolerance and true regularized potential; no fitted potential is assumed.
$\widehat{Q}_i, \mathcal{Z}_i$	Fitted joint reward and error $\mathcal{Z}_i = \widehat{Q}_i - r_i$;
$\sigma_i^2, \mathcal{E}$	Logger population risk $\mathbb{E}_\mu \mathcal{Z}_i^2$ and its sum $\sum_i \sigma_i^2$; both depend on the observed sample.
$Q_i^{\pi_{-i}}, \hat{Q}_i(x,\cdot,\pi_{-i})$	True and fitted reward vectors after averaging over the same opponents π_{-i}.
$d_i(x;\pi)$	$\|\hat{Q}_i(x,\cdot,\pi_{-i}) - Q_i^{\pi_{-i}}(x,\cdot)\|_\infty$; $d_i^t(x) = d_i(x;\pi^{(t)})$.
$g_i, \varepsilon_{\mathrm{pot},i}$	Pointwise true and empirical unilateral gaps; $\widehat{G}_t = \sum_i \mathbb{E}_\rho \varepsilon_{\mathrm{pot},i}^{(t)}(x)$.
h_i^t	Bounded score in $\pi_i^{(t)} \propto \pi_i^{\mathrm{ref}} e^{\eta h_i^t}$; it is not the reward estimate $\widehat{Q}_i$.
$C_{\mathrm{shift}}C_{\mathrm{uni}}, B_{\eta,\gamma,m}$	$C_{\mathrm{uni}}e^{\eta(m-1)}$ and $\eta^2(m-1)/\beta$, respectively.
X_t, Y_t, D_t	Sums over players and contexts of forward step KL, reverse step KL, and squared marginal errors; see (10).

Marginalization shorthand. For any opponent profile π_{-i} and player i, define the marginalized true and empirical Q-values as

$$Q_i^{\pi_{-i}}(x,a_i) := \mathbb{E}_{\boldsymbol{a}_{-i}\sim\pi_{-i}(\cdot|x)}\big[r_i(x,a_i,\boldsymbol{a}_{-i})\big],$$
$$\hat{Q}_i(x,a_i,\pi_{-i}) := \mathbb{E}_{\boldsymbol{a}_{-i}\sim\pi_{-i}(\cdot|x)}\big[\hat{Q}_i(x,a_i,\boldsymbol{a}_{-i})\big].$$

Both are functions of a_i alone after marginalization, and the marginalized regression error inherits the same convention, $\mathcal{Z}_i(x,a_i,\pi_{-i}) := \mathbb{E}_{\boldsymbol{a}_{-i}\sim\pi_{-i}}[\mathcal{Z}_i(x,a_i,\boldsymbol{a}_{-i})]$. At iteration t, $\bar{Q}_i^{(t)}(x,a_i) = \hat{Q}_i(x,a_i,\pi_{-i}^{(t)})$. In particular,

$$\mathcal{Z}_i(x,a_i,\pi_{-i}) = \hat{Q}_i(x,a_i,\pi_{-i}) - Q_i^{\pi_{-i}}(x,a_i). \tag{3}$$

True and Empirical Values. At a fixed player and context, suppress (i, x) and set

$$f_q(p) = \langle q, p\rangle - \eta^{-1}\mathrm{KL}(p\|\pi_i^{\mathrm{ref}}),$$

$$V_i^{\pi}(x) = f_{Q_i^{\pi_{-i}}}(\pi_i), \qquad \widehat{V}_i^{\pi}(x) = f_{\widehat{Q}_i(x,\cdot,\pi_{-i})}(\pi_i),$$

$$V_i^{\dagger,\pi_{-i}}(x) = \max_{\pi_i'\in\Delta(\mathcal{A}_i)} f_{Q_i^{\pi_{-i}}}(\pi_i'), \qquad \widehat{V}_i^{\dagger,\pi_{-i}}(x) = \max_{\pi_i'\in\Delta(\mathcal{A}_i)} f_{\widehat{Q}_i(x,\cdot,\pi_{-i})}(\pi_i').$$

The dagger indicates optimization of player i only, with opponents fixed; the hat indicates fitted rewards. In particular, the empirical best response is generally different from the current policy π_i.

The pointwise maximizer is the measurable Gibbs policy $b_q(a_i) = \pi_i^{\mathrm{ref}}(a_i)e^{\eta q(a_i)}/\sum_b \pi_i^{\mathrm{ref}}(b)e^{\eta q(b)}$. Consequently, integrating g_i also gives $\sup_{\pi_i'\in\Pi_i} J_i(\pi_i', \pi_{-i}) - J_i(\pi)$; this is the global unilateral improvement appearing in the Nash definition.

Per-iterate decomposition. Throughout the appendix we write the per-iterate gap of player i as a sum of three named quantities:

$$V_i^{\dagger,\hat{\pi}_{-i}}(x) - V_i^{\hat{\pi}}(x) \;=\; \varepsilon_{\mathrm{pot},i}(x) + \delta_{\mathrm{BR},i}(x) + \delta_{\mathrm{Iter},i}(x), \tag{4}$$

where, using the shorthand $V_i^{\dagger} \equiv V_i^{\dagger,\hat{\pi}_{-i}}$ and $\hat{V}_i^{\dagger} \equiv \hat{V}_i^{\dagger,\hat{\pi}_{-i}}$,

$$\varepsilon_{\mathrm{pot},i}(x) \coloneqq \hat{V}_i^{\dagger}(x) - \hat{V}_i^{\hat{\pi}}(x), \qquad \text{(empirical equilibrium gap)}$$

$$\delta_{\mathrm{BR},i}(x) \coloneqq V_i^{\dagger}(x) - \hat{V}_i^{\dagger}(x), \qquad \text{(evaluation bias of best response)}$$

$$\delta_{\mathrm{Iter},i}(x) \coloneqq \hat{V}_i^{\hat{\pi}}(x) - V_i^{\hat{\pi}}(x). \qquad \text{(iterate noise)}$$

For a generic profile π, we append $;\pi$ to these three quantities to make the evaluated profile explicit. Identity (4) is verified by adding and subtracting $\hat{V}_i^{\dagger}(x)$ and $\hat{V}_i^{\hat{\pi}}(x)$ in the gap:

$$\begin{aligned} V_i^{\dagger}(x) - V_i^{\hat{\pi}}(x) &= \big(V_i^{\dagger}(x) - \hat{V}_i^{\dagger}(x)\big) + \big(\hat{V}_i^{\dagger}(x) - \hat{V}_i^{\hat{\pi}}(x)\big) + \big(\hat{V}_i^{\hat{\pi}}(x) - V_i^{\hat{\pi}}(x)\big) \\ &= \delta_{\mathrm{BR},i}(x) + \varepsilon_{\mathrm{pot},i}(x) + \delta_{\mathrm{Iter},i}(x), \end{aligned}$$

which matches the algebraic decomposition appearing in the offline-RL literature (cf. Chen & Zhang (2026)). The empirical equilibrium gap $\varepsilon_{\mathrm{pot},i}$ is zero whenever $\hat{\pi}$ is an exact regularized empirical equilibrium of the estimated game (the ROPE setting); in the OPMD setting it is tracked by the true-potential progress and score-tracking bounds (Section B.1). The evaluation bias and the iterate noise are statistical quantities that we control through the regression error $\mathcal{Z}_i(x, \boldsymbol{a}) = \widehat{Q}_i(x, \boldsymbol{a}) - r_i(x, \boldsymbol{a})$.

For the OPMD analysis we attach an iteration index and write $\varepsilon_{\mathrm{pot},i}^{(t)}, \delta_{\mathrm{BR},i}^{(t)}, \delta_{\mathrm{Iter},i}^{(t)}$.

Fitted games and potential structure. Independent reward estimation generally does not preserve the potential structure of the original game. Our OPMD analysis uses the original game's potential to control the updates and treats fitted-reward errors explicitly.

Probability and Domain Conventions. We first establish deterministic inequalities for fixed fitted reward functions and policies. The only high-probability event is the least-squares risk bound in Lemma B.10; after conditioning on that event, all times and data-dependent policies are handled by a uniform deterministic coverage bound. Potential differences use Π so that no subtraction of infinite returns occurs. Full-support Gibbs policies make the pointwise KL identities well-defined, and their bounded-score representation gives finite integrated KL.

B.1 Structural Bias of the Potential Landscape

Proposition B.1 (Potential maximizers)**.** *If $\Phi(\widehat{\pi}) \geq \sup_{\Pi}\Phi - \epsilon$, then $g_i(\widehat{\pi}) \leq \epsilon + \alpha$ for every player and $\mathrm{Gap}_{\mathrm{NE}}(\widehat{\pi}) \leq m(\epsilon + \alpha)$.*

Proof. For every admissible π_i', (1) bounds its improvement by $\Phi(\pi_i', \widehat{\pi}_{-i}) - \Phi(\widehat{\pi}) + \alpha \leq \epsilon + \alpha$. Taking the supremum proves the claim. The Gibbs best response is admissible. □

We analyze one simultaneous OPMD step by interpolating between the old and new profiles through m unilateral changes. Applying the α-potential condition along this path relates the players' true payoff improvements to progress of the global potential, with total discrepancy at most $m\alpha$.

B.1.1 True-Potential Progress for Simultaneous Updates

$$0 < \gamma \le \frac{1}{\max\{1, m-1\}}, \qquad \beta := \frac{\gamma}{\eta + \gamma}, \qquad B_{\eta,\gamma,m} = \frac{\eta^2(m-1)}{\beta}. \tag{5}$$

Lemma B.2 (Perturbed finite-difference progress)**.** *Under Assumption 3.1 and (5), let* $X_t = \sum_i \mathbb{E}_\rho \mathrm{KL}(\pi_i^{(t)} \| \pi_i^{(t+1)})$ *and* $D_t = \sum_i \mathbb{E}_\rho (d_i^t)^2$*. Then (6) holds.*

$$\Phi(\pi^{(t+1)}) - \Phi(\pi^{(t)}) \ge \frac{X_t}{\eta\beta} - \gamma D_t - m\alpha. \tag{6}$$

Proof.

