Offline Learning in Markov Games with General Function Approximation

Yuheng Zhang [1] **Yu Bai** [2] **Nan Jiang** [1]

Abstract

We study offline multi-agent reinforcement learning (RL) in Markov games, where the goal is to learn an approximate equilibrium—such as Nash equilibrium and (Coarse) Correlated Equilibrium—from an offline dataset pre-collected from the game. Existing works consider relatively restricted tabular or linear models and handle each equilibria separately. In this work, we provide the first framework for sample-efficient offline learning in Markov games under general function approximation, handling all 3 equilibria in a unified manner. By using Bellman-consistent pessimism, we obtain interval estimation for policies' returns, and use both the upper and the lower bounds to obtain a relaxation on the gap of a candidate policy, which becomes our optimization objective. Our results generalize prior works and provide several additional insights. Importantly, we require a data coverage condition that improves over the recently proposed "unilateral concentrability". Our condition allows selective coverage of deviation policies that optimally trade-off between their greediness (as approximate best responses) and coverage, and we show scenarios where this leads to significantly better guarantees. As a new connection, we also show how our algorithmic framework can subsume seemingly different solution concepts designed for the special case of two-player zero-sum games.

1. Introduction

Offline RL aims to learn a good policy from a pre-collected historical dataset. It has emerged as an important paradigm for bringing RL to real-life scenarios due to its non-interative nature, especially in applications where deploying adaptive algorithms in the real system is financially costly and/or ethically problematic (Levine et al., 2020). While offline RL has been extensively studied in the single-agent setting, many real-world applications involve the strategic interactions between multiple agents. This renders the necessity of bringing in game-theoretic reasoning, often modeled using *Markov games* (Shapley, 1953) in the RL theory literature. Markov games can be viewed as the multi-agent extension of Markov Decision Processes (MDPs), where agents share the same state information and the dynamics is determined by the joint action of all agents.

While online RL in Markov games has seen significant developments in recent years (Bai & Jin, 2020; Liu et al., 2021; Song et al., 2022; Jin et al., 2021b), offline learning in Markov games has only started to attract attention from the community. Earlier works (Cui & Du, 2022b; Zhong et al., 2022) focus on tabular cases or linear function approximation, which cannot handle complex environments that require advanced function-approximation techniques. Although there has been a rich literature on single-agent RL with general function approximation (Jiang et al., 2017; Jin et al., 2021a; Wang et al., 2020; Huang et al., 2021), whether and how they can be extended to offline Markov games remains largely unclear. In addition, the learning goal in Markov games is no longer return optimization, but instead finding an *equilibrium*. However, there are multiple popular notions of equilibria, and prior results for the offline setting mainly focus on one of them (Nash) (Cui & Du, 2022a;b; Zhong et al., 2022). These considerations motivate us to study the following question:

Can we design sample-efficient algorithms for offline Markov games with general function approximation, and handle different equilibria in a unified framework?

Unified framework In this paper, we provide information-theoretic results that answer the question in the positive. We first express the equilibrium gap—the objective we wish to minimize—in a unified manner for 3 popular notions of equilibria: Nash Equilibrium (NE), Correlated Equilibrium (CE), and Coarse Correlated Equilibrium (CCE) (Section 3). Then, we build on top of the Bellman-consistent pessimism framework from single-agent offline RL (Xie et al., 2021a), which allows us to construct confidence sets for policy evaluation and obtain the confidence intervals of policies' returns. An important difference is that Xie et al. (2021a) only needs

[1]University of Illinois at Urbana-Champaign [2] Salesforce AI Research. Correspondence to: Nan Jiang <nanjiang@illinois.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

pessimistic evaluations in the single-agent case; in contrast, we need both *optimistic* and *pessimistic* evaluations to further compute a surrogate upper bound on the *equilibrium gap* of each candidate policy, which provably leads to strong offline learning guarantees (Section 4).

New insights on data conditions Our algorithm and analyses also shed new light on the offline learnability of Markov games. In single-agent offline RL, it is understood that a good policy can be learned as long as the data covers one, and this condition is generally known as "single-policy concentrability/coverage" (Jin et al., 2021c; Zhan et al., 2022). In contrast, in Markov games, data covering an equilibrium is intuitively insufficient, as a fundamental aspect of equilibrium is reasoning about what would happen if other agents were to *deviate*. To address this discrepancy, a notion of "unilateral concentrability" is proposed as a sufficient data condition for offline Markov games (Cui & Du, 2022a) (see also Zhong et al. (2022)), which asserts that the equilibrium as well as its all unilateral deviations are covered. While this is sufficient and in the worst-case necessary, it remains unclear whether less stringent conditions may also suffice. Our work relaxes the assumption and provide more flexible guarantees. Instead of depending on the worst-case estimation error of all unilateral deviation policies, our error bound exhibits the trade-off between a policy coverage error term and a policy suboptimality term. It automatically adapts to the optimal trade-off, and we show scenarios in Section 4.3 where the bound significantly improves over unilateral coverage results (Cui & Du, 2022b).

V-type variant Our main algorithm estimates the policies' Q-functions, which takes all agents' actions as inputs. When specialized to the tabular setting, this would incur an exponentially dependence on the number of agents. While this can be avoided by using strong function approximation to generalize over the joint action space (Zhong et al., 2022), it prevents us from reproducing and subsuming the prior works (Cui & Du, 2022a;b). To address this issue, we propose a V-type variant of our algorithm, which estimates state-value functions instead and uses importance sampling to correct for action mismatches. It naturally avoids the exponential dependence, and reproduces the rate (up to minor differences) of (Cui & Du, 2022b) whose analysis is specialized to tabular settings (Section 5).

New connection for two-player zero-sum games As an additional discovery, we show interesting connection between our work and prior algorithmic ideas (Jin et al., 2022; Cui & Du, 2022b) that are specifically designed for two-player zero-sum games. While they seem very different at the first glance, we show in Appendix B that these ideas can be subsumed by our algorithmic framework and our analyses and guarantees extend straightforwardly.

1.1. Related Work

Offline RL Offline RL aims to learn a good policy from a pre-collected dataset without direct interaction with the environment. There are many prior works studying single-agent offline RL problem in both the tabular (Yin et al., 2021b;a; Yin & Wang, 2021; Rashidinejad et al., 2021; Xie et al., 2021b; Shi et al., 2022; Li et al., 2022) and function approximation setting (Antos et al., 2008; Precup, 2000; Chen & Jiang, 2019; Xie & Jiang, 2020; 2021; Xie et al., 2021a; Jin et al., 2021c; Uehara & Sun, 2022; Zhan et al., 2022). Notably, Xie et al. (2021a) introduces the notion of Bellman-consistent pessimism and our techniques are built on it.

Markov games Markov games is a widely used framework for multi-agent reinforcement learning. Online learning equilibria of Markov games has been extensively studied, including two-player zero-sum Markov games (Wei et al., 2017; Bai & Jin, 2020; Bai et al., 2020; Liu et al., 2021; Dou et al., 2022), and multi-player general-sum Markov games (Liu et al., 2021; Song et al., 2022; Jin et al., 2021b; Mao & Başar, 2022). Three equilibria are usually considered as the learning goal—Nash Equilibrium (NE), Correlated Equilibrium (CE) and Coarse Correlated Equilibrium (CCE). Recently, a line of works consider solving Markov games with function approximation, including linear (Xie et al., 2020; Chen et al., 2022) and general function approximation (Huang et al., 2022; Jin et al., 2022). A closely related work is Jin et al. (2022), where a multi-agent version of the Bellman-Eluder dimension is introduced to solve zero-sum Markov games under general function approximation. However, they focus on the online setting which is different from our offline setting.

Offline Markov games Since Cui & Du (2022a)'s initial work on offline tabular zero-sum Markov games, there have been several follow-up works on offline Markov games, either for tabular zero-sum / general-sum Markov games (Cui & Du, 2022b; Yan et al., 2022) or linear function approximation (Zhong et al., 2022; Xiong et al., 2023). In this work, we study general function approximation for multi-player general-sum Markov games, which is a more general framework. Technically, we differ from these prior works in how we handle uncertainty quantification in policy evaluation, an important technical aspect of offline learning: we use initial state optimism/pessimism for policy evaluation, whereas previous works rely on pre-state pessimism with bonus terms. In addition, previous works require the so-called "unilateral concentrability" assumption of data coverage.[1] Although this assumption is unavoidable for the worst-case, our approach requires a condition that is

[1] Zhong et al. (2022) proposes the notion of "relative uncertainty, which is the linear version of "unilateral concentrability".

never worse (and coincides in the worst-case) and can be significantly better on certain instances.

2. Preliminaries

Notations We use $\Delta(\cdot)$ to denote the probability simplex. We use bold letters to denote vectors such as $\mathbf{a}$ and the jth element of $\mathbf{a}$ is denoted by $\mathbf{a}_j$. We use $-i$ to denote all the players except player i. For a positive integer m, $[m]$ denotes the set $\{1, 2, \cdots, m\}$. $\|f\|_{2,d}^2$ represents $\mathbb{E}_d[f^2]$ and $f(s, \pi)$ stands for $\mathbb{E}_{a\sim\pi(\cdot|s)}[f(s, a)]$. We use $\mathcal{O}(\cdot)$ to hide absolute constants and use $\widetilde{\mathcal{O}}(\cdot)$ to hide logarithmic factors.

2.1. Multi-player General-sum Markov Games

We consider multi-player general-sum Markov games in the infinite-horizon discounted setting. Such a Markov game is specified by $(\mathcal{S}, \mathcal{A} = \prod_{i\in[m]} \mathcal{A}_i, P, r, \gamma, s_0)$, where $\mathcal{S}$ is the state space with $|\mathcal{S}| = S$, $\mathcal{A}_i$ is the action space for player i with $|\mathcal{A}_i| = A_i$, $\mathbf{a} \in \mathcal{A}$ is the joint action taken by all m players, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function and $P(\cdot|s, \mathbf{a})$ describes the probability distribution over the next state when joint action $\mathbf{a}$ is taken at state s, $r = \{r_i\}_{i\in[m]}$ is the collection of reward functions where $r_i : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$ is the deterministic reward function for player i, $\gamma \in [0, 1)$ is the discount factor, and s_0 is the fixed initial state which is without loss of generality.

Product and correlated policies A Markov joint policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ specifies the decision-making strategies of all players and induces a trajectory $s_0, \mathbf{a}_0, \mathbf{r}_0, s_1, \mathbf{a}_1, \mathbf{r}_1, \cdots, s_t, \mathbf{a}_t, \mathbf{r}_t, \cdots$, where $\mathbf{a}_t \sim \pi(\cdot|s_t)$, $\mathbf{r}_{t,i} = r_i(s_t, \mathbf{a}_t)$, and $s_{t+1} \sim P(\cdot|s_t, \mathbf{a}_t)$. For a joint policy π, π_i is the marginalized policy of player i and π_{-i} is the marginalized policy for the remaining players. A joint policy π is a product policy if $\pi = \pi_1 \times \pi_2 \times \cdots \pi_m$ where each player i takes actions independently according to π_i. If π is not a joint policy, sometimes we say π is correlated, and the players need to depend their actions on public randomness.

Value function and occupancy For player i and joint policy π, we define the value function $V_i^\pi(s) := \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t r_i(s_t, \mathbf{a}_t)|s_0 = s]$ and the Q-function $Q_i^\pi(s, \mathbf{a}) := \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t r_i(s_t, \mathbf{a}_t)|s_0 = s, a_0 = \mathbf{a}]$, they are bounded in $[0, V_{\max}]$ where $V_{\max} = R_{\max}/(1-\gamma)$. For each joint policy π, the policy-specific Bellman operator of the ith player $\mathcal{T}_i^\pi : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is defined as

$$(\mathcal{T}_i^\pi f)(s, \mathbf{a}) = r_i(s, \mathbf{a}) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,\mathbf{a})}[f(s', \pi)],$$

and Q_i^π is the unique fixed point of $\mathcal{T}_i^\pi$. Note that once a policy is fixed, the game-theoretic considerations are no longer relevant and the value functions are defined in familiar manners similar to the single-agent setting, with the only difference that each player i has its own value function due to the player-specific reward function r_i. Similar to the single-agent case, we also consider the discounted state-action occupancy $d^\pi(s, \mathbf{a}) \in \Delta(\mathcal{S} \times \mathcal{A})$ which is defined as $d^\pi(s, \mathbf{a}) = (1 - \gamma)\mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t \mathbb{I}[s_t = s, \mathbf{a}_t = \mathbf{a}]]$.

2.2. Offline learning of Markov games

In the offline learning setting, we assume access to a pre-collected dataset and cannot have further interactions with the environment. The offline dataset $\mathcal{D}$ consists of n independent tuples $(s, \mathbf{a}, \mathbf{r}, s')$, which are generated as $(s, \mathbf{a}) \sim d_D$, $\mathbf{r}_i \sim r_i(s, \mathbf{a})$ and $s' \sim P(\cdot|s, \mathbf{a})$ with some data distribution $d_D \in \Delta(\mathcal{S} \times \mathcal{A})$.[2]

Policy class In practical problems with large state spaces, the space of all possible Markov joint policies is prohibitively large and intractable to work with. To address this, we assume we have a pre-specified policy class $\Pi \subset (\mathcal{S} \to \Delta(\mathcal{A}))$, from which we seek a policy that is approximately an equilibrium under a given criterion.[3] Let $\Pi_i = \{\pi_i : \pi \in \Pi\}$ denote the class of induced marginalized policies for player i, and define Π_{-i} similarly.

The extended class As we will see in Section 3, a fundamental aspect of equilibria is the counterfactual reasoning of how other agents would deviate and respond to a given policy. After considering the possible deviation behaviors of player i in response to each policy $\pi \in \Pi$, we arrive at an extended class $\Pi_i^{\text{ext}} \supseteq \Pi$ for player i. The concrete form of Π_i^{ext} will be defined in Section 3 and can depend on the notion of equilibrium under consideration, and for now it suffices to say that Π_i^{ext} is a superset of Π consisting of all policies that player i needs to reason about.

Value-function approximation We use $\mathcal{F}_i \subset (\mathcal{S} \times \mathcal{A} \to [0, V_{\max}])$ to approximate the Q-function Q_i^π for each player i. Following (Xie et al., 2021a), we make two standard

[2]For non-i.i.d. adaptive data we may use martingale concentration inequalities in our analyses. Without further mixing-type assumptions, our analyses extend if we change the d_D (which is a static object) in the definitions such as (1) and (2) to $\widehat{d}_D$, which is the empirical distribution over state-action pairs. The resulting definition of (2), for example, corresponds to quantities like $\widehat{C}(\pi)$ in Cui & Du (2022b, Definition 3) defined for the tabular setting.

[3]We only consider minimizing equilibrium gaps among a class of stationary Markov policies in this paper. See Daskalakis et al. (2022) and the references therein for how they suffice for standard notions of equilibria such as NE and CCE, and Nowak & Raghavan (1992) for the case of CE. Below we also only consider stationary Markov policies as *response policies* for NE/CCE, which is also justified by the fact that once a stationary Markov π_{-i} is fixed, optimizing player i's behavior for best response becomes a single-agent MDP problem.

assumptions on $\mathcal{F}_i$,

Assumption A (Approximate Realizability)**.** For any player $i \in [m]$ and any $\pi \in \Pi_i^{\text{ext}}$, we have

$$\inf_{f \in \mathcal{F}_i} \sup_{\text{admissible } d} \|f - \mathcal{T}_i^\pi f\|_{2,d}^2 \leq \varepsilon_{\mathcal{F}},$$

A data distribution d is admissible if $d \in \{d^{\pi'} : \pi' \in \Pi_i^{\text{ext}}\} \cup d_D$.

