Last-iterate Convergence of Decentralized Optimistic Gradient Descent/Ascent in Infinite-horizon Competitive Markov Games

Chen-Yu Wei
Chung-Wei Lee
Mengxiao Zhang
Haipeng Luo

University of Southern California

Editors: Mikhail Belkin and Samory Kpotufe

Abstract
We study infinite-horizon discounted two-player zero-sum Markov games, and develop a decentralized algorithm that provably converges to the set of Nash equilibria under self-play. Our algorithm is based on running an Optimistic Gradient Descent Ascent algorithm on each state to learn the policies, with a critic that slowly learns the value of each state. To the best of our knowledge, this is the first algorithm in this setting that is simultaneously rational (converging to the opponent’s best response when it uses a stationary policy), convergent (converging to the set of Nash equilibria under self-play), agnostic (no need to know the actions played by the opponent), symmetric (players taking symmetric roles in the algorithm), and enjoying a finite-time last-iterate convergence guarantee, all of which are desirable properties of decentralized algorithms.

1. Introduction
Multi-agent reinforcement learning studies how multiple agents should interact with each other and the environment, and has wide applications in, for example, playing board games (Silver et al., 2017) and real-time strategy games (Vinyals et al., 2019). To model these problems, the framework of Markov games (also called stochastic games) (Shapley, 1953) is often used, which can be seen as a generalization of Markov Decision Processes (MDPs) from a single agent to multiple agents. In this work, we focus on one fundamental class: two-player zero-sum Markov games.

In this setting, there are many centralized algorithms developed in a line of recent works for finding a Nash equilibrium (Wei et al., 2017; Sidford et al., 2020; Xie et al., 2020; Bai and Jin, 2020; Zhang et al., 2020a; Liu et al., 2021). These algorithms require a central controller that collects some global knowledge (such as the actions and the rewards of all players) and then jointly decides the policies for all players. Centralized algorithms are usually convergent (as defined in (Bowling and Veloso, 2001)), in the sense that the policies of the players converge to the set of Nash equilibria.

On the other hand, there is also a surge of studies on decentralized algorithms that run independently on each player, requiring only local information such as the player’s own action and the corresponding reward feedback (Zhang et al., 2019; Bai et al., 2020; Tian et al., 2021; Liu et al., 2020; Daskalakis et al., 2020). Compared to centralized ones, decentralized algorithms are usually more versatile and can potentially run in different environments (cooperative or competitive). Many of them enjoy the property of being rational (as defined in (Bowling and Veloso, 2001)), in the
sense that a player’s policy converges to the best response to the opponent no matter what stationary policy the opponent uses. However, it is also often more challenging to show the convergence to a Nash equilibrium when the two players execute the same decentralized algorithm.

It can be seen that a rational algorithm has different benefits compared to a convergent algorithm – the former satisfies individual player’s interests, while the latter might be better for achieving social good. Therefore, a single algorithm that possesses both properties is highly desirable. For example, in a market where “enforcing” all traders to follow the same rule is difficult, but “recommending” them to use a specific algorithm is possible, a rational and convergent algorithm would be a good candidate — if all traders follow the recommendation, then a social equilibrium is quickly attained; otherwise, those who follow the recommendation are still satisfied because they best respond to a stationary environment.

Based on this motivation, our main contribution is to develop the first decentralized algorithm that is simultaneously rational, last-iterate convergent (with a concrete finite-time guarantee),\(^1\) agnostic, and symmetric (more details to follow in Section 1.1) for two-player zero-sum Markov games. Our algorithm is based on Optimistic Gradient Descent/Ascent (OGDA) (Chiang et al., 2012; Rakhlin and Sridharan, 2013) and importantly relies on a critic that slowly learns a certain value function for each state. Following previous works on learning MDPs (Abbasi-Yadkori et al., 2019; Agarwal et al., 2020) or Markov games (Perolat et al., 2018), we present the convergence guarantee in terms of the number of iterations of the algorithm and the estimation error of some gradient information (along with other problem-dependent constants), where the estimation error can be zero in a full-information setting, or goes down to zero fast enough with additional structural assumptions (e.g. every stationary policy pair induces an irreducible Markov chain, similar to (Auer and Ortner, 2007)).