Step 1: Telescope Along Unilateral Changes. The actual algorithm updates simultaneously. The hybrid path below is an analytical device for applying the unilateral potential condition; it is not a different coordinate-update algorithm. Write $\pi = \pi^{(t)}$, $\pi^+ = \pi^{(t+1)}$ and form hybrid profiles $\pi^{[i]} = (\pi_1^+, \dots, \pi_i^+, \pi_{i+1}, \dots, \pi_m)$. Set $\pi^{[0]} = \pi$ and $\pi^{[m]} = \pi^+$. The hybrid profiles all have finite integrated KL by Lemma B.4. The potential condition applied to the ith edge is $\Phi(\pi^{[i]}) - \Phi(\pi^{[i-1]}) \ge J_i(\pi^{[i]}) - J_i(\pi^{[i-1]}) - \alpha$.

Summing these m inequalities cancels the intermediate potential values and gives

$$\Phi(\pi^+) - \Phi(\pi) \ge \sum_i [J_i(\pi^{[i]}) - J_i(\pi^{[i-1]})] - m\alpha. \tag{7}$$

Step 2: Bound the Effect of Moving Opponents. We now compare each gain to the gain with opponents frozen at π_{-i}. At a fixed context set $s_i = \|\pi_i^+ - \pi_i\|_1$ and $\Delta_i = \pi_i^+ - \pi_i$. The function of opponents' joint action $v_i(\boldsymbol{a}_{-i}) = \sum_{a_i} \Delta_i(a_i) r_i(x, \boldsymbol{a})$ lies in $[-s_i/2, s_i/2]$: the positive and negative parts of Δ_i each have mass $s_i/2$, and rewards lie in $[0, 1]$. Thus its oscillation is at most s_i. Replacing the updated opponents by old opponents changes its expectation by at most

$$s_i \,\mathrm{TV}\left(\prod_{j<i} \pi_j^+ \prod_{j>i} \pi_j, \prod_{j\ne i} \pi_j\right) \le \frac{s_i}{2} \sum_{j<i} s_j. \tag{8}$$

The signed vector Δ_i has zero total mass, so $\sum_a (\Delta_i(a))_+ = \sum_a (-\Delta_i(a))_+ = s_i/2$. For any function with range in $[a, b]$, changing a probability distribution from P to Q changes its expectation by at most $(b-a)\,\mathrm{TV}(P, Q)$. Applying this with $b - a = s_i$ gives the first inequality in (8). Here product total variation is bounded by the sum of marginal total variations, which follows by replacing one factor at a time and taking the triangle inequality. The own-policy KL difference is unchanged by the opponent profile. Summing (8) and using $s_i s_j \le (s_i^2 + s_j^2)/2$ yields

$$\frac{1}{2} \sum_i \sum_{j<i} s_i s_j \le \frac{m-1}{4} \sum_i s_i^2. \tag{9}$$

Each unordered pair occurs once in $\sum_i \sum_{j<i} s_i s_j$. After applying $2 s_i s_j \le s_i^2 + s_j^2$, each s_i^2 appears $m - 1$ times, which explains the coefficient $(m-1)/4$ in (9).

Step 3: Insert the Fitted-Reward Improvement. At context x, define the frozen-opponent gains

$$I_i = \langle Q_i^{\pi_{-i}}, \Delta_i \rangle - \eta^{-1} \big[\mathrm{KL}(\pi_i^+ \| \pi_i^{\mathrm{ref}}) - \mathrm{KL}(\pi_i \| \pi_i^{\mathrm{ref}})\big],$$
$$\widehat{I}_i = \langle \hat{Q}_i(x, \cdot, \pi_{-i}), \Delta_i \rangle - \eta^{-1} \big[\mathrm{KL}(\pi_i^+ \| \pi_i^{\mathrm{ref}}) - \mathrm{KL}(\pi_i \| \pi_i^{\mathrm{ref}})\big].$$

The KL terms cancel, and therefore $I_i - \widehat{I}_i = \langle Q_i^{\pi_{-i}} - \hat{Q}_i(x, \cdot, \pi_{-i}), \Delta_i \rangle \ge -d_i^t s_i$. The true and empirical frozen-opponent gains differ by at most $d_i^t s_i$. Use the exact proximal identity (12) and define

$$X_t = \sum_i \mathbb{E}_\rho \mathrm{KL}(\pi_i^{(t)} \| \pi_i^{(t+1)}), \qquad Y_t = \sum_i \mathbb{E}_\rho \mathrm{KL}(\pi_i^{(t+1)} \| \pi_i^{(t)}),$$
$$D_t = \sum_i \mathbb{E}_\rho (d_i^t)^2. \tag{10}$$

Substitution in (7) gives

$$\Phi(\pi^+) - \Phi(\pi) \geq (\eta^{-1} + \gamma^{-1})X_t + \gamma^{-1}Y_t - m\alpha - \sum_i \mathbb{E}_\rho \left[d_i^t s_i + \frac{m-1}{4} s_i^2 \right].$$

Step 4: Absorb Both Error Terms. Young's inequality gives $d_i^t s_i \leq s_i^2/(4\gamma) + \gamma(d_i^t)^2$; this follows by expanding $(s_i/(2\sqrt{\gamma}) - \sqrt{\gamma} d_i^t)^2 \geq 0$. Pinsker gives $s_i^2 \leq 2\mathrm{KL}(\pi_i^+ \| \pi_i)$. Since $\eta^{-1} + \gamma^{-1} = 1/(\eta\beta)$, we obtain

$$\begin{aligned}\Phi(\pi^+) - \Phi(\pi) &\geq \frac{X_t}{\eta\beta} + \left[\frac{1}{2\gamma} - \frac{m-1}{2} \right] Y_t - \gamma D_t - m\alpha \\ &\geq \frac{X_t}{\eta\beta} - \gamma D_t - m\alpha.\end{aligned}$$

The last coefficient is nonnegative because $(m-1)\gamma \leq 1$. This proves (6), including $m = 1$, where the cross-player sum is empty. The forward KL is the energy quantity in the score-tracking bound. □

B.1.2 One-Sided Endpoint Control

Lemma B.3 (One-sided endpoint bound)**.** *For every* $\pi \in \Pi$,

$$\Phi(\pi) - \Phi(\pi^{\mathrm{ref}}) \leq m(1+\alpha). \tag{11}$$

Proof. Define $\zeta^{[i]} = (\pi_1, \ldots, \pi_i, \pi_{i+1}^{\mathrm{ref}}, \ldots, \pi_m^{\mathrm{ref}})$, with $\zeta^{[0]} = \pi^{\mathrm{ref}}$ and $\zeta^{[m]} = \pi$. For the ith change,

$$\begin{aligned}J_i(\zeta^{[i]}) - J_i(\zeta^{[i-1]}) &= U_i(\zeta^{[i]}) - U_i(\zeta^{[i-1]}) - \eta^{-1} K_i(\pi_i) \\ &\leq 1 - \eta^{-1} K_i(\pi_i) \leq 1.\end{aligned}$$

The potential difference on each edge is at most that gain plus α by (1). Summing the m changes gives the endpoint bound. □

B.2 Geometric Stability of Mirror Descent

The OPMD update in Algorithm 2 is, in fact, a closed-form log-linear interpolation between the current iterate and the regularized empirical best response. Recall the explicit form in Algorithm 2, equivalently

$$\pi_i^{(t+1)} \propto (\pi_i^{(t)})^{1-\beta} (\widehat{\pi}_i^{\dagger,(t)})^{\beta}, \qquad \widehat{\pi}_i^{\dagger,(t)} \propto \pi_i^{\mathrm{ref}} \exp(\eta \bar{Q}_i^{(t)}).$$

We first establish the bounded-score representation and exact proximal identity, then control the accumulated score-tracking error.

B.2.1 Bounded Scores and Reference Densities

Lemma B.4 (Envelope)**.** *Every OPMD iterate can be written* $\pi_i^{(t)} \propto \pi_i^{\mathrm{ref}} e^{\eta h_i^t}$ *with* $0 \leq h_i^t \leq 1$. *The same bounded-score representation holds for empirical best responses and exact empirical equilibria. Consequently, pointwise,*

$$e^{-\eta} \pi_i^{\mathrm{ref}} \leq \pi_i^{(t)} \leq e^{\eta} \pi_i^{\mathrm{ref}}, \qquad \mathrm{KL}(\pi_i^{(t)} \| \pi_i^{\mathrm{ref}}) \leq \eta.$$

Proof.

Induction and Normalization. Write $N_i^t(x) = \sum_a \pi_i^{\mathrm{ref}}(a \mid x) e^{\eta h_i^t(x,a)}$ for the partition function. Initialize $h_i^1 = 0$. The normalizing constant of $\pi_i^{(t)}$ disappears when normalizing its next update, leaving $h_i^{t+1} = (1-\beta)h_i^t + \beta \bar{Q}_i^{(t)}$. Indeed, substituting $\pi_i^{(t)} = \pi_i^{\mathrm{ref}} e^{\eta h_i^t} / N_i^t$ into the update gives

$$(\pi_i^{\mathrm{ref}})^{\beta} (\pi_i^{(t)})^{1-\beta} e^{\eta\beta \bar{Q}_i^{(t)}} = (N_i^t)^{-(1-\beta)} \pi_i^{\mathrm{ref}} \exp\{\eta[(1-\beta)h_i^t + \beta \bar{Q}_i^{(t)}]\}.$$

The factor $(N_i^t)^{-(1-\beta)}$ is independent of the action and cancels from the normalized policy. This proves the score recursion, rather than requiring that normalized logits obey an uncorrected affine identity. The empirical marginalized rewards lie in $[0, 1]$, so induction proves the representation. Its partition function is between 1 and e^η, giving both density inequalities and $\log(\pi_i^{(t)}/\pi_i^{\mathrm{ref}}) \le \eta$. A best response simply uses score $\hat{Q}_i(x, \cdot, \pi_{-i})$; at an empirical equilibrium each player is such a best response. These inequalities imply finite integrated KL. □

B.2.2 Exact Proximal Improvement

Lemma B.5 (Two KL orientations)**.** *Fix i, x, t, write $p = \pi_i^{(t)}$, $p^+ = \pi_i^{(t+1)}$, and $f(\pi_i') = \langle \bar{Q}_i^{(t)}, \pi_i' \rangle - \eta^{-1}\mathrm{KL}(\pi_i' \| \pi_i^{\mathrm{ref}})$. Then*

$$p^+ = \arg\max_{\pi_i'} \{f(\pi_i') - \gamma^{-1}\mathrm{KL}(\pi_i' \| p)\},$$
$$f(p^+) - f(p) = (\eta^{-1} + \gamma^{-1})\mathrm{KL}(p \| p^+) + \gamma^{-1}\mathrm{KL}(p^+ \| p). \tag{12}$$

Proof.