For each player i and joint policy π, Assumption A requires that there exists $f \in \mathcal{F}_i$ such that f has small Bellman error under all possible distributions induced from the extended policy class Π_i^{ext} and the data distribution. When $Q_i^\pi \in \mathcal{F}_i$, $\forall \pi \in \Pi_i^{\text{ext}}, i \in [m]$, we have $\varepsilon_{\mathcal{F}} = 0$.

Assumption B (Approximate Completeness)**.** For any player $i \in [m]$ and any $\pi \in \Pi_i^{\text{ext}}$, we have

$$\sup_{f \in \mathcal{F}_i} \inf_{f' \in \mathcal{F}_i} \|f' - \mathcal{T}_i^\pi f\|_{2,d_D}^2 \leq \varepsilon_{\mathcal{F},\mathcal{F}}. \tag{1}$$

Assumption B requires that $\mathcal{F}_i$ is approximately closed under operator $\mathcal{T}_i^\pi$. Both assumptions are direct extensions of their counterparts that are widely used in the single-agent offline RL literature.

Distribution mismatch and data coverage Similar to Xie et al. (2021a), we use the discrepancy of Bellman error under π to measure the distribution mismatch between an arbitrary distribution d and data distribution d_D:

$$\mathscr{C}(d; d_D, \mathcal{F}_i, \pi) := \max_{f \in \mathcal{F}_i} \frac{\|f - \mathcal{T}_i^\pi f\|_{2,d}^2}{\|f - \mathcal{T}_i^\pi f\|_{2,d_D}^2}. \tag{2}$$

We remark that $\mathscr{C}(d; d_D, \mathcal{F}_i, \pi) \leq \sup_{s,\mathbf{a}} \frac{d(s,\mathbf{a})}{d_D(s,\mathbf{a})}$, which implies that $\mathscr{C}(d; d_D, \mathcal{F}_i, \pi)$ is a tighter measurement than the raw density ratio.

3. Equilibria

We consider three common equilibria in game theory: Nash Equilibrium (NE), Correlated Equilibrium (CE) and Coarse Correlated Equilibrium (CCE). We define the three equilibria in a unified fashion using the concept of *response class mappings*, so that each equilibrium is defined with respect to the relative best response within each corresponding response class.

A response class mapping $\Pi^\dagger(\cdot)$ maps a policy π to a policy *class*, $\Pi^\dagger(\pi) := \bigcup_{i \in [m]} \Pi_i^\dagger(\pi)$. Roughly speaking, $\Pi_i^\dagger(\pi)$ is obtained by taking a candidate policy π, considering various ways that player i would deviate its behavior from π_i to $\pi_i^\dagger$, and re-combining $\pi_i^\dagger$ and π_{-i} into joint policies.[4]

[4]For this reason, the policy class $\Pi_i^\dagger(\pi)$ always satisfies the following: for any $i \in [m]$ and any $\pi' \in \Pi_i^\dagger(\pi)$, $\pi'_{-i} = \pi_{-i}$.

The space of possible $\pi_i^\dagger$ which player i can choose from determines the mapping, and will take different forms under different notions of equilibria, as explained next.

1. A product policy is NE if it satisfies that no player can increase her gain by deviating from her own policy. Therefore, the response class mapping for NE is defined as $\Pi^{\dagger,\text{NE}}(\pi) := \{\Pi_i^{\dagger,\text{NE}}(\pi)\}_{i \in [m]}$, where $\Pi_i^{\dagger,\text{NE}}(\pi) := \{\pi_i^\dagger \times \pi_{-i} : \pi_i^\dagger \in \Pi_i\}$. Note that here $\Pi_i^{\dagger,\text{NE}}$ has no dependence on the input π_i, and player i simply considers using some $\pi_i^\dagger \in \Pi_i$ to replace π_i.
2. A CE is defined by a class of *strategy modifications* $\Phi = (\Phi_i)_{i \in [m]}$, where $\Phi_i \subseteq (\mathcal{S} \times \mathcal{A}_i \to \mathcal{A}_i)$ is a set of strategy modifications of the ith player, and each $\phi_i \in \Phi_i$ is a mapping $\phi_i : \mathcal{S} \times \mathcal{A}_i \to \mathcal{A}_i$. For any joint policy π, the modified policy $\phi_i \diamond \pi$ is defined as: at state $s \in \mathcal{S}$, all players sample $\mathbf{a} \sim \pi(\cdot|s)$, the ith player changes action $\mathbf{a}_i$ to $\phi_i(s, a_i)$ and $\mathbf{a}_{-i}$ remains the same. For CE, the response class mapping of each joint policy π is defined as $\Pi^{\dagger,\text{CE}}(\pi) := \{\Pi_i^{\dagger,\text{CE}}(\pi)\}_{i \in [m]}$, where $\Pi_i^{\dagger,\text{CE}}(\pi) = \{(\phi_i \diamond \pi_i) \odot \pi_{-i} : \phi_i \in \Phi_i\}$.
3. CCE is defined for general (i.e., possibly correlated) joint policies and is a relaxation of NE. The only difference is that CCE does not require the candidate policy π to be a product policy. Hence, the response class mapping of CCE is the same as that of NE.

With the definition of response class mapping, for $\text{EQ} \in \{\text{NE}, \text{CE}, \text{CCE}\}$, we define the gap of any joint policy π with respect to $\Pi^{\dagger,\text{EQ}}(\cdot)$ as

$$\text{Gap}^{\Pi^{\dagger,\text{EQ}}}(\pi) := \max_{i \in [m]} \max_{\pi^\dagger \in \Pi_i^{\dagger,\text{EQ}}(\pi)} V_i^{\pi^\dagger}(s_0) - V_i^\pi(s_0).$$

Now we are ready to present the definitions of three equilibria.

Definition 1 (Equilibria; NE, CE, and CCE)**.** For $\text{EQ} \in \{\text{NE}, \text{CE}, \text{CCE}\}$, a joint policy (product for NE) is an ε-EQ with respect to $\Pi^{\dagger,\text{EQ}}(\cdot)$, if for the response class $\Pi^{\dagger,\text{EQ}}(\pi)$,

$$\text{Gap}^{\Pi^{\dagger,\text{EQ}}}(\pi) \leq \varepsilon.$$

Definition 1 is defined with respect to the policy class Π and strategy modification class Φ (for CE). Throughout the paper, we focus on the theoretical guarantees of such "in-class" notion of gaps, which is a reasonable definition if we assume that all players have limited representation power and must work with restricted policy classes. Under additional assumptions (which we call "strategy completeness"; see Appendix A), such "in-class" gaps can be related to a stronger notion of gap where unrestricted deviation policies are considered for the best response.

With the response class mappings, we also define the extended policy class $\Pi_i^{\text{ext}} := (\bigcup_{\pi\in\Pi} \Pi_i^\dagger(\pi)) \bigcup \Pi$, which characterizes all possible policies with deviation from the ith player. In addition, we define $\Pi^{\text{ext}} := \bigcup_{i=1}^m \Pi_i^{\text{ext}}$.

4. Information-Theoretic Results for Multi-player General-sum Markov Games

4.1. Algorithm

As our learning goal is to find a policy $\pi \in \Pi$ with small equilibrium gap $\text{Gap}^{\Pi^\dagger,\text{EQ}}(\pi)$ (for EQ$\in$ {NE, CE, CCE}), a natural idea is to simply estimate the gap and minimize it over $\pi \in \Pi$. Unfortunately, we are in the offline setting and only have access to data sampled from an arbitrary data distribution d_D, which may not provide enough information for evaluating the gap of certain policies.

Since the gap is not always amendable to estimation, we instead seek a surrogate objective that will always be an *upper bound* on the equilibrium gap of each candidate policy $\pi \in \Pi$. The upper bound should also be *tight* when the policy is covered by the data and we have sufficient information to determine its gap accurately. To achieve this goal, we recall the definition of gap:

$$\text{Gap}^{\Pi^\dagger,\text{EQ}}(\pi) := \max_{i\in[m]} \max_{\pi^\dagger\in\Pi_i^{\dagger,\text{EQ}}(\pi)} V_i^{\pi^\dagger}(s_0) - V_i^\pi(s_0).$$

The key idea in our algorithm is that

$$V_i^{\pi^\dagger}(s_0) - V_i^\pi(s_0) \le \overline{V}_i^{\pi^\dagger}(s_0) - \underline{V}_i^\pi(s_0),$$

where

- $\overline{V}_i^{\pi^\dagger}(s_0) \ge V_i^{\pi^\dagger}(s_0)$ is an *optimistic* evaluation of $\pi^\dagger$.
- $\underline{V}_i^\pi(s_0) \le V_i^\pi(s_0)$ is an *pessimistic* evaluation of π.

With this relaxation, the problem reduces to optimistic and pessimistic policy evaluation, for which we can borrow existing techniques from single-agent RL.

Bellman-consistent pessimism & optimism We use the Bellman-consistent pessimism framework from Xie et al. (2021a) to construct optimistic and pessimistic policy evaluations. For each player i, we first use dataset $\mathcal{D}$ to compute an empirical Bellman error of all function $f_i \in \mathcal{F}_i$ under Bellman operator $\mathcal{T}_i^\pi$,

$$\mathcal{E}_i(f_i,\pi;\mathcal{D}) := \mathcal{L}_i(f_i,f_i,\pi;\mathcal{D}) - \min_{f_i'\in\mathcal{F}_i} \mathcal{L}_i(f_i',f_i,\pi;\mathcal{D}),$$

$$\mathcal{L}_i(f_i',f_i,\pi;\mathcal{D}) := \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \left(f_i'(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2.$$

Similar to the single-agent setting, $\mathcal{E}_i(f_i,\pi;\mathcal{D})$ is a good approximation of the true Bellman error of f_i w.r.t. π, i.e., $\mathcal{E}_i(f_i,\pi;\mathcal{D}) \approx \|f_i - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2$, so we can construct a version space $\mathcal{F}_i^{\pi,\beta_f}$ for each player i and policy $\pi \in \Pi_i^{\text{ext}}$ in (3). To ensure that the best approximation of Q_i^π is contained in $\mathcal{F}_i^{\pi,\beta_f}$, given a failure probability $\delta > 0$, we pick the threshold parameter β_f as follows,

$$\beta_f = \frac{80V_{\max}^2 \log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} + 30\varepsilon_{\mathcal{F}},$$

where $\mathcal{F} = \bigcup_{i=1}^m \mathcal{F}_i$. Then, optimistic and pessimistic evaluations can be obtained by simply taking the highest and the lowest prediction on the initial state s_0 across all functions in the version space ((4) and (5)).

With $\overline{V}_i^{\pi^\dagger}(s_0)$ and $\underline{V}_i^\pi(s_0)$ at hand, we calculate the estimated gap $\widehat{\text{Gap}}_{\text{EQ}}(\pi)$ for each $\pi \in \Pi$ in (6). We select the policy $\widehat{\pi}$ with the lowest estimated gap and the algorithm is summarized in Algorithm 1.

Algorithm 1 Bellman-Consistent Equilibrium Learning (BCEL) from an Offline Dataset

1: **Input:** Offline dataset $\mathcal{D}$, parameter β_f, equilibrium EQ $\in$ {NE, CE, CCE}

2: For each player $i \in [m]$ and policy $\pi \in \Pi_i^{\text{ext}}$, construct version space

$$\mathcal{F}_i^{\pi,\beta_f} = \{f_i \in \mathcal{F}_i : \mathcal{E}_i(f_i,\pi;\mathcal{D}) \le \beta_f\}. \quad (3)$$

3: For each player $i \in [m]$, compute

$$\overline{V}_i^{\pi^\dagger}(s_0) = \max_{f\in\mathcal{F}_i^{\pi^\dagger,\beta_f}} f(s_0,\pi^\dagger), \quad \forall \pi^\dagger \in \Pi_i^{\text{ext}}. \quad (4)$$

$$\underline{V}_i^\pi(s_0) = \min_{f\in\mathcal{F}_i^{\pi,\beta_f}} f(s_0,\pi), \quad \forall \pi \in \Pi. \quad (5)$$

4: For each policy $\pi \in \Pi$, compute the estimated gap

$$\widehat{\text{Gap}}_{\text{EQ}}(\pi) := \max_{i\in[m]} \max_{\pi^\dagger\in\Pi_i^{\dagger,\text{EQ}}(\pi)} \overline{V}_i^{\pi^\dagger}(s_0) - \underline{V}_i^\pi(s_0). \quad (6)$$

5: Output $\widehat{\pi} \leftarrow \min_{\pi\in\Pi} \widehat{\text{Gap}}_{\text{EQ}}(\pi)$.

4.2. Theoretical Guarantees

Before presenting the theoretical guarantee, we introduce the interval width Δ_i^π of $\mathcal{F}_i^{\pi,\beta_f}$, which will play a key role in our main theorem statement:

$$\Delta_i^\pi := \max_{f_i\in\mathcal{F}_i^{\pi,\beta_f}} f_i(s_0,\pi) - \min_{f_i\in\mathcal{F}_i^{\pi,\beta_f}} f_i(s_0,\pi).$$

As we will see, Δ_i^π is a measure of how well the data distribution d_D covers d^π, the state-action occupancy of π. The better coverage, the smaller Δ_i^π. This is formalized by the following proposition:

Proposition 2 (Bound on interval width)**.** *With probability at least $1-\delta$, for any player $i \in [m]$ and any $\pi \in \Pi_i^{\mathrm{ext}}$, we have*

$$\Delta_i^\pi \le \min_d \frac{1}{1-\gamma}\sqrt{\mathscr{C}(d; d_D, \mathcal{F}_i, \pi)}\varepsilon_{\mathrm{apx}} + \frac{1}{1-\gamma}\times \sum_{s,\mathbf{a}}(d^\pi \backslash d)(s,\mathbf{a})\left[\Delta f_i^\pi(s,\mathbf{a}) - \gamma(P^\pi \Delta f_i^\pi)(s,\mathbf{a})\right], \quad (7)$$

where $\varepsilon_{\mathrm{apx}} = \mathcal{O}\left(V_{\max}\sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\mathrm{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}\right)$, $(d^\pi \backslash d)(s,\mathbf{a}) := \max(d^\pi(s,\mathbf{a}) - d(s,\mathbf{a}), 0)$, $\Delta f_i^\pi(s,\mathbf{a}) := f_i^{\pi,\max}(s,\mathbf{a}) - f_i^{\pi,\min}(s,\mathbf{a})$, *and* $(P^\pi f)(s,\mathbf{a}) = \mathbb{E}_{s'\sim P(\cdot|s,\mathbf{a})}[f(s',\pi)]$.

Here, a distribution $d \in \Delta(\mathcal{S}\times\mathcal{A})$ in (7) is introduced to handle the discrepancy between d_D and d^π. The first term in (7) captures the distribution mismatch between d and d_D, and the second term represents the off-support Bellman error under π. When the data distribution d_D has a full coverage on d^π, d can be chosen as d^π and the second term becomes zero. Therefore, for the purpose of developing intuitions, one can always choose $d = d^\pi$ and treat $\Delta_i^\pi \propto \sqrt{\mathscr{C}(d^\pi; d_D, \mathcal{F}_i, \pi)}$, though in general some $d \neq d^\pi$ may achieve a better trade-off and tighter bound.

With an intuitive understanding of Δ_i^π, we are ready to show the following theorem for our proposed algorithm.

Theorem 3. *With probability at least $1-\delta$, for any $\pi \in \Pi$ and* $\mathrm{EQ} \in \{\mathrm{NE}, \mathrm{CE}, \mathrm{CCE}\}$*, the output policy $\widehat{\pi}$ of Algorithm 1 satisfies that*

$$\mathrm{Gap}^{\Pi^{\dagger,\mathrm{EQ}}}(\widehat{\pi}) \le \mathrm{Gap}^{\Pi^{\dagger,\mathrm{EQ}}}(\pi) + \frac{4\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} + \max_{i\in[m]} \min_{\widetilde{\pi}_i \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} \left(\Delta_i^{\widetilde{\pi}_i} + \Delta_i^\pi + \mathrm{subopt}_i^\pi(\widetilde{\pi}_i)\right),$$

where $\mathrm{subopt}_i^\pi(\widetilde{\pi}_i) := \max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} \overline{V}_i^{\pi^\dagger}(s_0) - \overline{V}_i^{\widetilde{\pi}_i}(s_0)$.