While the OGDA algorithm, first studied in (Popov, 1980) under a different name, has been extensively used in recent years for learning matrix games (a special case of Markov games with one state), to the best of our knowledge, no previous work has applied it to learning Markov games and derived a concrete last-iterate convergence rate. Several recent works derive last-iterate convergence of OGDA for matrix games (Hsieh et al., 2019; Liang and Stokes, 2019; Mokhtari et al., 2020; Golowich et al., 2020; Wei et al., 2021), and our analysis is heavily inspired by the approach of (Wei et al., 2021). However, the extension to infinite-horizon Markov games is highly non-trivial as there is additional “instability penalty” in the system that we need to handle; see Section 4 for detailed discussions.

\[\text{1.1. Related Work}\]

In this section, we discuss and compare related works on learning two-player zero-sum Markov games. We refer the readers to a thorough survey by (Zhang et al., 2020b) for other topics in multi-agent reinforcement learning.

Shapley (1953) first introduces the Markov game model and proposes an algorithm analogous to value iteration for solving two-player zero-sum Markov games (with all parameters known). Later, Hoffman and Karp (1966) propose a policy iteration algorithm, and Pollatschek and Avi-Itzhak (1969) propose another policy iteration variant that works better in practice but cannot always  

\(^1\) Note that while average-iterate convergence is possible (and standard) for stateless convex-concave games (e.g. (Syrgkanis et al., 2015)), it does not work for Markov games since the problem is nonconvex-nonconcave in the space of policies (Daskalakis et al., 2020).
converge. With the efforts of Van Der Wal (1978) and Filar and Tolwinski (1991), a slight variant of the (Pollatschek and Avi-Itzhak, 1969) algorithm is proposed in (Filar and Tolwinski, 1991) and proven to converge. In such a full-information setting where all parameters are known, our algorithm has no estimation error and can also be viewed as a new policy-iteration algorithm.

Littman (1994) initiates the study of competitive reinforcement learning under the framework of Markov games and proposes an extension of the single-player Q-learning algorithm, called minimax-Q, which is later proven to converge under some conditions (Szepesvári and Littman, 1999). While minimax-Q can run in a decentralized manner, it is conservative and only converges to the minimax policy but not the best response to the opponent.

To fix this issue, the work of Bowling and Veloso (2001) argues that a desirable multi-agent learning algorithm should have the following two properties simultaneously: rational and convergent. By their definition, a rational algorithm converges to its opponent’s best response if the opponent converges to a stationary policy, while a convergent algorithm converges to a Nash equilibrium if both agents use it. They propose the WoLF (Win-or-Learn-Fast) algorithm to achieve this goal, albeit only with empirical evidence. Subsequently, Conitzer and Sandholm (2007); Perolat et al. (2018); Sayin et al. (2020) design decentralized algorithms that provably enjoy these two properties, but only with asymptotic guarantees.

Recently, there is a surge of works that provide finite-time guarantees and characterize the tight sample complexity for finding Nash equilibria (Perolat et al., 2015; Pérolat et al., 2016; Wei et al., 2017; Sidford et al., 2020; Xie et al., 2020; Zhang et al., 2020a; Bai and Jin, 2020; Liu et al., 2021). These algorithms are all essentially centralized. Below, we focus on comparisons with several recent works that propose decentralized algorithms and provide finite-time guarantees.

Comparison with R-Max (Brafman and Tennenholtz, 2002), UCSG-online (Wei et al., 2017) and OMNI-VI-online (Xie et al., 2020) These algorithms, like minimax-Q, converge to the minimax policy instead of the best response to the opponent, even when the opponent is weak (i.e., not using its best policy). In other words, these algorithms are not rational. Another drawback of these algorithms is that the learner has to observe the actions taken by the opponent. Our algorithm, on the other hand, is both rational and agnostic to what the opponent plays.