Step 1: Derive the Normalized Update. Let $q = \bar{Q}_i^{(t)}$ and introduce a multiplier for $\sum_a \pi_i'(a) = 1$. The stationary condition in the relative interior is

$$q(a) - \eta^{-1}\big(\log(\pi_i'(a)/\pi_i^{\mathrm{ref}}(a)) + 1\big) - \gamma^{-1}\big(\log(\pi_i'(a)/p(a)) + 1\big) = \lambda.$$

Collecting the action-dependent terms gives

$$(\eta^{-1} + \gamma^{-1}) \log \pi_i'(a) = q(a) + \eta^{-1} \log \pi_i^{\mathrm{ref}}(a) + \gamma^{-1} \log p(a) + c,$$

where c is independent of a. After exponentiating and normalizing,

$$p^+(a) = \frac{\pi_i^{\mathrm{ref}}(a)^\beta p(a)^{1-\beta} e^{\eta\beta q(a)}}{\sum_b \pi_i^{\mathrm{ref}}(b)^\beta p(b)^{1-\beta} e^{\eta\beta q(b)}}.$$

Strict concavity gives uniqueness. The displayed policy is positive, and the entropy terms ensure the optimum is interior when π_i^{ref}, p are positive. Lemma B.4 supplies this positivity at every iteration.

Step 2: Retain the Exact Entropy Remainder. For $K(\pi_i') = \mathrm{KL}(\pi_i' \| \pi_i^{\mathrm{ref}})$, direct expansion shows $K(p) - K(p^+) - \langle \nabla K(p^+), p - p^+ \rangle = \mathrm{KL}(p \| p^+)$.

The linear reward has zero remainder. Thus the next calculation is an identity, rather than a lower bound obtained by comparing just two objective values. At p^+ the vector $\nabla f(p^+) - \gamma^{-1}(\log(p^+/p) + \mathbf{1})$ is a scalar multiple of $\mathbf{1}$. Negative entropy has exact Bregman remainder, hence

$$\begin{aligned} f(p^+) - f(p) &= \langle \nabla f(p^+), p^+ - p \rangle + \eta^{-1}\mathrm{KL}(p \| p^+) \\ &= \gamma^{-1}\langle \log(p^+/p), p^+ - p \rangle + \eta^{-1}\mathrm{KL}(p \| p^+). \end{aligned}$$

Step 3: Identify Both KL Orientations. Expanding the last inner product term by term gives

$$\sum_a (p^+(a) - p(a)) \log \frac{p^+(a)}{p(a)} = \sum_a p^+(a) \log \frac{p^+(a)}{p(a)} + \sum_a p(a) \log \frac{p(a)}{p^+(a)}.$$

The inner product equals $\mathrm{KL}(p^+ \| p) + \mathrm{KL}(p \| p^+)$; normalizing constants vanish because the displacement sums to zero. □

B.2.3 Accumulated Score-Tracking Error

The bounded-score representation permits a cumulative comparison between the current policy and its changing empirical best response. Auxiliary-logit tracking also appears in Cen et al. (2022, Section 4.2 of the extended arXiv v2 manuscript). Here we use a squared tracking bound and a quadratic log-partition remainder, together with true-potential progress, to control offline estimation error.

Lemma B.6 (Score tracking and empirical exploitability)**.** *For the simultaneous OPMD updates with fixed fitted joint rewards in* $[0,1]$*, initial scores* $h_i^1 = 0$*, and* $\beta \in (0,1)$*,*

$$\frac{1}{T}\sum_{t=1}^{T} \widehat{G}_t \le \frac{\eta m}{\beta(2-\beta)T} + \frac{\eta(m-1)}{2\beta^2 T}\sum_{t=1}^{T} X_t.$$

The bound follows from the bounded-score recursion for fixed fitted rewards.

Proof.

Step 1: Expand the Score Recursion. Fix a context and player. Lemma B.4 gives $h^{t+1} = (1-\beta)h^t + \beta\bar{Q}^{(t)}$ and $\pi^{(t)} \propto \pi^{\mathrm{ref}}\exp(\eta h^t)$, with $h^1 = 0$. Define vectors $r_t = \bar{Q}^{(t)} - h^t$, $v_t = \bar{Q}^{(t+1)} - \bar{Q}^{(t)}$, and $a = 1-\beta \in (0,1)$. Direct subtraction, followed by induction, gives

$$r_{t+1} = ar_t + v_t, \qquad r_t = a^{t-1}r_1 + \sum_{k=1}^{t-1} a^{t-1-k}v_k.$$

The recursion tracks unnormalized score vectors. Since $\bar{Q}^{(1)} \in [0,1]^{|\mathcal{A}_i|}$, $\|r_1\|_\infty \le 1$. Applying the triangle inequality, then $(u+v)^2 \le 2u^2 + 2v^2$ and weighted Cauchy–Schwarz, yields

$$\begin{aligned}
\|r_t\|_\infty^2 &\le 2a^{2(t-1)} + 2\left(\sum_{k<t} a^{t-1-k}\|v_k\|_\infty\right)^2 \\
&\le 2a^{2(t-1)} + 2\left(\sum_{k<t} a^{t-1-k}\right)\left(\sum_{k<t} a^{t-1-k}\|v_k\|_\infty^2\right) \\
&\le 2a^{2(t-1)} + \frac{2}{\beta}\sum_{k<t} a^{t-1-k}\|v_k\|_\infty^2.
\end{aligned}$$

For $t = 1$, the inner sums are empty. For all $T \ge 1$, summing the finite geometric series and interchanging only finite, nonnegative sums gives

$$\begin{aligned}
\sum_{t=1}^{T}\|r_t\|_\infty^2 &\le \frac{2}{1-a^2} + \frac{2}{\beta}\sum_{k=1}^{T-1}\|v_k\|_\infty^2 \sum_{t=k+1}^{T} a^{t-1-k} \\
&\le \frac{2}{\beta(2-\beta)} + \frac{2}{\beta^2}\sum_{k=1}^{T-1}\|v_k\|_\infty^2. \qquad (13)
\end{aligned}$$

The geometric factor comes from score averaging; the second term accounts for payoff motion.

Step 2: Bound Payoff Motion by Joint Opponent KL. At a fixed own action, the same fitted joint reward in $[0,1]$ is averaged against two successive opponent distributions. Hence, uniformly in that action,

$$\begin{aligned}
\|\bar{Q}_i^{(t+1)} - \bar{Q}_i^{(t)}\|_\infty^2 &\le \mathrm{TV}(\pi_{-i}^{(t+1)}, \pi_{-i}^{(t)})^2 \\
&\le \tfrac{1}{2}\mathrm{KL}(\pi_{-i}^{(t)}\|\pi_{-i}^{(t+1)}) = \tfrac{1}{2}\sum_{j\ne i}\mathrm{KL}(\pi_j^{(t)}\|\pi_j^{(t+1)}).
\end{aligned}$$

The equality is additivity of KL for product distributions: expanding the log likelihood ratio gives a sum over j, and integrating out all other action coordinates leaves their marginal KL terms. Summing over i counts each player's KL exactly $m-1$ times, so

$$\sum_i \mathbb{E}_\rho\|\bar{Q}_i^{(t+1)} - \bar{Q}_i^{(t)}\|_\infty^2 \le \frac{m-1}{2}X_t. \qquad (14)$$

Pinsker was applied to the joint opponent distribution before using additivity. For $m = 1$, there are no moving opponents, and both sides are zero.

Step 3: Convert Squared Tracking Error to Gap. For fixed (i, x), let $L(z) = \eta^{-1} \log \sum_b \pi_i^{\mathrm{ref}}(b \mid x) \exp(\eta z_b)$. Its gradient at h_i^t is $\pi_i^{(t)}$, and $f_{\bar{Q}_i^{(t)}}(\pi_i^{(t)}) = L(h_i^t) + \langle \pi_i^{(t)}, \bar{Q}_i^{(t)} - h_i^t \rangle$. Therefore the empirical gap is

$$\varepsilon_{\mathrm{pot},i}^{(t)}(x) = L(\bar{Q}_i^{(t)}) - L(h_i^t) - \langle \nabla L(h_i^t), \bar{Q}_i^{(t)} - h_i^t \rangle.$$

For any vector w, the Hessian satisfies $w^\top \nabla^2 L(z) w = \eta \operatorname{Var}_{b_z}(w) \le \eta \|w\|_\infty^2$. Taylor's formula with integral remainder and $\int_0^1 (1-u)\, du = 1/2$ now give

$$\varepsilon_{\mathrm{pot},i}^{(t)}(x) \le \frac{\eta}{2} \|\bar{Q}_i^{(t)} - h_i^t\|_\infty^2.$$

Combining (13) and (14), summing over players, and integrating contexts yields

$$\sum_{t=1}^{T} \widehat{G}_t \le \frac{\eta m}{\beta(2-\beta)} + \frac{\eta(m-1)}{2\beta^2} \sum_{t=1}^{T-1} X_t.$$

Adding the nonnegative X_T term and dividing by T proves the lemma. Bounded scores and full-support references ensure the pointwise identities and finite integrated KL. □

B.3 Synthesis of Statistical and Potential Rates

This section proves the two main theorems. We first develop three reusable ingredients: a Bregman bound on the regularized best-response gap (Section B.3.1), a transfer bound for $\delta_{\mathrm{BR},i} + \delta_{\mathrm{Iter},i}$ (Section B.3.2), and a single concentrability-shift argument (Section B.3.3). The two theorems are then short corollaries (Sections B.3.5 and B.3.6).