We remark that since our algorithm requires enumeration over the entire policy set, it is computationally inefficient and our results are information-theoretic; we leave the investigation of computationally-efficient algorithms to future work.

4.3. Improvement over Unilateral Coverage

To interpret Theorem 3 and compare it to existing guarantees, we first introduce a direct corollary of Theorem 3 + Proposition 2, which is a relaxed form of our result that is closer to existing guarantees by Cui & Du (2022a;b).

Corollary 4. *For Nash equilibrium policy $\pi^\star \in \Pi$, suppose there exists an unilateral coefficient $C(\pi^\star)$ such that the following inequality holds*

$$\max_{i\in[m]} \max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{NE}}(\pi^\star)} \mathscr{C}(d^{\pi^\dagger}; d_D, \mathcal{F}_i, \pi^\star) \le C(\pi^\star). \quad (8)$$

With probability at least $1-\delta$, we have

$$\mathrm{Gap}^{\Pi^{\dagger,\mathrm{NE}}}(\widehat{\pi}) \le \mathcal{O}\left(\frac{V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\mathrm{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}}{1-\gamma}\sqrt{C(\pi^\star)}\right).$$

The gap bound in Corollary 4 takes a simple form: the first part of it has an $O(1/\sqrt{n})$ statistical error (scaled by the complexities of function and policy classes, as fully expected), and an approximation error term that depends on $\varepsilon_{\mathcal{F}}, \varepsilon_{\mathcal{F},\mathcal{F}}$, which goes to 0 when our function classes are exactly realizable and Bellman-complete.

The key item in the bound is the $\sqrt{C(\pi^\star)}$ factor, which measures distribution mismatch and implicitly determines the data coverage condition. $C(\pi^\star)$ is defined in (8). As we can see, having a small $C(\pi^\star)$ requires that data not only covers $\pi^\star$ itself[5], but also *all policies in* $\Pi_i^{\dagger,\mathrm{NE}}(\pi^\star) = \{\pi_i \times \pi_{-i}^\star : \pi_i \in \Pi_i\}$. This is the notion of *unilateral coverage* proposed by Cui & Du (2022a) and Zhong et al. (2022). Visualizing this in Figure 1(a) with a simplified setting of a two-player matrix game, such a condition corresponds to data covering the entire "cross" centered at the NE.

Although (Cui & Du, 2022a) argues that unilateral coverage is "sufficient and necessary" in the worst case, their argument does not exclude an improved version that can be substantially relaxed under certain conditions, and we show that our Theorem 3 is such a version. We now provide a breakdown of the bound in Theorem 3:

1. First, the RHS of the bound depends on Δ_i^π, where π is the policy we compete with and correspond to $\pi^\star$ in Corollary 4. Recalling that $\Delta_i^\pi \propto \sqrt{\mathscr{C}(d^\pi; d_D, \mathcal{F}_i, \pi)}$, this term corresponds to data coverage on $\pi^\star$, which is always needed if we wish to compete with $\pi^\star$.
2. The RHS also depends on $\Delta_i^{\widetilde{\pi}_i} + \mathrm{subopt}_i^\pi(\widetilde{\pi}_i)$, where $\widetilde{\pi}_i$ is *minimized* over $\Pi_i^{\dagger,\mathrm{NE}}(\pi)$ when EQ=NE (and (8) *maximizes* over $\pi^\dagger$). In particular, we can always choose $\widetilde{\pi}_i$ as the policy that maximizes $\overline{V}_i$, i.e., the *optimistic best response*. This would set $\mathrm{subopt}_i^\pi(\widetilde{\pi}_i) = 0$, showing that we only need coverage for the optimistic best response policy, instead of all policies in $\Pi_i^{\dagger,\mathrm{NE}}(\pi^\star)$ as required by the unilateral assumption.
3. Finally, our bound provides a further relaxation: when the optimistic best response is poorly covered, we may choose some other well-covered $\widetilde{\pi}_i$ instead, and pay an

[5] Note that $\pi^\star \in \Pi_i^{\dagger,\mathrm{NE}}(\pi^\star)$.

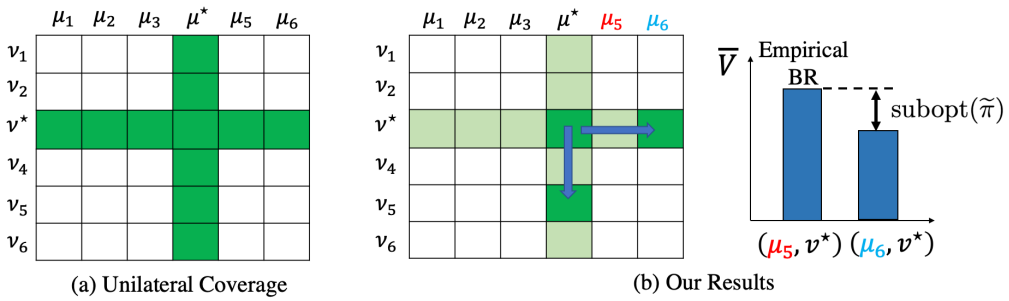

Figure 1. Illustration of unilateral coverage and our results on a zero-sum example. **(a)** Unilateral coverage requires the dataset to cover all unilateral pairs $(\mu^\star, \nu')$ and $(\mu', \nu^\star)$ where $(\mu^\star, \nu^\star)$ is NE. **(b)** Our approach enjoys an adaptive property and relaxes the condition. To begin with, we can already achieve a good sample complexity if the data were to cover the optimistic best response ($(\mu_5, \nu^\star)$ (in this example) only, i.e. when $\Delta^{\mu_5,\nu^\star}$ was small. Even when the dataset has a poor coverage on $(\mu_5, \nu^\star)$, there may exists some other μ_6 so that $\Delta^{\mu_5,\nu^\star} \gg \Delta^{\mu_6,\nu^\star}$. Instead of suffering $\Delta^{\mu_5,\nu^\star}$, our approach automatically adapts to the policy $\widetilde{\pi} = (\mu_6, \nu^\star)$ which achieves a better trade-off between the policy coverage term $\Delta^{\widetilde{\pi}}$ and suboptimality term $\mathrm{subopt}(\widetilde{\pi})$.

extra term $\mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i)$ that measures to what extent $\widetilde{\pi}_i$ is an approximate $\overline{V}_i$-based best response.

Again, we illustrate the flexibility of our bound in Figure 1(b). Below we also show a concrete example, where our guarantee leads to significantly improved sample rates compared to that provided by the unilateral condition.

Example Consider a simple two-player zero-sum matrix game with payoff matrix:

	b_1	b_2	b_3
a_1	0.5	0.75	0.75
a_2	0.25	0	0
a_3	0.25	0	0

where the column player aims to maximize the reward and the row player aims to minimize it. It is clear to see (a_1, b_1) is NE. The offline dataset $\mathcal{D}$ is collected from the following distribution,

	b_1	b_2	b_3
a_1	p_1	p_2	p_2
a_2	p_2	p_3	p_3
a_3	p_2	p_3	p_3

where $0 < p_2 \ll p_1$ and $p_3 = \frac{1-p_1-4p_2}{4}$. Under Corollary 4 (i.e., unilateral coverage (Cui & Du, 2022a)), the sample complexity bound is $\widetilde{\mathcal{O}}(\frac{1}{p_2\epsilon^2})$. However, when $n > \widetilde{\mathcal{O}}(\frac{1}{p_2})$, we already identify (a_1, b_2), (a_1, b_3), (a_2, b_1), and (a_3, b_1) as suboptimal actions with high probability. On this event, Theorem 3 shows that we only suffer the coverage coefficient on the optimistic best response (which is (a_1, b_1) itself), so that the sample complexity bound becomes $\widetilde{\mathcal{O}}(\max\{\frac{1}{p_2}, \frac{1}{p_1\epsilon^2}\}) \ll \widetilde{\mathcal{O}}(\frac{1}{p_2\epsilon^2})$.

5. V-type Function Approximation

A potential caveat of our approach in Section 4 is that we model Q-functions which take joint actions as inputs. In the tabular setting, the complexity of the full Q-function class has exponential dependence on the number of agents m, whereas prior results specialized to tabular settings do not suffer such a dependence.

While it is known that jointly featurizing the actions can avoid such an exponential dependence (Zhong et al., 2022) (in a way similar to how linear MDP results do not incur $|\mathcal{A}|$ dependence in the single-agent setting (Jin et al., 2020)), in this section we provide an alternative approach that directly subsumes the prior tabular results and produces the same rate (up to minor differences to be discussed). We propose a V-type variant algorithm of BCEL, which directly models the state-value function V_i^π with the help of a function class $\mathcal{G}_i \subset (\mathcal{S} \to [0, V_{\max}])$ for each player i.

As before, we assume that the tuples $(s, \mathbf{a}, \mathbf{r}, s') \in \mathcal{D}$ are generated as $(s, \mathbf{a}) \sim d_D$, $\mathbf{r}_i \sim r_i(s, \mathbf{a})$ and $s' \sim P(\cdot|s, \mathbf{a})$. In this section, we write $d_D = d_S \times d_A$, i.e., $(s, \mathbf{a}) \sim d_D \Leftrightarrow s \sim d_S, a \sim d_A(\cdot|s)$. We additionally assume that (1) $d_A(\mathbf{a}|s) > 0, \forall (s, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$, [6] and (2) the behavior policy $d_A(\cdot|s)$ is known to the learner. We use the behavior policy to perform importance weighting on the actions to correct the mismatch between d_A and π, and modify the loss function $\mathcal{L}$ as follows: for any function $g_i \in \mathcal{G}_i$, define

$$\mathcal{L}_i(g_i', g_i, \pi; \mathcal{D}) := \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s') \in \mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(g_i'(s) - \mathbf{r}_i - \gamma g_i(s')\right)^2.$$

[6] This assumption is w.l.o.g. and just for technical convenience, so that the action importance weights are always well defined. Otherwise, we can simply ignore any policy π where $\pi(\mathbf{a}|s)/d_A(\mathbf{a}|s)$ goes unbounded and assume maximum Δ_i^π for such π.

Similarly as before, we compute empirical Bellman error $\mathcal{E}_i(g_i, \pi; \mathcal{D}) := \mathcal{L}_i(g_i, g_i, \pi; \mathcal{D}) - \min_{g_i' \in \mathcal{G}_i} \mathcal{L}_i(g_i', g_i, \pi; \mathcal{D})$ and construct version space $\mathcal{G}_i^{\pi,\varepsilon} = \{g_i \in \mathcal{G}_i : \mathcal{E}_i(g_i, \pi; \mathcal{D}) \le \beta_g\}$. What is slightly different is that we set parameter β_g as

$$\beta_g := \frac{80 C_A(\pi) V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\mathrm{ext}}|}{\delta}}{n} + 30\varepsilon_{\mathcal{F}},$$

where $C_A(\pi) := \max_{s,\mathbf{a}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}$. Compared to β_f in Algorithm 1, the extra $C_A(\pi)$ term comes from importance weighting. With $\mathcal{G}_i^{\pi,\beta_g}$ at hand, we define

$$g_i^{\pi,\max} := \operatorname*{argmax}_{g_i \in \mathcal{G}_i^{\pi,\beta_g}} g_i(s_0), \quad g_i^{\pi,\min} := \operatorname*{argmin}_{g_i \in \mathcal{G}_i^{\pi,\beta_g}} g_i(s_0).$$

We then compute $\widehat{\mathrm{Gap}}_{\mathrm{EQ}}(\pi)$ which is an upper bound on equilibrium gap for any $\pi \in \Pi$:

$$\widehat{\mathrm{Gap}}_{\mathrm{EQ}}(\pi) := \max_{i \in [m]} \max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} g_i^{\pi^\dagger,\max}(s_0) - g_i^{\pi,\min}(s_0).$$

We select the policy by minimizing the estimated equilibrium gap:

$$\widehat{\pi} = \operatorname*{argmin}_{\pi \in \Pi} \widehat{\mathrm{Gap}}_{\mathrm{EQ}}(\pi), \tag{9}$$

whose performance guarantee is shown as follows.

Theorem 5 (V-type guarantee)**.** *With probability at least $1-\delta$, for any $\pi \in \Pi$ and $\mathrm{EQ} \in \{\mathrm{NE}, \mathrm{CE}, \mathrm{CCE}\}$, the output policy $\widehat{\pi}$ from (9) satisfies that*

$$\mathrm{Gap}^{\Pi^{\dagger,\mathrm{EQ}}}(\widehat{\pi}) \le \mathrm{Gap}^{\Pi^{\dagger,\mathrm{EQ}}}(\pi) + \frac{4\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} + \max_{i \in [m]} \min_{\widetilde{\pi}_i \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} \left(\Delta_i^{\widetilde{\pi}_i} + \Delta_i^{\pi} + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i) \right),$$

where $\Delta_i^\pi = g_i^{\pi,\max}(s_0) - g_i^{\pi,\min}(s_0)$ and $\mathrm{subopt}_i^{\widetilde{\pi}_i} = \max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} g_i^{\pi^\dagger,\max}(s_0) - g_i^{\widetilde{\pi}_i,\max}(s_0)$. In addition, with probability at least $1-\delta$, for any player $i \in [m]$ and any $\pi \in \Pi_i^{\mathrm{ext}}$, we have

$$\Delta_i^\pi \le \min_{d \in \Delta(S)} \frac{1}{1-\gamma} \sqrt{\mathscr{C}(d; d_S, \mathcal{G}_i, \pi)} \varepsilon_{\mathrm{apx}} + \frac{1}{1-\gamma} \sum_{s \in \mathcal{S}} (d^\pi \setminus d)(s) \left[\Delta g_i^\pi(s) - \gamma (P_i^\pi \Delta g_i^\pi)(s) \right],$$

where $\varepsilon_{\mathrm{apx}} = \mathcal{O}\left(V_{\max} \sqrt{C_A(\pi) \frac{\log \frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}} \right)$, $(d^\pi \setminus d)(s) := \max(d^\pi(s) - d(s), 0)$, $\Delta g_i^\pi(s) := g_i^{\pi,\max}(s) - g_i^{\pi,\min}(s)$, *and* $(P_i^\pi g)(s) = \mathbb{E}_{\mathbf{a} \sim \pi(\cdot|s), s' \sim P(\cdot|s,\mathbf{a})}[r_i(s,\mathbf{a}) + g(s')]$.

Similar to the results in Section 4, our bound enjoys an adaptive property and automatically selects the best policy $\widetilde{\pi}_i$, which achieves the trade-off between the suboptimality error $\mathrm{subopt}_i^\pi(\widetilde{\pi}_i)$ and the data coverage error $\Delta_i^{\widetilde{\pi}_i}$. Furthermore, when the dataset $\mathcal{D}$ satisfies the unilateral coverage assumption, we have the following corollary.