Comparison with Optimistic Nash V-Learning (Bai et al., 2020; Tian et al., 2021) The Optimistic Nash V-Learning algorithm handles the finite-horizon tabular case. It runs an exponential-weight algorithm on each state, with importance-weighted loss/reward estimators. It is unclear whether the dynamics of Optimistic Nash V-Learning leads to last iterate convergence. After training, however, Optimistic Nash V-Learning can output a near-optimal non-Markovian policy with size linear in the training time. In contrast, our algorithm exhibits last-iterate convergence, and the output is a simple Markovian policy.

Comparison with Smooth-FSP (Liu et al., 2020) The Smooth-FSP algorithm handles the function approximation setting. The objective function it optimizes is the original objective plus an entropy regularization term. Because of this additional regularization, the players are only guaranteed to converge to some neighborhood of the minimax policy pair (with a constant radius), even when their gradient estimation error is zero. In contrast, our algorithm converges to the true minimax policy pair when the gradient estimation error goes to zero.

---

2. It is tempting to consider an even stronger rationality notion, that is, having no regret against an arbitrary opponent. This is, however, known to be computationally hard (Radanovic et al., 2019; Bai et al., 2020).
Comparison with Independent PG (Daskalakis et al., 2020) Daskalakis et al. (2020) studies independent policy gradient in the tabular case. To achieve last-iterate convergence, the two players have to use asymmetric learning rates, and only the one with a smaller learning rate converges to the minimax policy. In contrast, the two players of our algorithm are completely symmetric, and they simultaneously converge to the equilibrium set.

2. Preliminaries

We consider a two-player zero-sum discounted Markov game defined by a tuple \((\mathcal{S}, \mathcal{A}, \mathcal{B}, \sigma, p, \gamma)\), where: 1) \(\mathcal{S}\) is a finite state space; 2) \(\mathcal{A}\) and \(\mathcal{B}\) are finite action spaces for Player 1 and Player 2 respectively; 3) \(\sigma\) is the loss (payoff) function for Player 1 (Player 2), with \(\sigma(s, a, b) \in [0, 1]\) specifying how much Player 1 pays to Player 2 if they are at state \(s\) and select actions \(a\) and \(b\) respectively; 4) \(p: \mathcal{S} \times \mathcal{A} \times \mathcal{B} \rightarrow \Delta_{\mathcal{S}}\) is the transition function, with \(p(s'|s, a, b)\) being the probability of transitioning to state \(s'\) after actions \(a\) and \(b\) are taken by the two players respectively at state \(s\) (\(\Delta_{\mathcal{S}}\) denotes the set of probability distributions over \(\mathcal{S}\)); 5) and \(\frac{1}{2} \leq \gamma < 1\) is a discount factor.\(^3\)

A stationary policy of Player 1 can be described by a function \(\pi: \mathcal{S} \rightarrow \Delta_{\mathcal{A}}\) that maps each state to an action distribution. We use \(x^s \in \Delta_{\mathcal{A}}\) to denote the action distribution for Player 1 on state \(s\), and use \(x = \{x^s\}_{s \in \mathcal{S}}\) to denote the complete policy. We define \(y^s\) and \(y = \{y^s\}_{s \in \mathcal{S}}\) similarly for Player 2. For notational convenience, we further define \(z^s = (x^s, y^s) \in \Delta_{\mathcal{A}} \times \Delta_{\mathcal{B}}\) as the concatenated policy of the players on state \(s\), and let \(z = \{z^s\}_{s \in \mathcal{S}}\).