B.3.1 Bregman Bound on the Regularized Best-Response Gap

For a fixed player and context, define

$$L(q) = \eta^{-1} \log \sum_a \pi_i^{\mathrm{ref}}(a) e^{\eta q(a)}, \qquad b_q(a) = \frac{\pi_i^{\mathrm{ref}}(a) e^{\eta q(a)}}{\sum_c \pi_i^{\mathrm{ref}}(c) e^{\eta q(c)}}. \tag{15}$$

Variational identity. Full support makes b_q positive. From (15), $\log(b_q(a)/\pi_i^{\mathrm{ref}}(a)) = \eta q(a) - \eta L(q)$. For any probability vector p, with $0 \log 0 = 0$, substitution gives

$$\begin{aligned} \mathrm{KL}(p \| b_q) &= \sum_a p(a) \log \frac{p(a)}{\pi_i^{\mathrm{ref}}(a)} - \sum_a p(a) \log \frac{b_q(a)}{\pi_i^{\mathrm{ref}}(a)} \\ &= \mathrm{KL}(p \| \pi_i^{\mathrm{ref}}) - \eta \langle q, p \rangle + \eta L(q). \end{aligned}$$

Consequently,

$$L(q) - f_q(p) = \eta^{-1} \mathrm{KL}(p \| b_q). \tag{16}$$

This also proves $L(q) = \max_p f_q(p)$, with unique maximizer b_q, since KL is nonnegative and vanishes exactly at $p = b_q$.

Lemma B.7 (Bregman bound on the regularized best-response gap)**.** *Let $\Psi(\theta) \coloneqq \log \sum_{a \in \mathcal{A}_i} \exp(\theta_a)$ denote the log-partition function. Then Ψ is convex and 1-smooth with respect to $\|\cdot\|_\infty$:*

$$D_\Psi(\theta' \,\|\, \theta) \;\coloneqq\; \Psi(\theta') - \Psi(\theta) - \langle \nabla \Psi(\theta), \theta' - \theta \rangle \;\le\; \tfrac{1}{2} \|\theta' - \theta\|_\infty^2. \tag{17}$$

Consequently, for any product profile $\hat{\pi}$, with the natural-parameter logits $\theta_i^\dagger(x, \cdot) \coloneqq \eta\, Q_i^{\hat{\pi}_{-i}}(x, \cdot) + \log \pi_i^{\mathrm{ref}}(\cdot|x)$ and $\hat{\theta}_i(x, \cdot) \coloneqq \eta\, \hat{Q}_i(x, \cdot, \hat{\pi}_{-i}) + \log \pi_i^{\mathrm{ref}}(\cdot|x)$, and the empirical regularized best response $\hat{\pi}_i^\dagger(\cdot|x) \coloneqq \mathrm{softmax}(\hat{\theta}_i(x, \cdot))$, the unilateral best-response value gap $\Delta_i(x) \coloneqq V_i^{\dagger, \hat{\pi}_{-i}}(x) - \hat{V}_i^{\dagger, \hat{\pi}_{-i}}(x)$ satisfies

$$\Delta_i(x) \;\le\; \mathbb{E}_{a_i \sim \hat{\pi}_i^\dagger(\cdot|x)} \Big[Q_i^{\hat{\pi}_{-i}}(x, a_i) - \hat{Q}_i(x, a_i, \hat{\pi}_{-i}) \Big] + \tfrac{\eta}{2} \big\| Q_i^{\hat{\pi}_{-i}}(x, \cdot) - \hat{Q}_i(x, \cdot, \hat{\pi}_{-i}) \big\|_\infty^2. \tag{18}$$

The expectation is taken under $\hat{\pi}_i^\dagger$, the regularized best response to $\hat{\pi}_{-i}$ on the empirical game, which equals the iterate $\hat{\pi}_i$ whenever $\hat{\pi}$ is an exact regularized NE of the empirical game (as in ROPE); for OPMD iterates the two differ in general.

Proof. Step 1 — A quadratic log-partition remainder. Differentiation of the finite sum gives

$$\nabla L(q) = b_q, \qquad \nabla^2 L(q) = \eta\big(\operatorname{diag}(b_q) - b_q b_q^{\mathsf{T}}\big).$$

For a vector v, therefore,

$$v^{\mathsf{T}}\nabla^2 L(q)v = \eta\left[\sum_a b_q(a)v(a)^2 - \left(\sum_a b_q(a)v(a)\right)^2\right] = \eta \operatorname{Var}_{b_q}(v) \le \eta\|v\|_\infty^2.$$

Set $v = q - \widehat{q}$ and $H(s) = L(\widehat{q} + sv)$ for $s \in [0,1]$. The fundamental theorem of calculus applied twice yields

$$\begin{aligned} L(q) - L(\widehat{q}) - \langle b_{\widehat{q}}, v\rangle &= \int_0^1 \int_0^s H''(u)\, du\, ds \\ &= \int_0^1 (1-u)H''(u)\, du \le \frac{\eta}{2}\|v\|_\infty^2. \end{aligned}$$

This is the Bregman remainder of L; L is distinct from the game's potential Φ. Equivalently,

$$L(q) \le L(\widehat{q}) + \langle b_{\widehat{q}}, q - \widehat{q}\rangle + \frac{\eta}{2}\|q - \widehat{q}\|_\infty^2. \tag{19}$$

For $\eta = 1$ and a uniform reference, $L(\theta) = \Psi(\theta) - \log|\mathcal{A}_i|$. Thus the same Hessian bound proves (17).

Step 2 — Applying the bound to value space. By the closed form of the regularized best response, $V_i^{\dagger,\hat{\pi}_{-i}}(x) = \eta^{-1}\Psi(\theta_i^\dagger)$ and $\hat{V}_i^{\dagger,\hat{\pi}_{-i}}(x) = \eta^{-1}\Psi(\hat{\theta}_i)$, with $\theta_i^\dagger - \hat{\theta}_i = \eta\big(Q_i^{\hat{\pi}_{-i}}(x,\cdot) - \hat{Q}_i(x,\cdot,\hat{\pi}_{-i})\big)$ and $\nabla\Psi(\hat{\theta}_i) = \operatorname{softmax}(\hat{\theta}_i)(\cdot) = \hat{\pi}_i^\dagger(\cdot|x)$ — the empirical regularized best response to $\hat{\pi}_{-i}$, not the iterate $\hat{\pi}_i$. Applying (17) to $(\theta_i^\dagger, \hat{\theta}_i)$ and dividing by η yields (18). □

Lemma B.8 (True gap from empirical gap)**.** *Let $g = L(q) - f_q(p)$, $\widehat{g} = L(\widehat{q}) - f_{\widehat{q}}(p)$, and $d = \|q - \widehat{q}\|_\infty$. Then*

$$g \le 2\widehat{g} + \eta d^2.$$

If $p = b_{\widehat{q}}$, the sharper conclusion is $g \le \eta d^2/2$.

Proof.

Bounding the remaining linear mismatch. The same policy is evaluated under the true and fitted vectors, so its KL penalty cancels when these two evaluations are subtracted. Subtract $f_q(p) = f_{\widehat{q}}(p) + \langle p, q - \widehat{q}\rangle$ from (19). The result is $g \le \widehat{g} + \langle b_{\widehat{q}} - p, q - \widehat{q}\rangle + \eta d^2/2$. Equation (16) and Pinsker give $\|p - b_{\widehat{q}}\|_1 \le \sqrt{2\eta\widehat{g}}$. Therefore the inner product satisfies

$$\begin{aligned} \langle b_{\widehat{q}} - p, q - \widehat{q}\rangle &\le \|b_{\widehat{q}} - p\|_1 \|q - \widehat{q}\|_\infty \\ &\le \sqrt{2\eta\widehat{g}}\, d \le \widehat{g} + \frac{\eta}{2}d^2. \end{aligned} \tag{20}$$

The last step follows from $(\sqrt{\widehat{g}} - \sqrt{\eta/2}\, d)^2 \ge 0$. Adding the previous quadratic remainder gives $g \le 2\widehat{g} + \eta d^2$. If $p = b_{\widehat{q}}$, both the empirical gap and inner product vanish instead. □

B.3.2 Transfer from fitted to true rewards

We use Lemma B.7 to convert the linear part of $\delta_{\mathrm{BR},i}$ into an expectation of the regression error $\mathcal{Z}_i$, and combine it with the on-policy expansion of $\delta_{\mathrm{Iter},i}$ to expose the algebraic cancellation at an empirical regularized equilibrium.

Fix i, x, π, and abbreviate $p = \pi_i$, $q = Q_i^{\pi_{-i}}$, $\widehat{q} = \hat{Q}_i(x,\cdot,\pi_{-i})$, $z = \widehat{q} - q$, and $b = b_{\widehat{q}} = \widehat{\pi}_i^\dagger$.

Pointwise expansion of the marginalized Q-difference. For any opponent profile π_{-i} and any action a_i, the marginalized true and empirical Q-values satisfy

$$\begin{aligned} Q_i^{\pi_{-i}}(x,a_i) - \hat{Q}_i(x,a_i,\pi_{-i}) &= \mathbb{E}_{\boldsymbol{a}_{-i}\sim\pi_{-i}}[r_i(x,a_i,\boldsymbol{a}_{-i}) - \widehat{Q}_i(x,a_i,\boldsymbol{a}_{-i})] \\ &= -\mathbb{E}_{\boldsymbol{a}_{-i}\sim\pi_{-i}}\mathcal{Z}_i(x,a_i,\boldsymbol{a}_{-i}) = -z(a_i). \end{aligned}$$

Pointwise expansion of $\delta_{\mathrm{BR},i}$. Substituting (3) into the linear part of the Bregman bound (18),

$$\begin{aligned}\delta_{\mathrm{BR},i}(x;\pi) &\le \mathbb{E}_{a_i\sim b}[Q_i^{\pi_{-i}}(x,a_i) - \hat{Q}_i(x,a_i,\pi_{-i})] + \tfrac{\eta}{2}d_i(x;\pi)^2 \\ &= \mathbb{E}_{a_i\sim b}\big[\mathbb{E}_{\boldsymbol{a}_{-i}\sim\pi_{-i}}[-\mathcal{Z}_i(x,a_i,\boldsymbol{a}_{-i})]\big] + \tfrac{\eta}{2}d_i(x;\pi)^2 \\ &= \mathbb{E}_{\boldsymbol{a}\sim b\pi_{-i}}[-\mathcal{Z}_i(x,\boldsymbol{a})] + \tfrac{\eta}{2}d_i(x;\pi)^2. \end{aligned} \tag{21}$$

The nested expectation becomes a joint expectation under the product profile $b\pi_{-i}$. The opponents remain π_{-i} throughout this step.