Corollary 6. *For Nash equilibrium policy $\pi^\star \in \Pi$, if there exists an unilateral coefficient $C_S(\pi^\star)$ such that the following inequality holds*

$$\max_{i \in [m]} \max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{NE}}(\pi^\star)} \mathscr{C}(d^{\pi^\dagger}; d_S, \mathcal{G}_i, \pi^\star) \le C_S(\pi^\star).$$

With probability at least $1-\delta$, we have

$$\mathrm{Gap}^{\Pi^{\dagger,\mathrm{NE}}}(\widehat{\pi}) \le \mathcal{O}\left(\frac{V_{\max} \sqrt{\frac{\log \frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}}{1-\gamma} \sqrt{C_A(\pi^\star) C_S(\pi^\star)} \right).$$

Compared to Corollary 4, our bound here depends logarithmically on the V-function class $\mathcal{G}$ instead of the Q-function class $\mathcal{F}$. In the tabular setting when we use fully expressive (and stationary) function classes, $\log|\mathcal{G}| \approx \mathcal{O}(S)$ (via a simple covering argument) and thus our bound avoids the exponential dependence on m (i.e., $\prod_{i=1}^m A_i$ dependence). In comparison, Cui & Du (2022b) established $\widetilde{\mathcal{O}}\left(\sqrt{H^4 S^2 \log(\mathcal{N}(\Pi)) C(\pi^\star)/n}\right)$ error bound for finite-horizon tabular Markov games, where H is the horizon length and $\mathcal{N}(\Pi)$ roughly corresponds to our $|\Pi^{\mathrm{ext}}|$. While finite-horizon and discounted results are generally incomparable, under a standard translation,[7] our bound has the same rate $\widetilde{\mathcal{O}}(n^{-1/2})$; $\sqrt{\log|\mathcal{G}|} \approx \sqrt{SH}$ [8] which results in a better dependence on S (saving a $\sqrt{S}$ factor) and a worse overall dependence on H (we have $\sqrt{H^5}$). The slight downside is that Corollary 6 measures distribution mismatch on actions and states separately (instead of doing them jointly as $C(\pi^*)$ in Corollary 4), which is looser.

6. Discussion and Conclusion

Algorithms for two-player zero-sum games For most part of this paper we consider the general case of multi-player general-sum Markov games. We discover that when our algorithm is specialized to the special case of two-player zero-sum (2p0s), it seemingly differs from another sample-efficient algorithm specifically designed for 2p0s inspired by Jin et al. (2022); Cui & Du (2022b). In Appendix B, we

[7] Cui & Du (2022b) assume that rewards are in $[0,1]$, thus we treat $V_{\max} = 1/(1-\gamma) = H$. When using fully expressive tabular classes, $\varepsilon_{\mathcal{F}} = \varepsilon_{\mathcal{F},\mathcal{F}} = 0$.

[8] In finite-horizon problems we need to use a non-stationary function class, therefore the extra H factor.

show that this difference is superficial, and these specialized algorithms can be subsumed as small variants of our algorithm.

Conclusion and open problems In this work, we study offline learning in Markov games. We design a framework that learn three popular equilibrium notions in a unified manner under general function approximation. The adaptive property of our framework enables us to relax and achieve significant improvement over the "unilateral concentrability" condition under certain situations.

One open problem is whether one can design a computational efficient algorithm for learning CE/CCE in offline Markov games, even in the tabular setting. A potential direction is to adapt the computationally efficient V-Learning algorithm (Song et al., 2022; Jin et al., 2021b)—which runs no-regret learning dynamics at each state—to the offline setting, which may require new ideas.

Acknowledgements

NJ acknowledges funding support from NSF IIS-2112471 and NSF CAREER IIS-2141781.

References

Antos, A., Szepesvári, C., and Munos, R. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

Aumann, R. J. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1): 67–96, 1974.

Bai, Y. and Jin, C. Provable self-play algorithms for competitive reinforcement learning. In *International Conference on Machine Learning*, pp. 551–560. PMLR, 2020.

Bai, Y., Jin, C., and Yu, T. Near-optimal reinforcement learning with self-play. *Advances in Neural Information Processing Systems*, 33:2159–2170, 2020.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Chen, Z., Zhou, D., and Gu, Q. Almost optimal algorithms for two-player zero-sum linear mixture markov games. In *International Conference on Algorithmic Learning Theory*, pp. 227–261. PMLR, 2022.

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

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

Daskalakis, C., Golowich, N., and Zhang, K. The complexity of markov equilibrium in stochastic games. *arXiv preprint arXiv:2204.03991*, 2022.

Dou, Z., Yang, Z., Wang, Z., and Du, S. Gap-dependent bounds for two-player markov games. In *International Conference on Artificial Intelligence and Statistics*, pp. 432–455. PMLR, 2022.

Huang, B., Huang, K., Kakade, S., Lee, J. D., Lei, Q., Wang, R., and Yang, J. Going beyond linear rl: Sample efficient neural function approximation. *Advances in Neural Information Processing Systems*, 34:8968–8983, 2021.

Huang, B., Lee, J. D., Wang, Z., and Yang, Z. Towards general function approximation in zero-sum markov games. *International Conference on Learning Representations*, 2022.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34:13406–13418, 2021a.

Jin, C., Liu, Q., Wang, Y., and Yu, T. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021b.

Jin, C., Liu, Q., and Yu, T. The power of exploiter: Provable multi-agent rl in large state spaces. In *International Conference on Machine Learning*, pp. 10251–10279. PMLR, 2022.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021c.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022.

Liu, Q., Yu, T., Bai, Y., and Jin, C. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pp. 7001–7010. PMLR, 2021.

Mao, W. and Başar, T. Provably efficient reinforcement learning in decentralized general-sum markov games. *Dynamic Games and Applications*, pp. 1–22, 2022.

Nash Jr, J. Non-cooperative games. In *Essays on Game Theory*, pp. 22–33. Edward Elgar Publishing, 1996.

Nowak, A. S. and Raghavan, T. Existence of stationary correlated equilibria with symmetric information for discounted stochastic games. *Mathematics of Operations Research*, 17(3):519–526, 1992.

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Shapley, L. S. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100, 1953.

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. *International Conference on Machine Learning*, pp. 19967–20025, 2022.

Song, Z., Mei, S., and Bai, Y. When can we learn general-sum markov games with a large number of players sample-efficiently? *International Conference on Learning Representations*, 2022.

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. *International Conference on Learning Representations*, 2022.

Wang, R., Salakhutdinov, R., and Yang, L. F. Provably efficient reinforcement learning with general value function approximation. *arXiv preprint arXiv:2005.10804*, 2020.

Wei, C.-Y., Hong, Y.-T., and Lu, C.-J. Online reinforcement learning in stochastic games. *Advances in Neural Information Processing Systems*, 30, 2017.

Xie, Q., Chen, Y., Wang, Z., and Yang, Z. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. In *Conference on Learning Theory*, pp. 3674–3682. PMLR, 2020.

Xie, T. and Jiang, N. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pp. 550–559. PMLR, 2020.

Xie, T. and Jiang, N. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pp. 11404–11413. PMLR, 2021.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:6683–6694, 2021a.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in Neural Information Processing Systems*, 34:27395–27407, 2021b.

Xiong, W., Zhong, H., Shi, C., Shen, C., Wang, L., and Zhang, T. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. *International Conference on Learning Representations*, 2023.

Yan, Y., Li, G., Chen, Y., and Fan, J. Model-based reinforcement learning is minimax-optimal for offline zero-sum markov games. *arXiv preprint arXiv:2206.04044*, 2022.

Yin, M. and Wang, Y.-X. Towards instance-optimal offline reinforcement learning with pessimism. *Advances in Neural Information Processing Systems*, 34:4065–4078, 2021.

Yin, M., Bai, Y., and Wang, Y.-X. Near-optimal offline reinforcement learning via double variance reduction. *Advances in Neural Information Processing Systems*, 34: 7677–7688, 2021a.

Yin, M., Bai, Y., and Wang, Y.-X. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1567–1575. PMLR, 2021b.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Zhong, H., Xiong, W., Tan, J., Wang, L., Zhang, T., Wang, Z., and Yang, Z. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. In *International Conference on Machine Learning*, pp. 27117–27142. PMLR, 2022.

A. Connection Between In-class Gap and Real Gap

In this paper we consider "in-class" equilibrium gaps that are defined w.r.t. certain deviation policy classes (Section 3). It is also common to consider stronger notions of equilibrium gap, which we denote simply as $\mathrm{Gap}^{\mathrm{EQ}}$, where the deviation policies are unrestricted, e.g., for NE and CCE, the deviation policies can take arbitrary policies (Nash Jr, 1996; Aumann, 1974).

To establish the connection between our in-class gap and the stronger notion of gap, we have the following strategy completeness assumption,

Assumption C (Strategy completeness)**.** For any player $i \in [m]$ and any $\mathrm{EQ} \in \{\mathrm{NE}, \mathrm{CCE}\}$, we have

$$\max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} V_i^{\pi^\dagger}(s_0) \geq \max_{\pi' : \pi'_{-i} = \pi_{-i}} V_i^{\pi'}(s_0) - \varepsilon_\Pi.$$

For CE, we have

$$\max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{CE}}(\pi)} V_i^{\pi^\dagger}(s_0) \geq \max_{\phi_i} V_i^{(\phi_i \diamond \pi_i) \odot \pi_{-i}}(s_0) - \varepsilon_\Pi.$$

Assumption C requires that the (unrestricted) best-response policy is contained in Π (and its counterpart for CE contained in Φ, respectively). Under Assumption C, it is clear that for any $\mathrm{EQ} \in \{\mathrm{NE}, \mathrm{CE}, \mathrm{CCE}\}$ and π, $\mathrm{Gap}^{\mathrm{EQ}}(\pi) \leq \mathrm{Gap}^{\Pi^{\dagger,\mathrm{EQ}}}(\pi) + \varepsilon_\Pi$.

B. A Connection in 2-player-0-sum Games

For the most part of this paper, we have considered the general case of multi-player general-sum Markov games. When we are in a specialized setting, such as two-player zero-sum Markov games (2p0s), it is often the case that we can exploit the special structure and come up with alternative algorithms (Yan et al., 2022; Jin et al., 2022).

In particular, Cui & Du (2022b, Section 3) design an offline 2p0s algorithm for the tabular setting, and extending their algorithm to the function approximation setting (using uncertainty quantification techniques from our paper) results in an algorithm that seemingly looks very different from our Algorithm 1. However, below we show that despite the superficial difference, the two algorithms are actually quite similar and can be derived using optimism/pessimism in the same way as in our Algorithm 1, with only one minor difference of minimizing the duality gap versus our $\mathrm{Gap}^{\mathrm{NE}}$. Consequently, for their algorithm, we can give guarantees similar to our Theorem 3, by slightly adapting our algorithm and analysis.

2p0s setup We now introduce some notation specialized to 2p0s games. We consider two players, where x-player aims to maximize the total reward while y-player aims to minimize it. The policy sets for x-player and y-player are denoted as $\Pi^{\max}$ and $\Pi^{\min}$ respectively. We consider the policy payoff $V \in [-1, 1]^{|\Pi^{\max}| \times |\Pi^{\min}|}$, where $V^{\mu,\nu}$ denotes the utility/loss for x-player/y-player when they follow policy μ and policy ν respectively. We use $\overline{V}$ and $\underline{V}$ to denote the UCB and the LCB estimation of V respectively. To connect these symbols with those in the main text, $V^{\mu,\nu}$ is essentially $V_1^\pi(s_0)(= -V_2^\pi(s_0))$ for $\pi = \mu \times \nu$, assuming player 1 is the max player x and player 2 is the min player y. Furthermore, we have $\overline{V}^{\mu,\nu} = \overline{V}_1^\pi(s_0) = -\underline{V}_2^\pi(s_0)$ and $\underline{V}^{\mu,\nu} = \underline{V}_1^\pi(s_0) = -\overline{V}_2^\pi(s_0)$ due to the 0-sum nature of the game.

Duality gap For 2p0s game, a common learning objective is the duality gap, which is defined as:

$$\text{Dual-Gap}(\mu, \nu) = \max_{\mu^\dagger \in \Pi^{\max}} V^{\mu^\dagger,\nu} - \min_{\nu^\dagger \in \Pi^{\min}} V^{\mu,\nu^\dagger}.$$

Since $\mu^\dagger$ and $\nu^\dagger$ can be chosen as μ and ν, the duality gap is always non-negative. It measures how close the policy is to NE policy and NE policy always has zero duality gap. Inspired by the tabular 2p0s algorithm from Cui & Du (2022b), one can design an offline algorithm that selects the two policies *independently* with adversarial opponent under pessimistic estimation:

$$\mu = \operatorname*{argmax}_{\mu} \min_{\nu^\dagger \in \Pi^{\min}} \underline{V}^{\mu,\nu^\dagger} \quad \text{and} \quad \nu = \operatorname*{argmin}_{\nu} \max_{\mu^\dagger \in \Pi^{\max}} \overline{V}^{\mu^\dagger,\nu}. \tag{10}$$

Similar ideas can also be found in Jin et al. (2022), who design online algorithms for 2p0s games. By flipping their optimism (for online) to pessimism (for offline), we can similarly arrive at (10).

Recover (10) in our algorithmic framework (10) looks very different from our Algorithm 1 at the first glance, as (10) chooses the players' policies independently whereas our Algorithm 1 requires joint optimization. We now show, however, that it is simply a minor variant of our algorithm, for which our analysis and guarantees straightforwardly extend.

First, note that the duality gap is not the same as our objective $\mathrm{Gap}^{\Pi^{\dagger,\mathrm{EQ}}}(\pi)$ when specialized to 2p0s games. Recall that

$$\mathrm{Gap}^{\Pi^{\dagger,\mathrm{EQ}}}(\pi) := \max_{i\in[m]} \max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} V_i^{\pi^\dagger}(s_0) - V_i^{\pi}(s_0).$$

To recover duality gap, we can simply replace the $\max_i$ in the above objective with $\sum_i$, and obtain the following in the 2p0s case:

$$\begin{aligned}\sum_{i\in[m]} \max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} V_i^{\pi^\dagger}(s_0) - V_i^{\pi}(s_0) &= (\max_{\mu^\dagger\in\Pi^{\max}} V^{\mu^\dagger,\nu} - V^{\mu,\nu}) + (V^{\mu,\nu} - \min_{\nu^\dagger\in\Pi^{\min}} V^{\mu,\nu^\dagger}) \\ &= \max_{\mu^\dagger\in\Pi^{\max}} V^{\mu^\dagger,\nu} - \min_{\nu^\dagger\in\Pi^{\min}} V^{\mu,\nu^\dagger} = \text{Dual-Gap}(\mu,\nu).\end{aligned}$$

From the above equation, we can see that our objective in Algorithm 1 is almost the same as the duality gap, up to a multiplicative factor of 2, as for non-negative a, b we have $\max(a,b) \le a+b \le 2\max(a,b)$. Therefore, our Algorithm 1 directly enjoys duality-gap guarantees.

However, remember that our goal here is to recover (10), so we choose to directly work with the duality gap and relax it in the same spirit as in our Algorithm 1: since $V \le \overline{V}$ and $-V \le -\underline{V}$, we have

$$\text{Dual-Gap}(\mu,\nu) = \max_{\mu^\dagger\in\Pi^{\max}} V^{\mu^\dagger,\nu} - \min_{\nu^\dagger\in\Pi^{\min}} V^{\mu,\nu^\dagger} \le \max_{\mu^\dagger\in\Pi^{\max}} \overline{V}^{\mu^\dagger,\nu} - \min_{\nu^\dagger\in\Pi^{\min}} \underline{V}^{\mu,\nu^\dagger}. \tag{11}$$

Now, (10) is recovered by noticing that μ and ν can be optimized independently on the RHS of (11) and the optima are exactly (10).