For a pair of stationary policies \((x, y)\) and an initial state \(s\), the expected discounted value that the players pay/gain can be represented as

\[
V^s_{x,y} = \mathbb{E} \left[ \sum_{t=1}^{\infty} \gamma^{t-1} \sigma(s_t, a_t, b_t) \right| s_1 = s, \quad a_t \sim x^s, b_t \sim y^s, s_{t+1} \sim p(.|s_t, a_t, b_t), \forall t \geq 1 \right].
\]

The minimax game value on state \(s\) is then defined as

\[
V^s_* = \min_x \max_y V^s_{x,y} = \max_y \min_x V^s_{x,y}.
\]

It is known that a pair of stationary policies \((x_*, y_*)\) attaining the minimax value on state \(s\) is necessarily attaining the minimax value on all states (Filar and Vrieze, 2012), and we call such \(x_*\) a minimax policy, such \(y_*\) a maximin policy, and such a pair a Nash equilibrium. Further define \(X^*_* = \{x^s_* \in X^s_* : x^s_*\) is a minimax policy\} and similarly \(Y^*_* = \{y^s_* \in Y^s_* : y^s_*\) is a maximin policy\}, and denote \(Z^*_* = X^*_* \times Y^*_*\). It is also known that any \(x = \{x^s\}_{s \in \mathcal{S}}\) with \(x^s \in X^*_*\) for all \(s\) is a minimax policy (similarly for \(y\)) (Filar and Vrieze, 2012).

For any \(x^s\), we denote its distance from \(X^*_*\) as \(\text{dist}_x(x^s) = \min_{x^* \in X^*_*} \|x^s - x^*\|\), where \(\|v\|\) for a vector \(v\) denotes its \(L_2\) norm throughout the paper; similarly, \(\text{dist}_y(y^s) = \min_{y^* \in Y^*_*} \|y^s - y^*\|\) and \(\text{dist}_z(z^s) = \min_{z^* \in Z^*_*} \|z^s - z^*\| = \sqrt{\text{dist}_x^2(z^s) + \text{dist}_y^2(z^s)}\).\(^4\) The projection operator for a convex set \(\mathcal{U}\) is defined as \(\Pi_{\mathcal{U}}(v) = \arg\min_{u \in \mathcal{U}} \|u - v\|\).

\(^3\) The discount factor is usually some value close to 1, so we assume that it is no less than \(\frac{1}{2}\) for simplicity. Also note that we consider the discounted setting instead of the finite-horizon episodic setting because the former captures more challenges of this problem (and is also the original setting considered in (Bowling and Veloso, 2001)). Indeed, in the episodic setting where states have a layered structure, convergence can be directly shown in a layer-by-layer manner; see (Lee et al., 2020), an early version of (Wei et al., 2021).

\(^4\) Note the slight abuse of notation here: the meaning of \(\text{dist}_\ast(\cdot)\) depends on its input.
We also define the \textit{Q-function} on state \( s \) under policy pair \((x, y)\) as
\[
Q_s^{x,y}(a, b) = \sigma(s, a, b) + \gamma \mathbb{E}_{s' \sim P(s, a, b)} \left[ V_{s', y}^{x'} \right],
\]
which can be compactly written as a matrix \( Q_s^{x,y} \in \mathbb{R}^{|A| \times |B|} \) such that \( V_{s,y}^{x} = x^\top Q_{s,y}^{x,y} \). We write \( Q_s^* = Q_{s,x,y}^* \) for any minimax/mimimax policy pair \((x_*, y_*)\) (which is unique even if \((x_*, y_*)\) is not). Finally, \( \|Q\| \) for a matrix \( Q \) is defined as \( \max_{i,j} |Q_{i,j}| \).