Pointwise expansion of $\delta_{\mathrm{Iter},i}$. The KL-regularized value functions satisfy $V_i^\pi(x) = \mathbb{E}_{\boldsymbol{a}\sim\pi}[r_i(x,\boldsymbol{a})] - \eta^{-1}\,\mathrm{KL}(\pi_i\|\pi_i^{\mathrm{ref}})$ and $\hat{V}_i^\pi(x) = \mathbb{E}_{\boldsymbol{a}\sim\pi}[\hat{Q}_i(x,\boldsymbol{a})] - \eta^{-1}\,\mathrm{KL}(\pi_i\|\pi_i^{\mathrm{ref}})$, so the entropy regularizers cancel in $\delta_{\mathrm{Iter},i}$ and the iterate noise unrolls into a pure expectation of $\mathcal{Z}_i$ under π:

$$\begin{aligned}\delta_{\mathrm{Iter},i}(x;\pi) &= \widehat{V}_i^\pi(x) - V_i^\pi(x) \\ &= \mathbb{E}_{\boldsymbol{a}\sim\pi}[\widehat{Q}_i(x,\boldsymbol{a}) - r_i(x,\boldsymbol{a})] \\ &= \mathbb{E}_{\boldsymbol{a}\sim\pi}\mathcal{Z}_i(x,\boldsymbol{a}) = \langle p, z\rangle. \end{aligned} \tag{22}$$

Combining the two terms. Summing (21)–(22) pointwise and integrating over $x\sim\rho$,

$$\mathbb{E}_\rho[\delta_{\mathrm{BR},i} + \delta_{\mathrm{Iter},i}] \le \underbrace{\mathbb{E}_\rho\big[\mathbb{E}_{\boldsymbol{a}\sim\pi}\mathcal{Z}_i - \mathbb{E}_{\boldsymbol{a}\sim b\pi_{-i}}\mathcal{Z}_i\big]}_{\text{linear mismatch}_i} + \tfrac{\eta}{2}\mathbb{E}_\rho d_i(x;\pi)^2. \tag{23}$$

The first bracket is the *linear mismatch* of player i; the second is the *squared evaluation error*, which we will bound through the concentrability shift in Section B.3.3.

Factoring the linear mismatch into an L_1 distance. Because both expectations in linear mismatch$_i$ share the same opponent marginal π_{-i}, only the player-i marginal differs. Writing the difference explicitly,

$$\begin{aligned}\mathbb{E}_\pi\mathcal{Z}_i - \mathbb{E}_{b\pi_{-i}}\mathcal{Z}_i &= \sum_{a_i,\boldsymbol{a}_{-i}}[\pi_i(a_i\mid x) - b(a_i)]\pi_{-i}(\boldsymbol{a}_{-i}\mid x)\mathcal{Z}_i(x,\boldsymbol{a}) \\ &= \sum_{a_i}[p(a_i) - b(a_i)]z(a_i) \\ &\le \|p-b\|_1\|z\|_\infty = \|p-b\|_1 d_i(x;\pi). \end{aligned}$$

Keeping the marginalized error d_i in this bound is essential for the squared-risk argument; replacing it by a constant loses that dependence.

Two Regimes for the Linear Mismatch.

1. *ROPE.* When $\hat{\pi}$ is the exact regularized NE returned by Algorithm 1, the first-order optimality of each $\hat{\pi}_i$ implies $\hat{\pi}_i(\cdot|x) \equiv \hat{\pi}_i^\dagger(\cdot|x)$ for every i and almost every x. Hence $\|\hat{\pi}_i - \hat{\pi}_i^\dagger\|_1 = 0$ pointwise, and the linear mismatch in (23) *vanishes identically*: the regression error $\mathcal{Z}_i$ cancels exactly between $\delta_{\mathrm{BR},i}$ and $\delta_{\mathrm{Iter},i}$.
2. *OPMD.* At an arbitrary iterate, p and b generally differ. Equation (16) and Pinsker give $\|p - b\|_1 \le \sqrt{2\eta\varepsilon_{\mathrm{pot},i}(x;\pi)}$. Consequently, (20) bounds the mismatch by $\varepsilon_{\mathrm{pot},i}(x;\pi) + \eta d_i(x;\pi)^2/2$. Combining this with the decomposition yields $g_i \le 2\varepsilon_{\mathrm{pot},i} + \eta d_i^2$. The additional empirical gap is controlled by true-potential progress; it is not asserted to vanish at each iterate.

B.3.3 Concentrability Shift to the Offline Distribution

We next transfer the marginalized squared reward error under any OPMD iterate to the logger distribution.

Lemma B.9 (Uniform squared-error shift)**.** *For any product policy whose opponents satisfy $\pi_j \le e^{\eta}\pi_j^{\mathrm{ref}}$, let $\hat{Q}_i(x,\cdot,\pi_{-i}) = \mathbb{E}_{\pi_{-i}}\widehat{Q}_i$ and $d_i(x;\pi) = \|\hat{Q}_i(x,\cdot,\pi_{-i}) - Q_i^{\pi_{-i}}\|_\infty$. Then*

$$\mathbb{E}_\rho d_i(x;\pi)^2 \le C_{\mathrm{shift}} C_{\mathrm{uni}} \sigma_i^2. \tag{24}$$

The assertion is deterministic and uniform over all such policies and bounded fitted functions.

Proof.

Step (i): Marginalization and Jensen's Inequality. For each fixed action a_i, the opponents' weights sum to one. Hence

$$\begin{aligned}|\hat{Q}_i(x,a_i,\pi_{-i}) - Q_i^{\pi_{-i}}(x,a_i)|^2 &= \left|\sum_{\boldsymbol{a}_{-i}} \pi_{-i}(\boldsymbol{a}_{-i} \mid x)\mathcal{Z}_i(x,a_i,\boldsymbol{a}_{-i})\right|^2 \\ &\le \sum_{\boldsymbol{a}_{-i}} \pi_{-i}(\boldsymbol{a}_{-i} \mid x)\mathcal{Z}_i(x,a_i,\boldsymbol{a}_{-i})^2.\end{aligned}$$

Step (ii): Shift the Opponents to the Reference. The product envelope gives $\pi_{-i}(\boldsymbol{a}_{-i} \mid x) \le e^{\eta(m-1)}\pi_{-i}^{\mathrm{ref}}(\boldsymbol{a}_{-i} \mid x)$. Since $\mathcal{Z}_i^2$ is nonnegative, maximizing the preceding bound over a_i gives

$$\begin{aligned}d_i(x;\pi)^2 &\le \max_{a_i} \mathbb{E}_{\pi_{-i}} \mathcal{Z}_i(x,a_i,\boldsymbol{A}_{-i})^2 \\ &\le e^{\eta(m-1)} \max_{a_i} \mathbb{E}_{\pi_{-i}^{\mathrm{ref}}} \mathcal{Z}_i(x,a_i,\boldsymbol{A}_{-i})^2.\end{aligned}$$

Step (iii): Shift a Unilateral Rule to the Logger. The remaining $\max_{a_i}$ is digested by introducing a deterministic deviation. Define $F_i(x,a_i) = \mathbb{E}_{\pi_{-i}^{\mathrm{ref}}} \mathcal{Z}_i(x,a_i,\boldsymbol{A}_{-i})^2$. This is measurable in x for each of the finitely many actions. Let $a_i^*(x)$ be the first maximizer of $F_i(x,\cdot)$ in a fixed ordering and $\pi_i'(a_i \mid x) = \mathbf{1}\{a_i = a_i^*(x)\}$. Then

$$\begin{aligned}\mathbb{E}_\rho \max_{a_i} F_i(x,a_i) &= \mathbb{E}_{x\sim\rho,\boldsymbol{a}\sim\pi_i'\pi_{-i}^{\mathrm{ref}}} \mathcal{Z}_i(x,\boldsymbol{a})^2 \\ &= \int \mathcal{Z}_i^2 \, d(\rho\,\pi_i'\pi_{-i}^{\mathrm{ref}}) = \int \mathcal{Z}_i^2 \frac{d(\rho\,\pi_i'\pi_{-i}^{\mathrm{ref}})}{d\mu}\, d\mu \le C_{\mathrm{uni}} \mathbb{E}_\mu \mathcal{Z}_i^2.\end{aligned}$$

The density ratios in the coverage assumptions denote Radon–Nikodym derivatives, so their existence gives the density equality. The rule π_i' need not have finite integrated KL. To justify applying Assumption 4.4, set $k_i(x) = \mathrm{KL}(\pi_i'(\cdot \mid x)\|\pi_i^{\mathrm{ref}}(\cdot \mid x))$. Full support and finite action sets give $k_i(x) < \infty$ for ρ-almost every x. For $L \ge 1$, let $\pi_i'^{,L}$ equal π_i' on $\{x : k_i(x) \le L\}$ and π_i^{ref} elsewhere. This is a measurable policy with $K_i(\pi_i'^{,L}) \le L$, so $\pi_i'^{,L} \in \Pi_i$. For every measurable context–action event A, the coverage assumption gives

$$(\rho\,\pi_i'^{,L}\pi_{-i}^{\mathrm{ref}})(A) \le C_{\mathrm{uni}}\mu(A). \tag{25}$$

Dominated convergence as $L \to \infty$ gives the same inequality for $\rho\,\pi_i'\pi_{-i}^{\mathrm{ref}}$. Thus the density ratio is bounded by the same C_{uni}, even when $K_i(\pi_i') = \infty$, justifying the preceding change of measure. Replacing π_{-i}^{ref} by π_{-i}^* in this argument gives the same extension with constant C_{uni}^* for (2).

Combining the three steps proves $\mathbb{E}_\rho d_i(x;\pi)^2 \le e^{\eta(m-1)} C_{\mathrm{uni}}\sigma_i^2 = C_{\mathrm{shift}} C_{\mathrm{uni}}\sigma_i^2$. The construction is valid even when the functions depend on the observed data. This pointwise argument covers every data-dependent iterate on the same regression event; no union bound over T and no policy/data independence are needed. □

B.3.4 Fast-Rate Generalization for Least Squares

The population squared regression error is controlled by the standard fast-rate guarantee for empirical risk minimization with bounded outputs.