We also provide a guarantee for the above algorithm:

Proposition 7. *Consider a two-player zero-sum Markov game with policy payoff $V \in [-1,1]^{|\Pi^{\max}|\times|\Pi^{\min}|}$, let*

$$J(\mu,\nu) = \max_{\mu^\dagger\in\Pi^{\max}} \overline{V}^{\mu^\dagger,\nu} - \min_{\nu^\dagger\in\Pi^{\min}} \underline{V}^{\mu,\nu^\dagger}.$$

Let $\widehat{\mu},\widehat{\nu} = \mathrm{argmin}\, J(\mu,\nu)$, with high probability, we have

$$\text{Dual-Gap}(\widehat{\mu},\widehat{\nu}) \le \min_{\widetilde{\mu},\widetilde{\nu}\in\Pi^{\max}\times\Pi^{\min}} \Delta^{\widetilde{\mu},\nu^*} + \Delta^{\mu^*,\widetilde{\nu}} + \mathrm{subopt}^{\pi^\star}(\widetilde{\mu}) + \mathrm{subopt}^{\pi^\star}(\widetilde{\nu}),$$

where $\Delta^{\mu,\nu} := \overline{V}^{\mu,\nu} - \underline{V}^{\mu,\nu}$, $\mathrm{subopt}^{\pi^\star}(\widetilde{\mu}) := \max_{\mu^\dagger\in\Pi^{\max}} \overline{V}^{\mu^\dagger,\nu^*} - \overline{V}^{\widetilde{\mu},\nu^*}$ and $\mathrm{subopt}^{\pi^\star}(\widetilde{\nu}) := \underline{V}^{\mu^*,\widetilde{\nu}} - \min_{\nu^\dagger\in\Pi^{\min}} \underline{V}^{\mu^*,\nu^\dagger}$.

Proof. By standard concentration analysis, we guarantee that with high probability, $\overline{V}^{\mu,\nu} \ge V^{\mu,\nu}$ and $\underline{V}^{\mu,\nu} \le V^{\mu,\nu}$ hold for any $\mu,\nu \in \Pi^{\max}\times\Pi^{\min}$. This implies that for any $\mu,\nu\in\Pi^{\max}\times\Pi^{\min}$, $\text{Dual-Gap}(\mu,\nu) \le J(\mu,\nu)$. For Nash policy $\pi^\star = (\mu^*,\nu^*)$, let $\mu^\dagger = \mathrm{argmax}_{\mu^\dagger\in\Pi^{\max}} \overline{V}^{\mu^\dagger,\nu^*}$ and $\nu^\dagger = \mathrm{argmin}_{\nu^\dagger\in\Pi^{\min}} \underline{V}^{\mu^*,\nu^\dagger}$. We have

$$\begin{aligned}J(\mu^*,\nu^*) - \text{Dual-Gap}(\mu^*,\nu^*) &= \overline{V}^{\mu^\dagger,\nu^*} - \max_{\mu\in\Pi^{\max}} V^{\mu,\nu^*} + \min_{\nu\in\Pi^{\min}} V^{\mu^*,\nu} - \underline{V}^{\mu^*,\nu^\dagger} \\ &\le (\overline{V}^{\widetilde{\mu},\nu^*} - V^{\widetilde{\mu},\nu^*}) + (V^{\mu^*,\widetilde{\nu}} - \underline{V}^{\mu^*,\widetilde{\nu}}) + \mathrm{subopt}^{\pi^\star}(\widetilde{\mu}) + \mathrm{subopt}^{\pi^\star}(\widetilde{\nu}) \\ &\le \Delta^{\widetilde{\mu},\nu^*} + \Delta^{\mu^*,\widetilde{\nu}} + \mathrm{subopt}^{\pi^\star}(\widetilde{\mu}) + \mathrm{subopt}^{\pi^\star}(\widetilde{\nu}),\end{aligned} \tag{12}$$

where $\widetilde{\mu}$ and $\widetilde{\nu}$ are arbitrary polices from $\Pi^{\max}$ and $\Pi^{\min}$ respectively. By the optimality of $(\widehat{\mu},\widehat{\nu})$ and (12), we obtain

$$\text{Dual-Gap}(\widehat{\mu},\widehat{\nu}) \le J(\widehat{\mu},\widehat{\nu}) \le J(\mu^*,\nu^*) \le \min_{\widetilde{\mu},\widetilde{\nu}\in\Pi^{\max}\times\Pi^{\min}} \Delta^{\widetilde{\mu},\nu^*} + \Delta^{\mu^*,\widetilde{\nu}} + \mathrm{subopt}^{\pi^\star}(\widetilde{\mu}) + \mathrm{subopt}^{\pi^\star}(\widetilde{\nu}).$$

The proof is completed. □

C. Proofs for Section 4

In this section, we prove Theorem 3. We first show some concentration results.

Lemma 8. *With probability at least $1-\delta$, for any player $i \in [m]$, any $f_i, g_1, g_2 \in \mathcal{F}_i$, and any $\pi \in \Pi_i^{\text{ext}}$, we have*

$$\begin{aligned}
&\Bigg| \|g_1 - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 - \|g_2 - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 \\
&- \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s') \in \mathcal{D}} \left(g_1(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2 + \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s') \in \mathcal{D}} \left(g_2(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2 \Bigg| \\
\leq\ & 2V_{\max} \|g_1 - g_2\|_{2,d_D} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}.
\end{aligned}$$

Proof. For player i, we observe that

$$\begin{aligned}
&\|g_1 - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 - \|g_2 - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 \\
&= \mathbb{E}_{d_D}\left[\left(g_1(s,\mathbf{a}) - (\mathcal{T}_i^\pi f_i)(s,\mathbf{a})\right)^2\right] - \mathbb{E}_{d_D}\left[\left(g_2(s,\mathbf{a}) - (\mathcal{T}_i^\pi f_i)(s,\mathbf{a})\right)^2\right] \\
&= \mathbb{E}_{d_D}\left[(g_1(s,\mathbf{a}) - g_2(s,\mathbf{a}))(g_1(s,\mathbf{a}) + g_2(s,\mathbf{a}) - 2(\mathcal{T}_i^\pi f)(s,\mathbf{a}))\right] \\
&= \mathbb{E}_{d_D}\left[(g_1(s,\mathbf{a}) - g_2(s,\mathbf{a}))\mathbb{E}_{s' \sim P(\cdot|s,\mathbf{a})}\left[g_1(s,\mathbf{a}) + g_2(s,\mathbf{a}) - 2\mathbf{r}_i - 2\gamma f(s',\pi)|s,\mathbf{a}\right]\right] \\
&= \mathbb{E}_{d_D \times P}\left[\left(g_1(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2\right] - \mathbb{E}_{d_D \times P}\left[\left(g_2(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2\right]. \qquad (13)
\end{aligned}$$

Let random variable $X = (g_1(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi))^2 - (g_2(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi))^2$, X is drawn from $d_D \times P$. We know that $\mathbb{E}_{d_D \times P}[X] = \mathbb{E}_{d_D \times P}\left[\left(g_1(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2\right] - \mathbb{E}_{d_D \times P}\left[\left(g_2(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2\right]$. For the variance, we have

$$\begin{aligned}
\mathbb{V}_{d_D \times P}[X] &\leq \mathbb{E}_{d_D \times P}[X^2] \\
&\leq \mathbb{E}_{d_D \times P}[(g_1(s,\mathbf{a}) - g_2(s,\mathbf{a})^2 (g_1(s,\mathbf{a}) + g_2(s,\mathbf{a}) - 2\mathbf{r}_i - 2\gamma f_i(s',\pi))^2] \\
&\leq 4V_{\max}^2 \mathbb{E}_{d_D}\left[(g_1(s,\mathbf{a}) - g_2(s,\mathbf{a}))^2\right].
\end{aligned}$$

We proceed as follows

$$\begin{aligned}
&\Bigg| \|g_1 - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 - \|g_2 - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 \\
&- \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s') \in \mathcal{D}} \left(g_1(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2 + \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s') \in \mathcal{D}} \left(g_2(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2 \Bigg| \\
&= \left| \mathbb{E}_{d_D \times P}\left[\left(g_1(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2\right] - \mathbb{E}_{d_D \times P}\left[\left(g_2(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2\right] - \frac{1}{n}\sum_{j=1}^n X_j \right| \\
&\qquad\qquad \text{(By (13) and definition of } X) \\
&\leq \sqrt{\frac{4V_{\max}^2 \|g_1 - g_2\|_{d_D}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}. \qquad \text{(By Freedman's inequality)}
\end{aligned}$$

Taking a union bound over $i \in [m]$ finishes the proof. □

For any player $i \in [m]$ and $\pi \in \Pi_i^{\text{ext}}$, let us define

$$f_i^\pi := \operatorname*{argmin}_{f \in \mathcal{F}_i} \sup_{\text{admissible } d} \|f - \mathcal{T}_i^\pi f\|_{2,d}^2 \qquad (14)$$

$$g_i^\pi := \operatorname*{argmin}_{g \in \mathcal{F}_i} \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s') \in \mathcal{D}} \left(g(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i^\pi(s',\pi)\right)^2. \qquad (15)$$

We bound $\|f_i^\pi - g_i^\pi\|_{2,d_D}$ as follows.

Lemma 9. *Let f_i^π and g_i^π be defined as in Equations (14) and (15). Under the success event of Lemma 8, for any player $i \in [m]$ and $\pi \in \Pi_i^{\text{ext}}$, we have*

$$\|f_i^\pi - g_i^\pi\|_{2,d_D} \leq 6V_{\max}\sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + 2\sqrt{\varepsilon_{\mathcal{F}}}.$$

Proof. We know that

$$\begin{aligned}
&\|f_i^\pi - g_i^\pi\|_{2,d_D}^2 \\
&\leq 2\|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 + 2\|g_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 \\
&= 2\|g_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 - 2\|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 + 4\|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 \\
&\leq 2\|g_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 - 2\|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 + 4\varepsilon_{\mathcal{F}} && \text{(By Assumption A)} \\
&\overset{\text{(a)}}{\leq} 4V_{\max}\sqrt{\frac{\|g_i^\pi - f_i^\pi\|_{2,d_D}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{2V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} + 4\varepsilon_{\mathcal{F}}, && (16)
\end{aligned}$$

where (a) is from

$$\begin{aligned}
&\|g_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 - \|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 \\
&\leq \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \left(g_i^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i^\pi(s',\pi)\right)^2 - \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \left(f_i^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i^\pi(s',\pi)\right)^2 \\
&\quad + 2V_{\max}\sqrt{\frac{\|g_i^\pi - f_i^\pi\|_{2,d_D}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} && \text{(by Lemma 8)} \\
&\leq 2V_{\max}\sqrt{\frac{\|g_i^\pi - f_i^\pi\|_{2,d_D}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} && \text{(by the optimality of } g)
\end{aligned}$$

Solving (16) finishes the proof. □

In the following lemma, we show that the best approximation of Q_i^π is contained in $\mathcal{F}_i^{\pi,\beta_f}$.

Lemma 10. *Under the success event of Lemma 8, for any player $i \in [m]$ and $\pi \in \Pi_i^{\text{ext}}$, the following inequality for $\mathcal{E}_i(f_i^\pi, \pi; \mathcal{D})$ holds*

$$\mathcal{E}_i(f_i^\pi, \pi; \mathcal{D}) \leq \frac{80V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} + 30\varepsilon_{\mathcal{F}} =: \beta_f.$$

Proof. Applying Lemma 8 and Lemma 9, we obtain that

$$\begin{aligned}
&\Bigg| \|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 - \|g_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 - \\
&\frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \left(f_i^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i^\pi(s',\pi)\right)^2 + \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \left(g_i^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i^\pi(s',\pi)\right)^2 \Bigg| \\
&\leq 2V_{\max}\|f_i^\pi - g_i^\pi\|_{2,d_D}\sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \\
&\leq 4V_{\max}\sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F}}} + \frac{13V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}. && (17)
\end{aligned}$$

Then, we bound $\|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 - \|g_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2$ as follows,

$$\|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 - \|g_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2$$

$$\leq \left(\|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D} + \|g_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}\right) \left|\|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D} - \|g_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}\right|$$

$$\leq \left(2\|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D} + \|f_i^\pi - g_i^\pi\|_{2,d_D}\right) \|f_i^\pi - g_i^\pi\|_{2,d_D} \qquad \text{(By triangle inequality)}$$

$$\leq 36V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F}}} + 36V_{\max}^2\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} + 8\varepsilon_{\mathcal{F}}. \qquad \text{(By Assumption A and Lemma 9)}$$

Combining this with (17), we get

$$\frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}}(f_i^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i^\pi(s',\pi))^2 - \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}}(g_i^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i^\pi(s',\pi))^2$$

$$\leq \|f_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 - \|g_i^\pi - \mathcal{T}_i^\pi f_i^\pi\|_{2,d_D}^2 + 4V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F}}} + \frac{13V_{\max}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}$$

$$\leq 40V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F}}} + \frac{59V_{\max}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} + 8\varepsilon_{\mathcal{F}}$$

$$\leq \frac{80V_{\max}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} + 30\varepsilon_{\mathcal{F}}. \qquad (\sqrt{ab} \leq \tfrac{a+b}{2})$$

□

Then, we show that $|f_i^\pi(s_0,\pi) - V_i^\pi(s_0)|$ is upper bounded as follows

Lemma 11. *For any player $i \in [m]$ and $\pi \in \Pi_i^{\text{ext}}$, let f_i^π be defined as in (14), we have*

$$|f_i^\pi(s_0,\pi) - V_i^\pi(s_0)| \leq \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.$$

Proof. By invoking Lemma 20, we get

$$|f_i^\pi(s_0,\pi) - V_i^\pi(s_0)| \leq \frac{|\mathbb{E}_{s,\mathbf{a}\sim d^\pi}[f(s,\mathbf{a}) - (\mathcal{T}_i^\pi f)(s,\mathbf{a})]|}{1-\gamma}$$
$$\leq \frac{\|f - \mathcal{T}_i^\pi f\|_{2,d^\pi}}{1-\gamma} \leq \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.$$

The second inequality is from Jensen's inequality and the last inequality follows from Assumption A. □

For the version space $\mathcal{F}_i^{\pi,\beta_f}$, we define

$$f_i^{\pi,\max} := \operatorname*{argmax}_{f_i\in\mathcal{F}_i^{\pi,\beta_f}} f_i(s_0,\pi)$$
$$f_i^{\pi,\min} := \operatorname*{argmin}_{f_i\in\mathcal{F}_i^{\pi,\beta_f}} f_i(s_0,\pi).$$

We show that $f_i^{\pi,\max}(s_0,\pi)$ and $f_i^{\pi,\min}(s_0,\pi)$ are the upper bound and the lower bound on the value function $V_i^\pi(s_0)$ respectively.

Lemma 12. *Under the success event of Lemma 8, for any player $i \in [m]$ and any $\pi \in \Pi_i^{\text{ext}}$, the following two inequalities hold*

$$f_i^{\pi,\max}(s_0,\pi) \geq V_i^\pi(s_0) - \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}$$
$$f_i^{\pi,\min}(s_0,\pi) \leq V_i^\pi(s_0) + \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.$$

Proof. By Lemma 10, we know that under the success event of Lemma 8, $f_i^\pi \in \mathcal{F}_i^{\pi,\beta_f}$. Then, we invoke Lemma 11 and get

$$f_i^{\pi,\min}(s_0,\pi) \le f_i^\pi(s_0,\pi) \le Q_i^\pi(s_0,\pi) + \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} = V_i^\pi(s_0) + \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.$$

Similarly, we have

$$f_i^{\pi,\max}(s_0,\pi) \ge f_i^\pi(s_0,\pi) \ge Q_i^\pi(s_0,\pi) - \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} = V_i^\pi(s_0) - \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.$$

□

We now show that $\mathcal{E}_i(f_i,\pi;\mathcal{D})$ could effectively estimate $\|f_i - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2$.