\textbf{Optimistic Gradient Descent Ascent (OGDA)}

As mentioned, our algorithm is based on running an instance of the OGDA algorithm on each state with an appropriate loss/reward function. To this end, here, following the exposition of (Wei et al., 2021) we briefly review OGDA for a matrix game defined by a matrix \( Q \in \mathbb{R}^{|A| \times |B|} \). Specifically, OGDA maintains two sequences of action distributions \( \tilde{x}_1, \tilde{x}_2, \ldots \in \Delta_A \) and \( x_1, x_2, \ldots \in \Delta_A \) for Player 1, and similarly two sequences \( \tilde{y}_1, \tilde{y}_2, \ldots \in \Delta_B \) and \( y_1, y_2, \ldots \in \Delta_B \) for Player 2, following the updates below:
\[
\begin{align*}
\tilde{x}_{t+1} & = \Pi_{\Delta_A} \{ \tilde{x}_t - \eta Q y_t \}, & x_{t+1} & = \Pi_{\Delta_A} \{ \tilde{x}_{t+1} - \eta Q y_t \}, \\
\tilde{y}_{t+1} & = \Pi_{\Delta_B} \{ \tilde{y}_t + \eta Q^\top x_t \}, & y_{t+1} & = \Pi_{\Delta_B} \{ \tilde{y}_{t+1} + \eta Q^\top x_t \},
\end{align*}
\]
where \( \eta \) is some learning rate. As one can see, unlike the standard Gradient Descent Ascent algorithm which simply sets \((x_t, y_t) = (\tilde{x}_t, \tilde{y}_t)\), OGDA takes a further descent/ascent step using the latest gradient to obtain \((x_t, y_t)\), which is then used to evaluate the gradient (of the function \( f(x, y) = x^\top Q y \)). Wei et al. (2021) prove that the iterate \((\tilde{x}_t, \tilde{y}_t)\) (or \((x_t, y_t)\)) converges to the set of Nash equilibria of the matrix game at a linear rate, which motivates us to generalize it to Markov games. As we show in the following sections, however, the extensions of both the algorithm and the analysis are highly non-trivial.

We remark that while Wei et al. (2021) also analyze the last-iterate convergence of another algorithm called Optimistic Multiplicatively Weight Update (OMWU), which is even more commonly used in finite-action games, they also show that the theoretical guarantees of OMWU hold under more limited assumptions (e.g., requiring the uniqueness of the equilibrium), and its empirical performance is also inferior to that of OGDA. We therefore only extend the latter to Markov games.

\textbf{3. Algorithm and Main Results}

A natural idea to extend OGDA to Markov games is to run the same algorithm described in Section 2 for each state \( s \) with the game matrix \( Q \) being \( Q_{x_t, y_t}^s \). However, an important difference is that now the game matrix is \textit{changing} over time. Indeed, if the polices are changing rapidly for subsequent states, the game matrix \( Q_{x_t, y_t}^s \) will also be changing rapidly, which makes the update on state \( s \) highly unstable and in turn causes similar issues for previous states.

To resolve this issue, we propose to have a \textit{critic} slowly learn the value function for each state. Specifically, for each state \( s \), the critic maintains a sequence of values \( V_0^s = 0, V_1^s, V_2^s, \ldots \). During iteration \( t \), instead of using \( Q_{x_t, y_t}^s \) as the game matrix for state \( s \), we use \( Q_t^s \) defined via \( Q_t^s(a, b) = \sigma(s, a, b) + \gamma \mathbb{E}_{s' \sim P(s, a, b)} [V_{s', y}^{s'}] \). Ideally, OGDA would then take the role of an \textit{actor} and compute \( x_{t+1}^s \) and \( \tilde{x}_{t+1}^s \) using the gradient \( Q_t^s y_t^s \) (and similarly \( y_{t+1}^s \) and \( \tilde{y}_{t+1}^s \) using the gradient \( Q_t^s \tilde{x}_t^s \)). Since such exact gradient information is often unknown, we only require the algorithm to come up with estimations \( \ell_t^s \) and \( r_t^s \) such that \( \|\ell_t^s - Q_t^s y_t^s\| \leq \varepsilon \) and \( \|r_t^s - Q_t^s \tilde{x}_t^s\| \leq \varepsilon \) for some prespecified error \( \varepsilon \) (more discussions in Section 3.1). See updates Eq. (2)-Eq. (5) in Algorithm 1. Note that
Algorithm 1 Optimistic Gradient Descent/Ascent for Markov Games

Parameters: $\gamma \in \left[ \frac{1}{2}, 1 \right], \eta \leq \frac{1}{10\epsilon} \sqrt{\frac{(1-\gamma)^2}{S}}, \epsilon \in \left[ 0, \frac{1}{1-\gamma} \right].$

Parameters: a non-increasing sequence $\{\alpha_t\}_{t=1}^T$ that goes to zero.