Replacing r_i by $\widehat{Q}_i$ gives the empirical values $\widehat{V}_i^{\pi}$ and total gap $\widehat{\mathrm{Gap}}_{\mathrm{NE}}(\pi)$. Let $\sigma_i^2 = \mathbb{E}_\mu \mathcal{Z}_i^2$ and $\mathcal{E} = \sum_i \sigma_i^2$. These are expected squared regression errors under μ. For $0 < \delta < 1$, define the following upper bound on the total squared reward-estimation error:

$$\mathcal{E}_n(\delta) := \frac{30}{n} \sum_{i=1}^{m} \log \frac{2m|\mathcal{Q}_i|}{\delta}.$$

By Lemma B.10, $\mathcal{E} \le \mathcal{E}_n(\delta)$ with probability at least $1 - \delta$.

Lemma B.10 (Bounded realizable least squares)**.** *Under the offline learning model in Section 3.5, simultaneously for all i, with probability at least $1 - \delta$,*

$$\sigma_i^2 \le \frac{30}{n} \log \frac{2m|\mathcal{Q}_i|}{\delta}. \tag{26}$$

Proof. Fix a player and a deterministic $f \in \mathcal{Q}_i$. For one row $(x_\tau, \boldsymbol{a}_\tau, \mathbf{r}_\tau)$, abbreviate $X = x_\tau$ and $\boldsymbol{A} = \boldsymbol{a}_\tau$, and set $D = (f(X, \boldsymbol{A}) - r_{\tau,i})^2 - (r_i(X, \boldsymbol{A}) - r_{\tau,i})^2$ and $v = \mathbb{E}_\mu (f - r_i)^2$. The conditional mean model gives

$$\begin{aligned} D &= (f - r_i)^2 + 2(f - r_i)(r_i - r_{\tau,i}), \\ \mathbb{E}[D \mid X, \boldsymbol{A}] &= (f(X, \boldsymbol{A}) - r_i(X, \boldsymbol{A}))^2, \\ \mathbb{E}D &= v. \end{aligned}$$

Thus excess population squared loss equals prediction risk, although excess empirical loss can be negative. The function f is fixed in this calculation; we take a union bound before selecting the fitted function. Since $D = (f - r_i)(f + r_i - 2r_{\tau,i})$, $|D| \le 1$ and $\mathbb{E}D^2 \le 4v$. Also $|D - v| \le 2$.

Concentration of a Fixed Candidate. The identity below bounds the second moment, using $f, r_i, r_{\tau,i} \in [0, 1]$:

$$D^2 = (f - r_i)^2 (f + r_i - 2r_{\tau,i})^2 \le 4(f - r_i)^2, \qquad \mathrm{Var}(D) \le \mathbb{E}D^2 \le 4v.$$

For independent centered $W_\tau = v - D_\tau$, therefore, $|W_\tau| \le 2$ and $\sum_\tau \mathbb{E}W_\tau^2 \le 4nv$. The event of nonpositive empirical excess loss is $\sum_\tau W_\tau \ge nv$. For $v > 0$, the one-sided Bernstein inequality for independent centered variables, with variance sum at most $4nv$ and absolute bound 2, gives

$$\Pr\left\{ \frac{1}{n} \sum_{\tau=1}^{n} D_\tau \le 0 \right\} \le \exp\left\{ -\frac{nv^2}{8v + (4/3)v} \right\} = e^{-3nv/28}.$$

Uniform Selection of the Least-Squares Fit. For any $a > 0$, a union bound over the *fixed* class gives

$$\Pr\{\exists f \in \mathcal{Q}_i : \mathbb{E}_\mu (f - r_i)^2 > a,\ \sum_\tau D_\tau(f) \le 0\} \le |\mathcal{Q}_i| e^{-3na/28}.$$

Set $a = (28/(3n)) \log(m|\mathcal{Q}_i|/\delta)$. ERM and realizability imply the empirical excess loss of $\widehat{Q}_i$ is nonpositive. Union bounding over the fixed class excludes every selected function with $v > (28/3n) \log(m|\mathcal{Q}_i|/\delta)$, except on an event of probability δ/m. The zero-risk case is immediate. Sum these failure probabilities over players. Since $28/3 < 30$ and $\log(m|\mathcal{Q}_i|/\delta) < \log(2m|\mathcal{Q}_i|/\delta)$, (26) follows. No independence between players' rewards is used. □

B.3.5 Closing the Bound: Idealized ROPE Rate

Let $\hat{\pi}$ denote the policy returned by Algorithm 1. Because $\hat{\pi}$ is an exact regularized NE of the empirical game, every $\hat{\pi}_i$ is the empirical regularized best response to $\hat{\pi}_{-i}$, so the empirical equilibrium gap and the linear mismatch both vanish identically:

$$\varepsilon_{\mathrm{pot},i}(x) = \varepsilon_{\mathrm{pot},i}(x;\widehat{\pi}) = 0, \qquad \widehat{\pi}_i(\cdot \mid x) = \widehat{\pi}_i^{\dagger}(\cdot \mid x) \quad (\rho\text{-almost everywhere}). \tag{27}$$

Fix a dataset in the event of Lemma B.10. Summing the per-iterate decomposition (4) over players, integrating over $x \sim \rho$, and using (27),

$$\begin{aligned}
\mathrm{Gap}_{\mathrm{NE}}(\hat{\pi}) &= \sum_{i=1}^{m} \mathbb{E}_{x\sim\rho}\big[V_i^{\dagger,\hat{\pi}_{-i}}(x) - V_i^{\hat{\pi}}(x)\big] \\
&= \sum_{i=1}^{m} \mathbb{E}_{x\sim\rho}\big[\varepsilon_{\mathrm{pot},i}(x) + \delta_{\mathrm{BR},i}(x) + \delta_{\mathrm{Iter},i}(x)\big] \\
&\overset{(27)}{=} \sum_{i=1}^{m} \mathbb{E}_{x\sim\rho}\big[\delta_{\mathrm{BR},i}(x) + \delta_{\mathrm{Iter},i}(x)\big].
\end{aligned}$$

Apply the bias-cancellation identity (23) term by term. Since the linear mismatch is zero in the ROPE setting,

$$\begin{aligned}
\sum_{i=1}^{m} \mathbb{E}_{x\sim\rho}\big[\delta_{\mathrm{BR},i}(x) + \delta_{\mathrm{Iter},i}(x)\big] &\le \sum_{i=1}^{m} \Big\{ \underbrace{\mathbb{E}_{x\sim\rho}[\text{linear mismatch}_i]}_{=0 \text{ since } \hat{\pi}_i = \hat{\pi}_i^{\dagger}} \\
&\qquad + \tfrac{\eta}{2}\,\mathbb{E}_{x\sim\rho}\Big[\big\|Q_i^{\hat{\pi}_{-i}}(x,\cdot) - \hat{Q}_i(x,\cdot,\hat{\pi}_{-i})\big\|_{\infty}^2\Big]\Big\} \\
&= \tfrac{\eta}{2}\sum_{i=1}^{m} \mathbb{E}_{x\sim\rho}\Big[\big\|Q_i^{\hat{\pi}_{-i}}(x,\cdot) - \hat{Q}_i(x,\cdot,\hat{\pi}_{-i})\big\|_{\infty}^2\Big].
\end{aligned}$$

Apply the concentrability-shift bound (24) and the per-player concentration bound proved in Lemma B.10, with failure probability δ_i, to each summand:

$$\begin{aligned}
\tfrac{\eta}{2}\,\mathbb{E}_{x\sim\rho}\Big[\big\|Q_i^{\hat{\pi}_{-i}}(x,\cdot) - \hat{Q}_i(x,\cdot,\hat{\pi}_{-i})\big\|_{\infty}^2\Big] &\overset{(24)}{\le} \tfrac{\eta\, C_{\mathrm{shift}}\, C_{\mathrm{uni}}}{2}\,\mathbb{E}_{(x,\boldsymbol{a})\sim\mu}\big[\mathcal{Z}_i(x,\boldsymbol{a})^2\big] \\
&\le \frac{15\,\eta\, C_{\mathrm{shift}}\, C_{\mathrm{uni}}\, \log(2|\mathcal{Q}_i|/\delta_i)}{n},
\end{aligned}$$

each holding with probability $\ge 1-\delta_i$. Setting $\delta_i := \delta/m$ and applying a union bound across $i \in [m]$, with probability at least $1-\delta$,

$$\mathrm{Gap}_{\mathrm{NE}}(\hat{\pi}) \;\le\; \sum_{i=1}^{m} \frac{15\,\eta\, C_{\mathrm{shift}}\, C_{\mathrm{uni}}\, \log(2m|\mathcal{Q}_i|/\delta)}{n} \;\le\; \frac{15\,m\,\eta\, C_{\mathrm{shift}}\, C_{\mathrm{uni}}\, \log(2m|\mathcal{Q}|/\delta)}{n},$$

where $|\mathcal{Q}| := \max_i |\mathcal{Q}_i|$. Substituting $C_{\mathrm{shift}} = \exp(\eta(m-1))$ and absorbing $\log(2m/\delta)$ into the $\widetilde{\mathcal{O}}$ notation,

$$\mathrm{Gap}_{\mathrm{NE}}(\hat{\pi}) \;\le\; \widetilde{\mathcal{O}}\left(\frac{m\,\eta\, C_{\mathrm{uni}}\, \exp(\eta(m-1))\, \log|\mathcal{Q}|}{n}\right),$$

The sum form before replacing individual class sizes by $|\mathcal{Q}|$ is exactly $\frac{\eta}{2} C_{\mathrm{shift}} C_{\mathrm{uni}} \mathcal{E}_n(\delta)$, proving Theorem 5.1. This argument uses the empirical best-response property and does not require a potential for the fitted rewards.

Step 3: Allow an Approximate Empirical Oracle. If $\widehat{\mathrm{Gap}}_{\mathrm{NE}}(\pi) \le \epsilon_{\mathrm{emp}}$ and the opponent envelope holds, the general form of the transfer lemma instead gives

$$\mathrm{Gap}_{\mathrm{NE}}(\pi) \le 2\widehat{\mathrm{Gap}}_{\mathrm{NE}}(\pi) + \eta \sum_i \mathbb{E}_\rho d_i(x;\pi)^2 \le 2\epsilon_{\mathrm{emp}} + \eta C_{\mathrm{shift}} C_{\mathrm{uni}} \mathcal{E}_n(\delta).$$

The extra factor two is an explicit allowance for the residual empirical gap. An arbitrary numerical policy need not satisfy the envelope; verifying a small empirical gap alone does not remove that hypothesis.