Lemma 13. *Under the success event of Lemma 8, for any player $i \in [m]$ and any $\pi \in \Pi_i^{\text{ext}}$, given $\varepsilon > 0$, if $f_i \in \mathcal{F}_i$ satisfies that $\mathcal{E}_i(f_i,\pi;\mathcal{D}) \le \varepsilon$, we have*

$$\|f_i - \mathcal{T}_i^\pi f_i\|_{2,d_D} \le 8V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + 4\sqrt{\varepsilon_{\mathcal{F},\mathcal{F}}} + \sqrt{\varepsilon}.$$

Proof. Let g_i^π be defined as in (15), we first upper bound term $\|g_i^\pi - \mathcal{T}_i^\pi f_i\|_{2,d_D}$. Let us define

$$f_{i,d_D}^\pi \coloneqq \operatorname*{argmin}_{f_i' \in \mathcal{F}_i} \|f_i' - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2.$$

By invoking Lemma 8, we obtain that

$$\begin{aligned}
&\Bigg| \|g_i^\pi - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 - \left\|f_{i,d_D}^\pi - \mathcal{T}_i^\pi f_i\right\|_{2,d_D}^2 - \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \left(g_i^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2 \\
&\quad + \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \left(f_{i,d_D}^\pi(s,\mathbf{a}) - r - \gamma f_i(s',\pi)\right)^2 \Bigg| \\
&\le 2V_{\max}\|g_i^\pi - f_{i,d_D}^\pi\|_{2,d_D}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2 \log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}.
\end{aligned}$$

Rearranging the terms and we have

$$\begin{aligned}
&\|g_i^\pi - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 \\
&\le \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \left(g_i^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2 - \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \left(f_{i,d_D}^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2 \\
&\quad + \left\|f_{i,d_D}^\pi - \mathcal{T}_i^\pi f_i\right\|_{2,d_D}^2 + 2V_{\max}\|g_i^\pi - f_{i,d_D}^\pi\|_{2,d_D}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \\
&\le \left\|f_{i,d_D}^\pi - \mathcal{T}_i^\pi f_i\right\|_{2,d_D}^2 + 2V_{\max}\|g_i^\pi - f_{i,d_D}^\pi\|_{2,d_D}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \\
&\le \varepsilon_{\mathcal{F},\mathcal{F}} + 2V_{\max}\|g_i^\pi - \mathcal{T}_i^\pi f_i\|_{2,d_D}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + 2V_{\max}\sqrt{\varepsilon_{\mathcal{F},\mathcal{F}}}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \\
&\le 2V_{\max}\|g_i^\pi - \mathcal{T}_i^\pi f_i\|_{2,d_D}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{2V_{\max}^2\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} + 2\varepsilon_{\mathcal{F},\mathcal{F}}. \qquad (18)
\end{aligned}$$

The second inequality is from the optimality of g_i^π. The third inequality follows from Assumption B and $\|g_i^\pi - f_{i,d_D}^\pi\|_{2,d_D} \le \|g_i^\pi - \mathcal{T}_i^\pi f_i\|_{2,d_D} + \|f_{i,d_D}^\pi - \mathcal{T}_i^\pi f_i\|_{2,d_D}$. The last inequality is from $\sqrt{ab} \le \frac{a+b}{2}$. By solving (18), we get

$$\|g_i^\pi - \mathcal{T}_i^\pi f_i\|_{2,d_D} \le 3V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \sqrt{2\varepsilon_{\mathcal{F},\mathcal{F}}}. \qquad (19)$$

Then, we invoke Lemma 8 for f_i^π

$$
\begin{aligned}
&\Bigg| \|f_i - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 - \|g_i^\pi - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 \\
&\quad - \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \left(f_i(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2 + \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \left(g_i^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2 \Bigg| \\
&\le 2V_{\max} \|f_i - g_i^\pi\|_{2,d_D} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \\
&\le 2V_{\max} \left(\|f_i - \mathcal{T}_i^\pi f_i\|_{2,d_D} + \|g_i^\pi - \mathcal{T}_i^\pi f_i\|_{2,d_D}\right) \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \\
&\le 2V_{\max} \|f_i - \mathcal{T}_i^\pi f_i\|_{2,d_D} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + 3V_{\max} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \varepsilon_{\mathcal{F},\mathcal{F}}} + \frac{7V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}. \qquad \text{(By (19))}
\end{aligned}
$$

Rearranging the terms, we get

$$
\begin{aligned}
&\|f_i - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 \\
&\le \|g_i^\pi - \mathcal{T}_i^\pi f_i\|_{2,d_D}^2 + \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \left(f_i(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2 - \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \left(g_i^\pi(s,\mathbf{a}) - \mathbf{r}_i - \gamma f_i(s',\pi)\right)^2 \\
&\quad + 2V_{\max} \|f_i - \mathcal{T}_i^\pi f_i\|_{2,d_D} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + 3V_{\max} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \varepsilon_{\mathcal{F},\mathcal{F}}} + \frac{7V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \\
&\le \left(3V_{\max} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \sqrt{2\varepsilon_{\mathcal{F},\mathcal{F}}}\right)^2 + \varepsilon \qquad \text{(By (19) and } \mathcal{E}_i(f_i,\pi;\mathcal{D}) \le \varepsilon) \\
&\quad + 2V_{\max} \|f_i - \mathcal{T}_i^\pi f_i\|_{2,d_D} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + 3V_{\max} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \varepsilon_{\mathcal{F},\mathcal{F}}} + \frac{7V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \\
&= 2V_{\max} \|f_i - \mathcal{T}_i^\pi f_i\|_{2,d_D} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + 12V_{\max} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} \varepsilon_{\mathcal{F},\mathcal{F}}} + \frac{16V_{\max}^2 \log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} + 2\varepsilon_{\mathcal{F},\mathcal{F}} + \varepsilon. \qquad (20)
\end{aligned}
$$

Solving (20) and using AM-GM inequality finishes the proof. □

We upper bound Δ_i^π as follows.

Proposition 2 (Bound on interval width)**.** *With probability at least $1-\delta$, for any player $i \in [m]$ and any $\pi \in \Pi_i^{\text{ext}}$, we have*

$$
\begin{aligned}
\Delta_i^\pi \le \min_d \frac{1}{1-\gamma} \sqrt{\mathscr{C}(d; d_D, \mathcal{F}_i, \pi)} \varepsilon_{\text{apx}} + \frac{1}{1-\gamma} \times \\
\sum_{s,\mathbf{a}} (d^\pi \setminus d)(s,\mathbf{a}) \left[\Delta f_i^\pi(s,\mathbf{a}) - \gamma (P^\pi \Delta f_i^\pi)(s,\mathbf{a})\right], \qquad (7)
\end{aligned}
$$

where $\varepsilon_{\text{apx}} = \mathcal{O}\left(V_{\max} \sqrt{\frac{\log \frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}\right)$, $(d^\pi \setminus d)(s,\mathbf{a}) := \max(d^\pi(s,\mathbf{a}) - d(s,\mathbf{a}), 0)$, $\Delta f_i^\pi(s,\mathbf{a}) := f_i^{\pi,\max}(s,\mathbf{a}) - f_i^{\pi,\min}(s,\mathbf{a})$, *and* $(P^\pi f)(s,\mathbf{a}) = \mathbb{E}_{s'\sim P(\cdot|s,\mathbf{a})}[f(s',\pi)]$.

Proof. We apply Lemma 20 for $f_i^{\pi,\max}$ and $f_i^{\pi,\min}$ and obtain

$$
\begin{aligned}
&f_i^{\pi,\max}(s_0,\pi) - f_i^{\pi,\min}(s_0,\pi) \\
&= f_i^{\pi,\max}(s_0,\pi) - V_i^\pi(s_0) + V_i^\pi(s_0) - f_i^{\pi,\min}(s_0,\pi).
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{1-\gamma}\left(\mathbb{E}_{d_\pi}\left[f_i^{\pi,\max} - \mathcal{T}_i^\pi f_i^{\pi,\max}\right] - \mathbb{E}_{d_\pi}\left[f_i^{\pi,\min} - \mathcal{T}_i^\pi f_i^{\pi,\min}\right]\right) && \text{(By Lemma 20)}\\
&= \frac{1}{1-\gamma}\Big(\mathbb{E}_d\left[(f_i^{\pi,\max} - \mathcal{T}_i^\pi f_i^{\pi,\max}) - \left(f_i^{\pi,\min} - \mathcal{T}_i^\pi f_i^{\pi,\min}\right)\right]\\
&\quad + \mathbb{E}_{d_\pi}\left[(f_i^{\pi,\max} - \mathcal{T}_i^\pi f_i^{\pi,\max}) - (f_i^{\pi,\min} - \mathcal{T}_i^\pi f_i^{\pi,\min})\right]\\
&\quad - \mathbb{E}_d\left[(f_i^{\pi,\max} - \mathcal{T}_i^\pi f_i^{\pi,\max}) - (f_i^{\pi,\min} - \mathcal{T}_i^\pi f_i^{\pi,\min})\right]\Big)\\
&= \frac{1}{1-\gamma}\underbrace{\mathbb{E}_d\left[(f_i^{\pi,\max} - \mathcal{T}_i^\pi f_i^{\pi,\max}) - \left(f_i^{\pi,\min} - \mathcal{T}_i^\pi f_i^{\pi,\min}\right)\right]}_{\text{(I)}}\\
&\quad + \frac{1}{1-\gamma}\underbrace{\left(\mathbb{E}_{d_\pi}\left[\Delta f_i^\pi - \gamma P^\pi \Delta f_i^\pi\right] - \mathbb{E}_d\left[\Delta f_i^\pi - \gamma P^\pi \Delta f_i^\pi\right]\right)}_{\text{(II)}}, && (\Delta f_i^\pi := f_i^{\pi,\max} - f_i^{\pi,\min})
\end{aligned}
$$

where $d \in \Delta(\mathcal{S}\times\mathcal{A})$ is an arbitrary distribution. For the term (I), we have

$$
\begin{aligned}
\text{(I)} &\le \left|\mathbb{E}_d\left[(f_i^{\pi,\max} - \mathcal{T}_i^\pi f_i^{\pi,\max})\right]\right| + \left|\mathbb{E}_d\left[\left(f_i^{\pi,\min} - \mathcal{T}_i^\pi f_i^{\pi,\min}\right)\right]\right|\\
&\le \|f_i^{\pi,\max} - \mathcal{T}_i^\pi f_i^{\pi,\max}\|_{2,d} + \|f_i^{\pi,\min} - \mathcal{T}_i^\pi f_i^{\pi,\min}\|_{2,d} && \text{(By Jensen's inequality)}\\
&\le \sqrt{\mathscr{C}(d; d_D, \mathcal{F}_i, \pi)}\left(\|f_i^{\pi,\max} - \mathcal{T}_i^\pi f_i^{\pi,\max}\|_{2,d_D} + \|f_i^{\pi,\min} - \mathcal{T}_i^\pi f_i^{\pi,\min}\|_{2,d_D}\right).
\end{aligned}
$$

Recall that $f_i^{\pi,\max} := \operatorname{argmax}_{f_i\in\mathcal{F}_i^{\pi,\beta_f}} f_i(s_0,\pi)$ and $f_i^{\pi,\min} := \operatorname{argmin}_{f_i\in\mathcal{F}_i^{\pi,\beta_f}} f_i(s_0,\pi)$ and $\beta_f = \frac{80V_{\max}^2 \log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n} + 30\varepsilon_{\mathcal{F}}$. We invoke Lemma 13 and have

$$
\text{(I)} \le \sqrt{\mathscr{C}(d; d_D, \mathcal{F}_i, \pi)}\,\mathcal{O}\left(V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}\right). \tag{21}
$$

For term (II), we have

$$
\begin{aligned}
\text{(II)} &\le \sum_{(s,\mathbf{a})\in\mathcal{S}\times\mathcal{A}} (d_\pi \setminus d)(s,\mathbf{a})\left[\Delta f_i^\pi(s,\mathbf{a}) - \gamma(P^\pi \Delta f_i^\pi)(s,\mathbf{a})\right]\\
&\quad + \sum_{(s,\mathbf{a})\in\mathcal{S}\times\mathcal{A}} \mathbb{I}(d(s,\mathbf{a}) > d_\pi(s,\mathbf{a}))\left[d(s,\mathbf{a}) - d_\pi(s,\mathbf{a})\right]\left|\Delta f_i^\pi(s,\mathbf{a}) - \gamma(P^\pi \Delta f_i^\pi)(s,\mathbf{a})\right|\\
&\le \sum_{(s,\mathbf{a})\in\mathcal{S}\times\mathcal{A}} (d_\pi \setminus d)(s,\mathbf{a})\left[\Delta f_i^\pi(s,\mathbf{a}) - \gamma(P^\pi \Delta f_i^\pi)(s,\mathbf{a})\right]\\
&\quad + \mathbb{E}_d\left[|f_i^{\pi,\max} - \mathcal{T}_i^\pi f_i^{\pi,\max}| + \left|f_i^{\pi,\min} - \mathcal{T}_i^\pi f_i^{\pi,\min}\right|\right]\\
&\le \sum_{(s,\mathbf{a})\in\mathcal{S}\times\mathcal{A}} (d_\pi \setminus d)(s,\mathbf{a})\left[\Delta f_i^\pi(s,\mathbf{a}) - \gamma(P^\pi \Delta f_i^\pi)(s,\mathbf{a})\right]\\
&\quad + \sqrt{\mathscr{C}(d; d_D, \mathcal{F}_i, \pi)}\,\mathcal{O}\left(V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}\right).
\end{aligned} \tag{22}
$$

The last step is from the analysis of term (I). Combining (21) and (22), we get

$$
\begin{aligned}
f_i^{\pi,\max}(s_0,\pi) - f_i^{\pi,\min}(s_0,\pi) &\le \min_d \frac{1}{1-\gamma}\sqrt{\mathscr{C}(d; d_D, \mathcal{F}_i, \pi)}\,\mathcal{O}\left(V_{\max}\sqrt{\frac{\log\frac{|\mathcal{F}||\Pi^{\text{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}\right)\\
&\quad + \frac{1}{1-\gamma}\sum_{(s,\mathbf{a})\in\mathcal{S}\times\mathcal{A}} (d_\pi \setminus d)(s,\mathbf{a})\left[\Delta f_i^\pi(s,\mathbf{a}) - \gamma(P^\pi \Delta f_i^\pi)(s,\mathbf{a})\right].
\end{aligned}
$$

The proof is completed. □

Then, we show that $\mathrm{Gap}^{\Pi^\dagger,\mathrm{EQ}}(\pi)$ is upper bounded by the estimated gap $\widehat{\mathrm{Gap}}_{\mathrm{EQ}}(\pi)$.

Lemma 14. *Under the success event of Lemma 8, for any $\pi \in \Pi$, we have*

$$\mathrm{Gap}^{\Pi^\dagger,\mathrm{EQ}}(\pi) \le \widehat{\mathrm{Gap}}_{\mathrm{EQ}}(\pi) + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.$$

Proof. Let $\pi_i^\dagger = \mathrm{argmax}_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} V_i^{\pi^\dagger}(s_0)$. With Lemma 12, for any player $i \in [m]$, we have with probability at least $1-\delta$,

$$V_i^{\pi_i^\dagger}(s_0) \le \max_{f_i \in \mathcal{F}_i^{\pi_i^\dagger,\beta_f}} f_i(s_0, \pi_i^\dagger) + \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.$$

Recall the definition of $\mathrm{Gap}^{\Pi^\dagger,\mathrm{EQ}}(\pi)$, we obtain

$$\begin{aligned}
\mathrm{Gap}^{\Pi^\dagger,\mathrm{EQ}}(\pi) &= \max_{i\in[m]} \max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} V_i^{\pi^\dagger}(s_0) - V_i^{\pi}(s_0) \\
&\le \max_{i\in[m]} \left(\max_{f_i \in \mathcal{F}_i^{\pi_i^\dagger,\beta_f}} f_i(s_0,\pi_i^\dagger) - V_i^{\pi}(s_0) \right) + \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}. \\
&\le \max_{i\in[m]} \left(\max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} \overline{V}_i^{\pi^\dagger}(s_0) - V_i^{\pi}(s_0) \right) + \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(By definition of } \overline{V}_i^{\pi^\dagger}(s_0)) \\
&\le \max_{i\in[m]} \left(\max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} \overline{V}_i^{\pi^\dagger}(s_0) - \underline{V}_i^{\pi}(s_0) \right) + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} \qquad \text{(By definition of } \underline{V}_i^{\pi}(s_0) \text{ and Lemma 12)} \\
&= \widehat{\mathrm{Gap}}_{\mathrm{EQ}}(\pi) + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.
\end{aligned}$$

□

Now we are ready to prove Theorem 3.