Initialization: $\forall s \in S$, arbitrarily initialize $x_1^s = x_1^s \in \Delta_A$ and $y_1^s = y_1^s \in \Delta_B$, and set $V_0^s \leftarrow 0.$

for $t = 1, \ldots, T$ do

For all $s$, define $Q_t^s \in \mathbb{R}^{|A| \times |B|}$ as

\[ Q_t^s(a, b) \stackrel{\Delta}{=} \sigma(s, a, b) + \gamma E_{s' \sim p(s,a,b)} \left[ V_{t-1}^{s'} \right], \]

and update

\[ \hat{x}_{t+1}^s = \Pi_{\Delta_A} \left\{ x_t^s - \eta \ell_t^s \right\}, \]

\[ x_{t+1}^s = \Pi_{\Delta_A} \left\{ \hat{x}_{t+1}^s - \eta \ell_t^s \right\}, \]

\[ \tilde{y}_{t+1}^s = \Pi_{\Delta_B} \left\{ y_t^s + \eta r_t^s \right\}, \]

\[ y_{t+1}^s = \Pi_{\Delta_B} \left\{ \tilde{y}_{t+1}^s + \eta r_t^s \right\}, \]

\[ V_t^s = (1 - \alpha_t) V_{t-1}^s + \alpha_t \rho_t^s, \]

where $\ell_t^s, r_t^s,$ and $\rho_t^s$ are $\epsilon$-approximations of $Q_t^s, x_t^s, y_t^s$ respectively, such that $\|\ell_t^s - Q_t^s y_t^s\| \leq \epsilon, \|r_t^s - Q_t^s x_t^s\| \leq \epsilon,$ and $|\rho_t^s - x_t^s Q_t^s y_t^s| \leq \epsilon.$

end

similar to (Wei et al., 2021), we adopt a constant learning rate $\eta$ (independent of the number of iterations) in these updates.

At the end of each iteration $t$, the critic then updates the value function via $V_t^s = (1 - \alpha_t) V_{t-1}^s + \alpha_t \rho_t^s,$ where $\rho_t^s$ is an estimation of $x_t^s Q_t^s y_t^s$ such that $|\rho_t^s - x_t^s Q_t^s y_t^s| \leq \epsilon.$ To stabilize the game matrix, we require the learning rate $\alpha_t$ to decrease in $t$ and go to zero. Most of our analysis is conducted under this general condition, and the final convergence rate depends on the concrete form of $\alpha_t$, which we set to $\alpha_t = \frac{H+1}{H+2}$ with $H = \frac{2}{1-\gamma}$ inspired by (Jin et al., 2018) (there could be a different choice leading to a better convergence though).

Our main results are the following two theorems on the last-iterate convergence of Algorithm 1.

Theorem 1 (Average duality-gap convergence) Algorithm 1 with the choice of $\alpha_t = \frac{H+1}{H+2}$ where $H = \frac{2}{1-\gamma}$ guarantees

\[ \frac{1}{T} \sum_{t=1}^{T} \max_{s,x',y'} \left( V_t^s(x',y') - V_t^s(x',\tilde{y_t}) \right) = O \left( \frac{|S|}{\eta(1-\gamma)^2} \sqrt{\frac{\log T}{T}} + \frac{|S|\sqrt{\epsilon}}{\sqrt{\eta}(1-\gamma)^2} \right), \]

5. For simplicity, here we assume that the two players share the same estimator $\rho_t^s$ (and thus same $V_t^s$ and $Q_t^s$). However, our analysis works even if they maintain different versions of $\rho_t^s$, as long as they are $\epsilon$-close to $x_t^s Q_t^s y_t^s$ with respect to their own $Q_t^s$. 