B.3.6 Closing the Bound: Simultaneous OPMD Rate

We now turn to the iterate $\tilde{\pi} = \pi^{(t^*)}$ with $t^* \sim \mathrm{Unif}([T])$. We average the per-iterate decomposition (4) over t^* and over players.

Step 1: Sum the Potential Progress Inequality. On the regression event, $D_t \le C_{\mathrm{shift}} C_{\mathrm{uni}} \mathcal{E} \le C_{\mathrm{shift}} C_{\mathrm{uni}} \mathcal{E}_n(\delta)$ for every t. Rearranging and summing (6) gives

$$\begin{aligned}\frac{1}{\eta\beta}\sum_{t=1}^{T} X_t &\le \sum_{t=1}^{T}[\Phi(\pi^{(t+1)}) - \Phi(\pi^{(t)})] + \gamma \sum_{t=1}^{T} D_t + Tm\alpha \\ &= \Phi(\pi^{(T+1)}) - \Phi(\pi^{\mathrm{ref}}) + \gamma \sum_{t=1}^{T} D_t + Tm\alpha \\ &\le m(1+\alpha) + \gamma T C_{\mathrm{shift}} C_{\mathrm{uni}} \mathcal{E} + Tm\alpha.\end{aligned}$$

Here the last line uses the one-sided endpoint bound (11). Multiply by $\eta\beta/T$ and insert the result into Lemma B.6 to obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} \widehat{G}_t \le{}& \frac{\eta m}{\beta(2-\beta)T} \\ &+ \frac{\eta^2(m-1)}{2\beta}\left[\frac{m(1+\alpha)}{T} + m\alpha + \gamma C_{\mathrm{shift}} C_{\mathrm{uni}} \mathcal{E}\right].\end{aligned} \tag{28}$$

The first term is the initial score-tracking error, rather than a potential endpoint term. Retaining it is essential when $m = 1$, because opponent motion and the second term then vanish.

Step 2: Transfer the Average Gap to the True Game. For each iterate, Lemma B.8 gives $\mathrm{Gap}_{\mathrm{NE}}(\pi^{(t)}) \le 2\widehat{G}_t + \eta D_t$. Integrating and summing Lemma B.8, uniformly at each iterate, gives

$$\frac{1}{T}\sum_{t=1}^{T} \mathrm{Gap}_{\mathrm{NE}}(\pi^{(t)}) \le \frac{2}{T}\sum_{t=1}^{T} \widehat{G}_t + \eta C_{\mathrm{shift}} C_{\mathrm{uni}} \mathcal{E}.$$

Step 3: Identify the Output Randomness and Constants. For a fixed dataset $\mathcal{D}$, the sequence is determined by the specified fits and updates. A single shared index, independent of $\mathcal{D}$, satisfies

$$\mathbb{E}_{t^*}[\mathrm{Gap}_{\mathrm{NE}}(\pi^{(t^*)}) \mid \mathcal{D}] = \sum_{t=1}^{T} \frac{1}{T} \mathrm{Gap}_{\mathrm{NE}}(\pi^{(t)}).$$

This is the gap of a randomly chosen product profile, averaged over the choice. It is not the gap of $\frac{1}{T}\sum_t \pi^{(t)}$ or of the product of componentwise average policies. The left side is exactly the conditional expectation over a uniformly sampled shared index. Inserting (28), then $\mathcal{E} \le \mathcal{E}_n(\delta)$, proves Theorem 6.1 with $B_{\eta,\gamma,m} = \eta^2(m-1)/\beta$.

Explicitly, the coefficient multiplying the bracket in (28) after transfer is $2[\eta^2(m-1)/(2\beta)] = B_{\eta,\gamma,m}$. Thus

$$\begin{aligned}\mathbb{E}_{t^*}[\mathrm{Gap}_{\mathrm{NE}}(\pi^{(t^*)}) \mid \mathcal{D}] &\le \frac{2\eta m}{\beta(2-\beta)T} + B_{\eta,\gamma,m}\left[\frac{m(1+\alpha)}{T} + m\alpha\right] \\ &\quad + (B_{\eta,\gamma,m}\gamma + \eta) C_{\mathrm{shift}} C_{\mathrm{uni}} \mathcal{E} \\ &\le \frac{2\eta m}{\beta(2-\beta)T} + B_{\eta,\gamma,m}\left[\frac{m(1+\alpha)}{T} + m\alpha\right] \\ &\quad + (B_{\eta,\gamma,m}\gamma + \eta) C_{\mathrm{shift}} C_{\mathrm{uni}} \mathcal{E}_n(\delta).\end{aligned}$$

For the compact scaling display in Theorem 6.1, use $T \ge 1$, $m \ge 1$, $0 < \beta < 1$, and $\gamma \le 1$ to obtain

$$\begin{aligned}\frac{2\eta m}{\beta(2-\beta)T} + \frac{B_{\eta,\gamma,m} m}{T} &\le \frac{(2\eta + \eta^2)m^2}{\beta T}, \\ B_{\eta,\gamma,m} m\alpha(1 + T^{-1}) &\le \frac{2\eta^2 m^2 \alpha}{\beta}, \\ m(B_{\eta,\gamma,m}\gamma + \eta) &\le (\eta + \eta^2 + \eta^3)m^2.\end{aligned}$$

Combining these inequalities with $\mathcal{E}_n(\delta) \le 30m \log(2m|\mathcal{Q}|/\delta)/n$ gives the displayed three terms. Thus the hidden η factors can be taken as $2\eta + \eta^2$, $2\eta^2$, and $30(\eta + \eta^2 + \eta^3)$, respectively. The $\widetilde{\mathcal{O}}$ notation in Theorem 6.1 treats η as fixed.

The dataset event has probability at least $1 - \delta$ and supports every integer $T \ge 1$. For $\alpha = 0$, the structural term vanishes.

For $\gamma = \min\{\eta, 1/\max\{1, m-1\}\}$,

$$B_{\eta,\gamma,m}\gamma + \eta = \eta^2(m-1)(\eta+\gamma) + \eta \le 2\eta^3(m-1) + \eta.$$

When $\gamma = \eta$, $\beta = 1/2$ and $B_{\eta,\gamma,m} = 2\eta^2(m-1)$. The optimization and structural coefficients are polynomial in η and $1/\gamma$; $C_{\text{shift}}C_{\text{uni}} = C_{\text{uni}}e^{\eta(m-1)}$ remains in the statistical term. Choosing T proportional to n balances the optimization and statistical terms after accounting for their coefficients.

Computational Model. For S finite contexts and $K = \prod_i |\mathcal{A}_i|$ joint actions, dense reward storage costs $O(mSK)$ and exact simultaneous marginalization costs $O(mSK)$ per sweep. Local policy updates use these marginalized rewards and the opponents' current distributions. ERM and policy representation incur additional costs.

B.4 Regularized and Unregularized Equilibria

Our learning target is the KL-regularized game. As a supplementary comparison, we bound the unregularized Nash gap of the same policy. Let Gap_0 denote total exploitability under the unregularized rewards. If $B_i \coloneqq \mathbb{E}_\rho \log \frac{1}{\min_{a_i} \pi_i^{\text{ref}}(a_i|x)} < \infty$ for every player, then

$$\text{Gap}_0(\pi) \le \text{Gap}_{\text{NE}}(\pi) + \eta^{-1} \sum_i B_i.$$

For uniform references, $B_i = \log|\mathcal{A}_i|$. The bound follows from $U_i(\pi) = J_i(\pi) + \eta^{-1}K_i(\pi_i)$ and $0 \le K_i(\pi_i') \le B_i$ for every deviation. Thus the additional unregularized gap is controlled by the explicit term $\eta^{-1}\sum_i B_i$.

For example, in a one-player game with rewards $(0, 1)$ and a uniform reference, the regularized optimum chooses the rewarding action with probability $e^\eta/(1+e^\eta)$. Its regularized gap is zero, whereas its unregularized gap is $1/(1+e^\eta)$.

C Experimental Protocol and Reproducibility

All experiments use three players, two contexts with probabilities $(0.4, 0.6)$, binary actions, and independent Bernoulli observations conditional on context and joint action. The logger and reference are the same fixed nonuniform product policy. Population metrics use the true reward tables, with no evaluation sampling. Numerical checks are distinct from the offline-data experiment.

C.1 Games and controlled comparisons

Coordination. For $\boldsymbol{a} \in \{0,1\}^3$, define the common potential

$$\phi(x, \boldsymbol{a}) = 0.08 + 0.80\mathbf{1}\{a_1 = a_2 = a_3\} + \sum_{i=1}^{3} b_{x,i}a_i,$$

$$b_0 = (0.03, 0.02, 0.01), \qquad b_1 = (0.01, 0.02, 0.03).$$

The exact-potential game has $r_i = \phi$ for every player. The perturbed family is

$$r_i(x, \boldsymbol{a}) = \phi(x, \boldsymbol{a}) + s_i\delta(-1)^{a_1+a_2+a_3}, \qquad s = (1, -1, 1).$$

A unilateral switch changes the perturbation by at most 2δ. Thus the common potential, less the sum of the players' KL penalties, certifies $\alpha \le 2\delta$ for the regularized game. This is an upper bound, not a claim of a minimal potential tolerance. All rewards lie in $[0, 1]$ and on the fixed 0.001 grid for the selected perturbations.

Congestion. Each player chooses one of two resources. Let $N_b(\boldsymbol{a}) = \sum_i \mathbf{1}\{a_i = b\}$ be the load on resource b. In context 0, the utilities at loads 1, 2, and 3 are

$$(u_{0,0}(1), u_{0,0}(2), u_{0,0}(3)) = (0.90, 0.65, 0.35),$$
$$(u_{0,1}(1), u_{0,1}(2), u_{0,1}(3)) = (0.85, 0.60, 0.30).$$

Context 1 swaps the resource tables, $u_{1,b} = u_{0,1-b}$. The base reward is $u_{x,a_i}(N_{a_i}(\boldsymbol{a}))$, with exact potential

$$\phi_{\mathrm{cong}}(x, \boldsymbol{a}) = \sum_{b=0}^{1} \sum_{k=1}^{N_b(\boldsymbol{a})} u_{x,b}(k).$$

If a player switches from resource b to c, the potential changes by $u_{x,c}(N_c + 1) - u_{x,b}(N_b)$, exactly its reward change. The near-potential rewards are

$$r_i(x, \boldsymbol{a}) = u_{x,a_i}(N_{a_i}(\boldsymbol{a})) + s_i \delta (-1)^{a_1+a_2+a_3}, \qquad s = (1, -1, 1), \quad \delta = 0.02.$$

The same unilateral perturbation argument certifies $\alpha \le 0.04$ after subtracting the KL penalties from the potential. The rewards lie in $[0.28, 0.92]$ and on the fixed grid. Unlike coordination, these base rewards decrease when more players choose the same resource.