Theorem 3. *With probability at least $1-\delta$, for any $\pi \in \Pi$ and $\mathrm{EQ} \in \{\mathrm{NE}, \mathrm{CE}, \mathrm{CCE}\}$, the output policy $\widehat{\pi}$ of Algorithm 1 satisfies that*

$$\begin{aligned}
\mathrm{Gap}^{\Pi^\dagger,\mathrm{EQ}}(\widehat{\pi}) \le \mathrm{Gap}^{\Pi^\dagger,\mathrm{EQ}}(\pi) &+ \frac{4\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} \\
&+ \max_{i\in[m]} \min_{\widetilde{\pi}_i \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} \left(\Delta_i^{\widetilde{\pi}_i} + \Delta_i^{\pi} + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i) \right),
\end{aligned}$$

where $\mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i) \coloneqq \max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} \overline{V}_i^{\pi^\dagger}(s_0) - \overline{V}_i^{\widetilde{\pi}_i}(s_0)$.

Proof. Let $\pi_i^\dagger = \mathrm{argmax}_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} \overline{V}_i^{\pi^\dagger}(s_0)$. With probability at least $1-\delta$, for each player $i \in [m]$, we upper bound $\overline{V}_i^{\pi_i^\dagger}(s_0) - \underline{V}_i^{\pi}(s_0)$ as

$$\begin{aligned}
\overline{V}_i^{\pi_i^\dagger}(s_0) - \underline{V}_i^{\pi}(s_0) &= \overline{V}_i^{\widetilde{\pi}_i}(s_0) - \underline{V}_i^{\pi}(s_0) + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i) \qquad (\widetilde{\pi}_i \text{ is an arbitrary policy from } \Pi_i^{\mathrm{ext}}(\pi)) \\
&= f_i^{\widetilde{\pi}_i,\max}(s_0) - f_i^{\pi,\min}(s_0) + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i) \\
&\le V_i^{\widetilde{\pi}_i}(s_0) + \Delta_i^{\widetilde{\pi}_i} - f_i^{\pi,\max}(s_0) + \Delta_i^{\pi} + \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(By definition of } \Delta_i^{\widetilde{\pi}_i} \text{ and Lemma 12)}
\end{aligned}$$

$$\leq V_i^{\widetilde{\pi}_i}(s_0) - V_i^{\pi}(s_0) + \Delta_i^{\widetilde{\pi}_i} + \Delta_i^{\pi} + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i) \qquad \text{(By Lemma 12)}$$

$$\leq \max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} V_i^{\pi^\dagger}(s_0) - V_i^{\pi}(s_0) + \Delta_i^{\widetilde{\pi}_i} + \Delta_i^{\pi} + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i).$$

This directly implies that

$$\widehat{\mathrm{Gap}}_{\mathrm{EQ}}(\pi) \leq \mathrm{Gap}^{\Pi^{\dagger,\mathrm{EQ}}}(\pi) + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} + \max_{i\in[m]} \min_{\widetilde{\pi}_i \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} \left(\Delta_i^{\widetilde{\pi}_i} + \Delta_i^{\pi} + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i)\right). \tag{23}$$

By the optimality of $\widehat{\pi}$, for any $\pi \in \Pi$, we have

$$\begin{aligned}
\mathrm{Gap}^{\Pi^{\dagger,\mathrm{EQ}}}(\widehat{\pi}) &\leq \widehat{\mathrm{Gap}}_{\mathrm{EQ}}(\widehat{\pi}) + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} \\
&\leq \widehat{\mathrm{Gap}}_{\mathrm{EQ}}(\pi) + \frac{2\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} \\
&\leq \mathrm{Gap}^{\Pi^{\dagger,\mathrm{EQ}}}(\pi) + \frac{4\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} + \max_{i\in[m]} \min_{\widetilde{\pi}_i \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} \left(\Delta_i^{\widetilde{\pi}_i} + \Delta_i^{\pi} + \mathrm{subopt}_i^{\pi}(\widetilde{\pi}_i)\right). \qquad \text{(By (23))}
\end{aligned}$$

This completes the proof. □

D. Proofs for Section 5

In this section, we prove Theorem 5. We start with some concentration results.

Lemma 15. *With probability at least $1-\delta$, for any $g_1, g_2, h \in \mathcal{G}_i$ and $\pi \in \Pi_i^{\mathrm{ext}}$, we have*

$$\begin{aligned}
&\Bigg| \|g_1 - \mathcal{T}_i^{\pi} h\|_{2,d_S}^2 - \|g_2 - \mathcal{T}_i^{\pi} h\|_{2,d_S}^2 \\
&\quad - \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(g_1(s) - \mathbf{r}_i - \gamma h(s')\right)^2 + \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(g_2(s) - \mathbf{r}_i - \gamma h(s')\right)^2 \Bigg| \\
&\leq 2V_{\max} \|g_1 - g_2\|_{2,d_S} \sqrt{C_A(\pi) \frac{\log \frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n}} + \frac{C_A(\pi) V_{\max}^2 \log \frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n},
\end{aligned}$$

where $C_A(\pi) := \max_{s,\mathbf{a}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}$.

Proof. First, we observe that

$$\begin{aligned}
&\|g_1 - \mathcal{T}_i^{\pi} h\|_{2,d_S}^2 - \|g_2 - \mathcal{T}_i^{\pi} h\|_{2,d_S}^2 \\
&= \mathbb{E}_{s\sim d_S, \mathbf{a}\sim\pi(\cdot|s), s'\sim P(\cdot|s,\mathbf{a})}[(g_1(s) - r_i(s,\mathbf{a}) - \gamma h(s'))^2] - \mathbb{E}_{s\sim d_S, \mathbf{a}\sim\pi(\cdot|s), s'\sim P(\cdot|s,\mathbf{a})}[(g_2(s) - r_i(s,\mathbf{a}) - \gamma h(s'))^2].
\end{aligned}$$

Let random variable $X = \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}(g_1(s_i) - r_i(s_i,\mathbf{a}) - \gamma h(s_i'))^2 - \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}(g_2(s_i) - r_i(s_i,\mathbf{a}) - \gamma h(s_i'))^2$, X is drawn from $d_S \times d_A \times \mathcal{P}$. Then we obtain

$$\begin{aligned}
&\Bigg| \|g_1 - \mathcal{T}_i^{\pi} f\|_{2,d_S}^2 - \|g_2 - \mathcal{T}_i^{\pi} f\|_{2,d_S}^2 \\
&\quad - \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(g_1(s) - \mathbf{r}_i - \gamma h(s')\right)^2 + \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(g_2(s) - \mathbf{r}_i - \gamma h(s')\right)^2 \Bigg| \\
&= \left| \mathbb{E}_{d_S\times\pi\times\mathcal{P}} \left[(g_1(s) - r_i(s,\mathbf{a}) - \gamma h(s'))^2\right] - \mathbb{E}_{d_S\times\pi\times\mathcal{P}} \left[(g_2(s) - r_i(s,\mathbf{a}) - \gamma h(s'))^2\right] - \frac{1}{n}\sum_{i=1}^{n} X_i \right| \\
&\qquad \text{(By definition of } X\text{)}
\end{aligned}$$

Here $\mathbb{E}_{d_S\times d_A\times\mathcal{P}}[X] = \mathbb{E}_{d_S\times\pi\times\mathcal{P}}\left[(g_1(s)-r_i(s,\mathbf{a})-\gamma h(s'))^2\right] - \mathbb{E}_{d_S\times\pi\times\mathcal{P}}\left[(g_2(s)-r_i(s,\mathbf{a})-\gamma h(s'))^2\right]$. For the variance, we have

$$\begin{aligned}
&\mathbb{V}_{d_S\times d_A\times\mathcal{P}}[X] \\
&\leq \mathbb{E}_{d_S\times d_A\times\mathcal{P}}\left[\frac{\pi(\mathbf{a}|s)^2}{d_A(\mathbf{a}|s)^2}(g_1(s)-g_2(s))^2(g_1(s)+g_2(s)-2r_i(s,\mathbf{a})-2\gamma h(s'))^2\right] \\
&\leq 4V_{\max}^2\mathbb{E}_{d_S\times d_A}\left[\frac{\pi(\mathbf{a}|s)^2}{d_A(\mathbf{a}|s)^2}(g_1(s)-g_2(s))^2\right] \\
&\leq 4V_{\max}^2\max_{s,\mathbf{a}}\frac{\pi(\mathbf{a}|s)^2}{d_A(\mathbf{a}|s)}\mathbb{E}_{d_S}\left[(g_1(s)-g_2(s))^2\right]
\end{aligned}$$

Let $C_A(\pi) \coloneqq \max_{s,\mathbf{a}}\frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}$. By Freedman's inequality and union bound, we have with probability at least $1-\delta$,

$$\left\|\mathbb{E}[X]-\frac{1}{n}\sum_{i=1}^n X_i\right\| \leq \sqrt{\frac{4V_{\max}^2 C_A(\pi)\|g_1-g_2\|_{d_S}^2\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}} + \frac{C_A(\pi)V_{\max}^2\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}.$$

□

For any player $i\in[m]$ and $\pi\in\Pi_i^{\text{ext}}$, let us define

$$g_i^\pi \coloneqq \operatorname*{argmin}_{g\in\mathcal{G}_i}\sup_{\text{admissible } d}\|g-\mathcal{T}_i^\pi g\|_{2,d}^2 \tag{24}$$

$$h_i^\pi \coloneqq \operatorname*{argmin}_{h\in\mathcal{G}_i}\frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}}\frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}\left(h(s)-\mathbf{r}_i-\gamma g_i^\pi(s')\right)^2. \tag{25}$$

We bound $\|g_i^\pi-h_i^\pi\|_{2,d_S}$ as follows.

Lemma 16. *Let g_i^π and h_i^π be defined as in Equations (24) and (25). Under the success event of Lemma 15, for any player $i\in[m]$ and $\pi\in\Pi_i^{\text{ext}}$, we have*

$$\|g_i^\pi-h_i^\pi\|_{2,d_S} \leq 6V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}} + 2\sqrt{\varepsilon_{\mathcal{F}}}.$$

The proof is to invoke Lemma 15 for g_i^π and h_i^π and the calculation is the same as Lemma 9. Similar to Lemma 10, we show that the best approximation of V_i^π is contained in $\mathcal{G}_i^{\pi,\beta_g}$.

Lemma 17. *Under the success event of Lemma 15, for any player $i\in[m]$ and $\pi\in\Pi_i^{\text{ext}}$, the following inequality for $\mathcal{E}_i(g_i^\pi,\pi;\mathcal{D})$ holds*

$$\mathcal{E}_i(g_i^\pi,\pi;\mathcal{D}) \leq \frac{80C_A(\pi)V_{\max}^2\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n} + 30\varepsilon_{\mathcal{F}} =: \beta_g.$$

Proof. Applying Lemma 15 and Lemma 16, we obtain

$$\begin{aligned}
&\Bigg|\|g_i^\pi-\mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}^2 - \|h_i^\pi-\mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}^2 - \\
&\frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}}\frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}\left(g_i^\pi(s)-\mathbf{r}_i-\gamma g_i^\pi(s')\right)^2 + \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}}\frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}\left(h_i^\pi(s)-\mathbf{r}_i-\gamma g_i^\pi(s')\right)^2\Bigg| \\
&\leq 4V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F}}} + \frac{13C_A(\pi)V_{\max}^2\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}.
\end{aligned} \tag{26}$$

Similar to Lemma 10, we bound $\|g_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}^2 - \|h_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}^2$ as follows,

$$\begin{aligned}
&\|g_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}^2 - \|h_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}^2 \\
\leq\ & \left(\|g_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S} + \|h_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}\right) \left|\|g_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S} - \|h_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}\right| \\
\leq\ & 36V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F}}} + 36V_{\max}^2\frac{C_A(\pi)\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n} + 8\varepsilon_{\mathcal{F}}. \qquad \text{(By Lemma 16)}
\end{aligned}$$

Combining this with (26), we get

$$\begin{aligned}
&\frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}}\frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}\left(g_i^\pi(s) - \mathbf{r}_i - \gamma g_i^\pi(s')\right)^2 - \frac{1}{n}\sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}}\frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)}\left(h_i^\pi(s) - \mathbf{r}_i - \gamma g_i^\pi(s')\right)^2 \\
\leq\ & \|g_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}^2 - \|h_i^\pi - \mathcal{T}_i^\pi g_i^\pi\|_{2,d_S}^2 + 4V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}\varepsilon_{\mathcal{F}}} + \frac{13C_A(\pi)V_{\max}^2\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n} \\
\leq\ & \frac{80C_A(\pi)V_{\max}^2\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n} + 30\varepsilon_{\mathcal{F}}. \qquad \text{(By AM-GM inequality)}
\end{aligned}$$

□

We then prove that $g_i^{\pi,\max}(s_0)$ and $g_i^{\pi,\min}(s_0)$ are the upper bound and the lower bound on the value function $V_i^\pi(s_0)$ respectively.

Lemma 18. *Under the success event of Lemma 15, for any player $i \in [m]$ and any $\pi \in \Pi_i^{\text{ext}}$, the following two inequalities hold*

$$\begin{aligned}
g_i^{\pi,\max}(s_0) &\geq V_i^\pi(s_0) - \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} \\
g_i^{\pi,\min}(s_0) &\leq V_i^\pi(s_0) + \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.
\end{aligned}$$

Proof. Let g_i^π be defined as in (24), by invoking Lemma 21, we get

$$\begin{aligned}
|g_i^\pi(s_0) - V_i^\pi(s_0)| &\leq \frac{\mathbb{E}_{s,\mathbf{a}\sim d^\pi, s'\sim P(\cdot|s,\mathbf{a})}\left[g(s) - r_i(s,\mathbf{a}) - \gamma g(s')\right]}{1-\gamma} \\
&\leq \frac{\|g - \mathcal{T}_i^\pi g\|_{2,d^\pi}}{1-\gamma} \leq \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.
\end{aligned}$$

By Lemma 17, we know that $g_i^\pi \in \mathcal{G}_i^{\pi,\beta_g}$. Then, we obtain

$$g_i^{\pi,\max}(s_0) \geq g_i^\pi(s_0) \geq V_i^\pi(s_0) - \frac{\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma}.$$

The case for $g_i^{\pi,\min}$ is similar. □

We now show that $\mathcal{E}_i(g_i, \pi; \mathcal{D})$ could effectively estimate $\|g_i - \mathcal{T}_i^\pi g_i\|_{2,d_S}^2$.