Shared protocol. The action-zero reference probabilities are $(0.60, 0.55, 0.52)$ in context 0 and $(0.52, 0.55, 0.60)$ in context 1. The logger draws contexts from ρ and actions independently from these reference policies. We fix $\eta = 4$, $\gamma = 1/2$, and $\beta = 1/9$. Consequently the target and coverage remain fixed as n varies. For this logger, the reference-anchored coverage coefficient is $C_{\mathrm{uni}} = 1/0.4 = 2.5$.

The coordination statistical experiment uses $\delta = 0$ and 0.02, with

$$n \in \{128, 181, 256, 362, 512, 724, 1024, 1448, 2048, 2896, 4096, 5792, 8192\}$$

and 32 independent datasets per setting, for $2 \cdot 13 \cdot 32 = 832$ datasets. The nine additional sample sizes and their seeds were fixed before the corresponding datasets were generated. The congestion game and evaluation protocol were fixed before any congestion datasets were generated. The near-potential congestion experiment adds $13 \cdot 32 = 416$ independent datasets under the same sample-size and estimation protocol, giving 1248 datasets in total. The noiseless coordination optimization sweep evaluates both variants at $T \in \{1, 4, 16, 64, 256, 1024, 4096, 16384, 65536, T_0\}$.

C.2 Fixed optimization horizon

Before generating the offline datasets for each family, we selected $T_0 = 2^{23}$ using population-known controls, requiring the time-averaged regularized Nash gap to be below 10^{-7} for all included games. We held this common horizon fixed across all sample sizes, so increasing n does not increase the optimization effort.

C.3 Estimator, algorithms, and numerical certificates

Each fixed class contains all tables on the $2 \cdot 2^3 = 16$ context–action cells with entries in $\{j/1000 : 0 \le j \le 1000\}$, hence $|\mathcal{Q}_i| = 1001^{16}$. The true rewards lie on this grid. Squared-loss ERM separates across cells: round each observed cell mean to the nearest grid point, with deterministic ties, and use 0.5 in an unobserved cell. The independent fits need not form a potential game. For each context, each pair of players i, j, and each fixed action a_k of the third player, we calculate the face cross-difference

$$\begin{aligned} D_{ij}\widehat{Q}_\ell(x, a_k) = \widehat{Q}_\ell(x, a_i = 1, a_j = 1, a_k) - \widehat{Q}_\ell(x, a_i = 1, a_j = 0, a_k) \\ - \widehat{Q}_\ell(x, a_i = 0, a_j = 1, a_k) + \widehat{Q}_\ell(x, a_i = 0, a_j = 0, a_k). \end{aligned}$$

Exact potential consistency requires $D_{ij}\widehat{Q}_i(x, a_k) = D_{ij}\widehat{Q}_j(x, a_k)$ on every face. We save all $2 \cdot \binom{3}{2} \cdot 2 = 12$ mismatches and their largest absolute value. A mismatch above 10^{-12} certifies a violation numerically. All 416 exact-coordination fits violate at least one face condition, even though the true game has zero face mismatches. For the perturbed true games, the maximum face mismatch is 8δ, as also checked by direct enumeration. (The 8δ face mismatch is an exact-potential consistency diagnostic and is distinct from the unilateral potential-discrepancy bound $\alpha \le 2\delta$.)

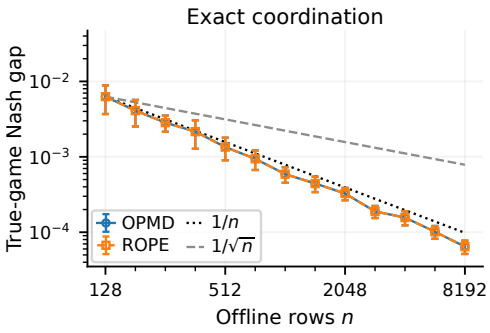


Figure 2: Exact-potential coordination control. Population regularized Nash gaps with means and 95% Student intervals across 32 independent datasets per sample size, using the same T_0 as the main figure. Both rate guides are anchored at the OPMD mean at $n = 128$. Gaps average all pre-update iterates with no discarded burn-in.

All OPMD runs use the bounded-score recursion in Lemma B.4, initialized at the reference. The population regularized Nash gap of the shared uniform-index output is evaluated by averaging the T pre-update gaps. OPMD uses the prescribed horizon T_0. If the floating-point state becomes exactly unchanged, identical future iterates are counted algebraically when computing the time average; no tolerance-based early stopping or burn-in is used.

Numerical certification. The empirical-equilibrium baseline solves the binary logit fixed-point equations numerically. We accept a solution only if its empirical regularized Nash gap is at most 10^{-10} and its fixed-point residual is at most 10^{-8}. All reported solutions satisfy these criteria. Full solver settings and numerical validation tests are provided with the code.

C.4 Exact-potential coordination control

Figure 2 shows the exact-potential coordination comparison. Mean OPMD Nash gaps decrease from 6.28×10^{-3} at $n = 128$ to 6.45×10^{-5} at $n = 8192$. Its descriptive log–log slope is -1.085, with bootstrap 95% interval $[-1.139, -1.018]$.

C.5 Optimization-only ablation

To separate optimization error from statistical error, we additionally give OPMD the true coordination reward tables and vary T (Figure 3). This controlled ablation removes offline reward estimation. With true rewards, the same OPMD updates achieve regularized uniform-index gaps of 7.68×10^{-8} and 7.88×10^{-8} at T_0. At the largest sample size, these controls are more than 800 times smaller than the fitted-data gaps.

Noncontraction. For the exact and near-potential coordination games, interior witnesses give best-response Jacobian spectral radii 1.600 and 1.597, respectively. The OPMD logit Jacobian has spectral radius $16/15 > 1$ in both games. These witnesses exclude global contraction as an explanation of convergence, while allowing local convergence. The ablation does not establish a causal advantage of potential structure. Analytical Jacobians and independent numerical checks are provided with the code.

C.6 Robustness to larger potential deviations

As an additional ablation, Table 1 reports the known-reward results for the coordination family with $\delta \in \{0, 0.01, 0.02, 0.04\}$, giving certified bounds $\alpha \leq 2\delta$. OPMD reaches small regularized Nash gaps in each tested instance. This is a robustness check for this family, not evidence that the worst-case $O(\alpha)$ term is tight or that arbitrary large-α games are easy.

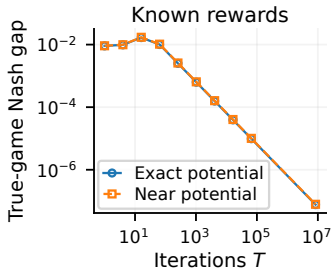


Figure 3: Optimization-only ablation with known rewards. OPMD uses the true reward tables in the exact-potential and near-potential coordination games. The population regularized Nash gap is averaged over the T pre-update iterates, with no discarded burn-in. These deterministic runs have no sampling error bars.

δ	Certified bound on α	Regularized Nash gap
0.00	$\alpha = 0$	7.68×10^{-8}
0.01	$\alpha \leq 0.02$	7.78×10^{-8}
0.02	$\alpha \leq 0.04$	7.88×10^{-8}
0.04	$\alpha \leq 0.08$	8.05×10^{-8}

Table 1: Known-reward OPMD at larger tested perturbations. All runs use $\eta = 4$ and $T_0 = 2^{23}$; the population regularized Nash gap averages all pre-update iterates. These deterministic controls have no sampling error bars.

C.7 Uncertainty, saved results, and reproduction

For each statistical mean we report a 95% Student interval across 32 independent datasets. Descriptive log–log slopes are obtained by regressing log mean gap on log sample size. Their 95% intervals use 2000 bootstrap resamples within each sample size. The near-potential coordination and congestion OPMD slopes are -1.113 and -1.071, with 95% bootstrap intervals $[-1.168, -1.052]$ and $[-1.113, -1.021]$, respectively.

These finite-range slopes do not prove asymptotic rates, and the fixed grid eventually changes the estimation regime. The optimization sweep uses known rewards and has no sampling error bars.

Reproducibility. We use 32 independent datasets at every sample size, with seeds fixed before sampling each batch.

C.8 Comparison with the unregularized objective

Our theoretical and algorithmic target throughout is the KL-regularized game. The KL penalty is part of each player's payoff, and we do not aim to learn an equilibrium of the unregularized game. For methods targeting unregularized equilibria, see Cui & Du (2022a;b). As a supplementary illustration, Table 2 evaluates the same learned policies under both payoff objectives at $n = 8192$. The regularized Nash gap measures error on our learning objective. For OPMD, both metrics average over the same T_0 pre-update iterates.

Game	Algorithm	Regularized	Unregularized
Exact coordination	OPMD	0.0645 [0.0515, 0.0774]	0.0954 [0.0944, 0.0965]
Exact coordination	ROPE oracle	0.0644 [0.0515, 0.0773]	0.0954 [0.0944, 0.0965]
Near coordination	OPMD	0.0645 [0.0502, 0.0787]	0.0928 [0.0918, 0.0939]
Near coordination	ROPE oracle	0.0644 [0.0501, 0.0786]	0.0928 [0.0918, 0.0939]
Near congestion	OPMD	0.3261 [0.2603, 0.3918]	0.0632 [0.0614, 0.0650]
Near congestion	ROPE oracle	0.3261 [0.2603, 0.3918]	0.0632 [0.0614, 0.0650]

Table 2: Same-policy objective comparison at $n = 8192$: mean [95% Student interval] across 32 independent datasets. Both columns report population Nash gaps. Regularized gaps are in units of 10^{-3}; unregularized gaps are unscaled.