Lemma 19. *Under the success event of Lemma 15, for any player $i \in [m]$ and any $\pi \in \Pi_i^{\text{ext}}$, given $\varepsilon > 0$, if $g_i \in \mathcal{G}_i$ satisfies that $\mathcal{E}_i(g_i, \pi; \mathcal{D}) \leq \varepsilon$, we have*

$$\|g_i - \mathcal{T}_i^\pi g_i\|_{2,d_S} \leq 8V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}} + 4\sqrt{\varepsilon_{\mathcal{F},\mathcal{F}}} + \sqrt{\varepsilon}.$$

Proof. Let h_i^π be defined as in (25), let us define

$$g_{i,d_S}^\pi \coloneqq \operatorname*{argmin}_{g_i' \in \mathcal{G}_i} \|g_i' - \mathcal{T}_i^\pi g_i\|_{2,d_S}^2 .$$

Similar to Lemma 13, we first upper bound $\|h_i^\pi - \mathcal{T}_i^\pi g_i\|_{2,d_S}$. By invoking Lemma 15, we obtain,

$$\begin{aligned}
&\Bigg| \|h_i^\pi - \mathcal{T}_i^\pi g_i\|_{2,d_S}^2 - \left\|g_{i,d_S}^\pi - \mathcal{T}_i^\pi g_i\right\|_{2,d_S}^2 - \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(h_i^\pi(s) - \mathbf{r}_i - \gamma g_i(s')\right)^2 \\
&+ \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(g_{i,d_S}^\pi(s) - \mathbf{r}_i - \gamma g_i(s')\right)^2 \Bigg| \\
&\le 2V_{\max} \|h_i^\pi - g_{i,d_S}^\pi\|_{2,d_S} \sqrt{C_A(\pi) \frac{\log \frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n}} + \frac{C_A(\pi) V_{\max}^2 \log \frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n}.
\end{aligned}$$

Rearranging the terms and by similar calculation to Lemma 13, we have

$$\begin{aligned}
&\|h_i^\pi - \mathcal{T}_i^\pi g_i\|_{2,d_S}^2 \\
&\le 2V_{\max} \|h_i^\pi - \mathcal{T}_i^\pi g_i\|_{2,d_S} \sqrt{C_A(\pi) \frac{\log \frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n}} + \frac{2C_A(\pi) V_{\max}^2 \log \frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n} + 2\varepsilon_{\mathcal{F},\mathcal{F}}.
\end{aligned} \tag{27}$$

By solving (27), we get

$$\|h_i^\pi - \mathcal{T}_i^\pi g_i\|_{2,d_S} \le 3V_{\max} \sqrt{C_A(\pi) \frac{\log \frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n}} + \sqrt{2\varepsilon_{\mathcal{F},\mathcal{F}}}. \tag{28}$$

Then, we invoke Lemma 15 for g_i^π and have

$$\begin{aligned}
&\Bigg| \|g_i - \mathcal{T}_i^\pi g_i\|_{2,d_S}^2 - \|g_i^\pi - \mathcal{T}_i^\pi g_i\|_{2,d_S}^2 \\
&- \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(g_i(s) - \mathbf{r}_i - \gamma g_i(s')\right)^2 + \frac{1}{n} \sum_{(s,\mathbf{a},\mathbf{r},s')\in\mathcal{D}} \frac{\pi(\mathbf{a}|s)}{d_A(\mathbf{a}|s)} \left(g_i^\pi(s) - \mathbf{r}_i - \gamma g_i(s')\right)^2 \Bigg| \\
&\le 2V_{\max} \|g_i - \mathcal{T}_i^\pi g_i\|_{2,d_S} \sqrt{C_A(\pi) \frac{\log \frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n}} + 3V_{\max} \sqrt{C_A(\pi) \frac{\log \frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n} \varepsilon_{\mathcal{F},\mathcal{F}}} + \frac{7C_A(\pi) V_{\max}^2 \log \frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n}.
\end{aligned}$$

With similar calculation to Lemma 13, we arrange the terms and have

$$\begin{aligned}
\|g_i - \mathcal{T}_i^\pi g_i\|_{2,d_S}^2 &= 2V_{\max} \|g_i - \mathcal{T}_i^\pi g_i\|_{2,d_S} \sqrt{C_A(\pi) \frac{\log \frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n}} + 12V_{\max} \sqrt{C_A(\pi) \frac{\log \frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n} \varepsilon_{\mathcal{F},\mathcal{F}}} \\
&\quad + \frac{16C_A(\pi) V_{\max}^2 \log \frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n} + 2\varepsilon_{\mathcal{F},\mathcal{F}} + \varepsilon.
\end{aligned} \tag{29}$$

Solving (29) and using AM-GM inequality finishes the proof. □

Now we are ready to prove Theorem 5.

Theorem 5 (V-type guarantee)**.** *With probability at least* $1-\delta$*, for any* $\pi \in \Pi$ *and* $\mathrm{EQ} \in \{\mathrm{NE}, \mathrm{CE}, \mathrm{CCE}\}$*, the output policy* $\widehat{\pi}$ *from (9) satisfies that*

$$\begin{aligned}
\mathrm{Gap}^{\Pi^{\dagger,\mathrm{EQ}}}(\widehat{\pi}) &\le \mathrm{Gap}^{\Pi^{\dagger,\mathrm{EQ}}}(\pi) + \frac{4\sqrt{\varepsilon_{\mathcal{F}}}}{1-\gamma} \\
&\quad + \max_{i\in[m]} \min_{\widetilde{\pi}_i \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} \left(\Delta_i^{\widetilde{\pi}_i} + \Delta_i^\pi + \mathrm{subopt}_i^\pi(\widetilde{\pi}_i)\right),
\end{aligned}$$

where $\Delta_i^\pi = g_i^{\pi,\max}(s_0) - g_i^{\pi,\min}(s_0)$ *and* $\mathrm{subopt}_i^{\widetilde{\pi}_i} = \max_{\pi^\dagger \in \Pi_i^{\dagger,\mathrm{EQ}}(\pi)} g_i^{\pi^\dagger,\max}(s_0) - g_i^{\widetilde{\pi}_i,\max}(s_0)$. *In addition, with probability at least* $1-\delta$*, for any player* $i \in [m]$ *and any* $\pi \in \Pi_i^{\mathrm{ext}}$*, we have*

$$\Delta_i^\pi \le \min_{d \in \Delta(S)} \frac{1}{1-\gamma}\sqrt{\mathscr{C}(d; d_S, \mathcal{G}_i, \pi)}\varepsilon_{\mathrm{apx}} + \frac{1}{1-\gamma} \sum_{s\in\mathcal{S}} (d^\pi \setminus d)(s)\left[\Delta g_i^\pi(s) - \gamma(P_i^\pi \Delta g_i^\pi)(s)\right],$$

where $\varepsilon_{\mathrm{apx}} = \mathcal{O}\left(V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}\right)$, $(d^\pi \setminus d)(s) := \max(d^\pi(s) - d(s), 0)$, $\Delta g_i^\pi(s) := g_i^{\pi,\max}(s) - g_i^{\pi,\min}(s)$, *and* $(P_i^\pi g)(s) = \mathbb{E}_{\mathbf{a}\sim\pi(\cdot|s), s'\sim P(\cdot|s,\mathbf{a})}[r_i(s,\mathbf{a}) + g(s')]$.

Proof. The proof for the first part is the same as Theorem 3. For the second part, we invoke Lemma 21 for $g_i^{\pi,\min}$ and $g_i^{\pi,\max}$

$$\begin{aligned}
& g_i^{\pi,\max}(s_0) - g_i^{\pi,\min}(s_0) \\
&= \frac{1}{1-\gamma}\underbrace{\mathbb{E}_d\left[(g_i^{\pi,\max} - \mathcal{T}_i^\pi g_i^{\pi,\max}) - \left(g_i^{\pi,\min} - \mathcal{T}_i^\pi g_i^{\pi,\min}\right)\right]}_{\text{(I)}} \\
&\quad + \frac{1}{1-\gamma}\underbrace{(\mathbb{E}_{d_\pi}\left[\Delta g_i^\pi - \gamma P_i^\pi \Delta g_i^\pi\right] - \mathbb{E}_d\left[\Delta g_i^\pi - \gamma P_i^\pi \Delta g_i^\pi\right])}_{\text{(II)}}, \qquad (\Delta g_i^\pi := g_i^{\pi,\max} - g_i^{\pi,\min})
\end{aligned}$$

where $d \in \Delta(\mathcal{S})$ is an arbitrary distribution. For the term (I), we have

$$\begin{aligned}
\text{(I)} &\le \left|\mathbb{E}_d\left[(g_i^{\pi,\max} - \mathcal{T}_i^\pi g_i^{\pi,\max})\right]\right| + \left|\mathbb{E}_d\left[\left(g_i^{\pi,\min} - \mathcal{T}_i^\pi g_i^{\pi,\min}\right)\right]\right| \\
&\le \|g_i^{\pi,\max} - \mathcal{T}_i^\pi g_i^{\pi,\max}\|_{2,d} + \|g_i^{\pi,\min} - \mathcal{T}_i^\pi g_i^{\pi,\min}\|_{2,d} \qquad \text{(By Jensen's inequality)} \\
&\le \sqrt{\mathscr{C}(d; d_S, \mathcal{G}_i, \pi)}\left(\|g_i^{\pi,\max} - \mathcal{T}_i^\pi g_i^{\pi,\max}\|_{2,d_S} + \|g_i^{\pi,\min} - \mathcal{T}_i^\pi g_i^{\pi,\min}\|_{2,d_S}\right).
\end{aligned}$$

Recall that $\beta_g = \frac{80 C_A(\pi) V_{\max}^2 \log\frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n} + 30\varepsilon_{\mathcal{F}}$. We invoke Lemma 13 and obtain with probability at least $1-\delta$

$$\text{(I)} \le \sqrt{\mathscr{C}(d; d_S, \mathcal{G}_i, \pi)}\mathcal{O}\left(V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}\right). \tag{30}$$

For term (II), we have

$$\begin{aligned}
\text{(II)} &\le \sum_{s\in\mathcal{S}} (d_\pi \setminus d)(s)\left[\Delta g_i^\pi(s) - \gamma(P_i^\pi \Delta g_i^\pi)(s)\right] \\
&\quad + \sum_{(s)\in\mathcal{S}} \mathbb{I}(d(s) > d_\pi(s))\left[d(s) - d_\pi(s)\right]\left|\Delta g_i^\pi(s) - \gamma(P_i^\pi \Delta g_i^\pi)(s)\right| \\
&\le \sum_{s\in\mathcal{S}} (d_\pi \setminus d)(s)\left[\Delta g_i^\pi(s) - \gamma(P_i^\pi \Delta g_i^\pi)(s)\right] \\
&\quad + \mathbb{E}_d\left[|g_i^{\pi,\max} - \mathcal{T}_i^\pi g_i^{\pi,\max}| + \left|g_i^{\pi,\min} - \mathcal{T}_i^\pi g_i^{\pi,\min}\right|\right] \\
&\le \sum_{s\in\mathcal{S}} (d_\pi \setminus d)(s)\left[\Delta g_i^\pi(s) - \gamma(P_i^\pi \Delta g_i^\pi)(s)\right] \\
&\quad + \sqrt{\mathscr{C}(d; d_S, \mathcal{G}_i, \pi)}\mathcal{O}\left(V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\mathrm{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}\right).
\end{aligned} \tag{31}$$

The last step is from the analysis of term (I). Combining (30) and (31), we get

$$g_i^{\pi,\max}(s_0) - g_i^{\pi,\min}(s_0) \leq \min_d \frac{1}{1-\gamma}\sqrt{\mathscr{C}(d; d_S, \mathcal{G}_i, \pi)}\mathcal{O}\left(V_{\max}\sqrt{C_A(\pi)\frac{\log\frac{|\mathcal{G}||\Pi^{\text{ext}}|}{\delta}}{n}} + \sqrt{\varepsilon_{\mathcal{F}} + \varepsilon_{\mathcal{F},\mathcal{F}}}\right)$$
$$+ \frac{1}{1-\gamma}\sum_{s\in\mathcal{S}}(d_\pi \setminus d)(s)\left[\Delta g_i^\pi(s) - \gamma(P_i^\pi \Delta g_i^\pi)(s)\right].$$

This completes the proof. □

E. Auxiliary Lemmas

Lemma 20 (Q-function Evaluation Error Lemma)**.** *For any player $i \in [m]$ and any $\pi \in \Pi_i^{\text{ext}}$, and any $f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$*

$$f(s_0,\pi) - V_i^\pi(s_0) = \frac{\mathbb{E}_{s,\mathbf{a}\sim d^\pi, s'\sim P(\cdot|s,\mathbf{a})}\left[f(s,\mathbf{a}) - r_i(s,\mathbf{a}) - \gamma f(s',\pi)\right]}{1-\gamma}$$

Proof. We observe that

$$\sum_{s,\mathbf{a}}\sum_{t=0}^{\infty}\gamma^{t+1}\Pr(s_t = s, \mathbf{a}_t = \mathbf{a}|s_0,\pi)\sum_{s'}\Pr(s_{t+1} = s'|s_t = s, \mathbf{a}_t = \mathbf{a})f(s',\pi)$$
$$= \sum_{s,\mathbf{a}}\sum_{t=1}^{\infty}\gamma^{t}\Pr(s_t = s, \mathbf{a}_t = \mathbf{a}|s_0,\pi)f(s,\mathbf{a})$$

Then, we have

$$\frac{\mathbb{E}_{s,\mathbf{a}\sim d^\pi, s'\sim P(\cdot|s,\mathbf{a})}\left[f(s,\mathbf{a}) - \gamma f(s',\pi)\right]}{1-\gamma}$$
$$= \sum_{s,\mathbf{a}}\sum_{t=0}^{\infty}\gamma^{t}\Pr(s_t = s, \mathbf{a}_t = \mathbf{a}|s_0,\pi)f(s,\mathbf{a}) - \sum_{s,\mathbf{a}}\sum_{t=1}^{\infty}\gamma^{t}\Pr(s_t = s, \mathbf{a}_t = \mathbf{a}|s_0,\pi)f(s,\mathbf{a})$$
$$= \sum_{\mathbf{a}}\Pr(\mathbf{a}_0 = \mathbf{a}|s_0,\pi)f(s_0,\mathbf{a}) = f(s_0,\pi).$$

Since $V_i^\pi(s_0) = \frac{\mathbb{E}_{d^\pi}[r_i(s,\mathbf{a})]}{1-\gamma}$, rearranging the terms finishes the proof. □

Lemma 21 (Value Function Evaluation Error Lemma)**.** *For any player $i \in [m]$ and any $\pi \in \Pi_i^{\text{ext}}$, and any $f \in \mathbb{R}^{\mathcal{S}}$*

$$f(s_0) - V_i^\pi(s_0) = \frac{\mathbb{E}_{s,\mathbf{a}\sim d^\pi, s'\sim P(\cdot|s,\mathbf{a})}\left[f(s) - r_i(s,\mathbf{a}) - \gamma f(s')\right]}{1-\gamma}$$

Proof. We observe that

$$\sum_{s,\mathbf{a}}\sum_{t=0}^{\infty}\gamma^{t+1}\Pr(s_t = s, \mathbf{a}_t = \mathbf{a}|s_0,\pi)\sum_{s'}\Pr(s_{t+1} = s'|s_t = s, \mathbf{a}_t = \mathbf{a})f(s')$$
$$= \sum_{s,\mathbf{a}}\sum_{t=1}^{\infty}\gamma^{t}\Pr(s_t = s, \mathbf{a}_t = \mathbf{a}|s_0,\pi)f(s)$$

Then, we have

$$\frac{\mathbb{E}_{s\sim d^\pi, s'\sim P(\cdot|s,\mathbf{a})}\left[f(s) - \gamma f(s')\right]}{1-\gamma}$$
$$= \sum_{s,\mathbf{a}}\sum_{t=0}^{\infty}\gamma^{t}\Pr(s_t = s, \mathbf{a}_t = \mathbf{a}|s_0,\pi)f(s) - \sum_{s,\mathbf{a}}\sum_{t=1}^{\infty}\gamma^{t}\Pr(s_t = s, \mathbf{a}_t = \mathbf{a}|s_0,\pi)f(s)$$

$$= \sum_{\mathbf{a}} \Pr(\mathbf{a}_0 = \mathbf{a} | s_0, \pi) f(s_0) = f(s_0).$$

Since $V_i^\pi(s_0) = \frac{\mathbb{E}_{d^\pi}[r_i(s,\mathbf{a})]}{1-\gamma}$, rearranging the terms finishes the proof. □