6
Theorem 2 (Last-iterate convergence) Algorithm 1 with the choice of $\alpha_t = \frac{H+1}{H+t}$ where $H = \frac{2}{1-\gamma}$ guarantees with $\tilde{z}_t^s = (\tilde{x}_t^s, \tilde{y}_t^s)$,

$$\frac{1}{|S|} \sum_{s \in S} \text{dist}_s^2(\tilde{z}_t^s) = \mathcal{O}\left(\frac{|S|^2}{\eta^4 C^4 (1-\gamma)^4} \epsilon^4 + \frac{\epsilon}{\eta C^2 (1-\gamma)^3}\right),$$

where $C > 0$ is a problem-dependent constant (that always exists) satisfying: for all state $s$ and all policy pair $z = (x, y)$, $\max_{x', y'} (x^s Q_s^s y_s^s - x'^s Q_s^s y_s^s) \geq C \text{dist}_s(z^s)$.

Theorem 1 shows that the average duality-gap for each state $s$ goes to zero when both $1/T$ and $\epsilon$ go to zero, though it does not show the convergence of the policy. Theorem 2, on the other hand, shows a concrete finite-time convergence rate on the distance of $\tilde{z}_t^s$ from the equilibrium set, which goes down at the rate of $1/T$ up to the estimation error $\epsilon$. The problem-dependent constant $C$ is similar to the matrix game case analyzed in (Wei et al., 2021), as we will discuss in Section 4. As far as we know, this is the first symmetric algorithm with finite-time last-iterate convergence for both players simultaneously.

3.1. Estimation

In the full-information setting where all parameters of the Markov game are given, we can calculate the exact value of $Q_t^s y_t^s$, $Q_t^s x_t^s$, and $x_t^s Q_t^s y_t^s$, making $\epsilon = 0$. In this case, our algorithm is essentially a new policy-iteration style algorithm for solving Markov games. However, in a learning setting where the parameters are unknown, the players need to estimate these quantities based on any feedback from the environments. Here, we discuss how to do so when the players only observe their current state and their loss/reward after taking an action.

Specifically, in iteration $t$ of our algorithm and with $(x_t, y_t)$ at hand, the two players interact with each other for a sequence of $L$ steps, following a mixed strategy with a certain amount of uniform exploration defined via: $\tilde{x}_t^a(a) = \left(1 - \frac{\epsilon'}{2}\right)x_t^a(a) + \frac{\epsilon'}{2|A|}$ and $\tilde{y}_t^b(b) = \left(1 - \frac{\epsilon'}{2}\right)y_t^b(b) + \frac{\epsilon'}{2|B|}$, where $\epsilon' = (1-\gamma)\epsilon$. This generates a sequence of observations $\{(s_i, a_i, \sigma(s_i, a_i, b_i))\}_{i=1}^L$ for Player 1 and similarly a sequence of observations $\{(s_i, b_i, \sigma(s_i, a_i, b_i))\}_{i=1}^L$ for Player 2, where $a_i \sim \tilde{x}_t^s$, $b_i \sim \tilde{y}_t^s$, and $s_{i+1} \sim p(\cdot|s_i, a_i, b_i)$. Then we construct the estimators as follows:

$$\ell_t^s(a) = \sum_{i=1}^L 1[s_i = s, a_i = a] \left(\sigma(s, a, b_i) + \gamma V_{t-1}^{s_i+1}\right),$$

$$r_t^s(b) = \sum_{i=1}^L 1[s_i = s, b_i = b] \left(\sigma(s, a_i, b) + \gamma V_{t-1}^{s_i+1}\right),$$

$$p_t^s = \sum_{i=1}^L 1[s_i = s] \left(\sigma(s, a_i, b_i) + \gamma V_{t-1}^{s_i+1}\right).$$

(If any of the denominator is zero, define the corresponding estimator as zero.) To make sure that these are accurate estimators for every state, we naturally need to ensure that every state is visited often enough. To this end, we make the following assumption similar to (Auer and Ortner, 2007), which essentially requires that the induced Markov chain under any stationary policy pair is irreducible.
Assumption 1 There exists $\mu > 0$ such that $\frac{1}{\mu} = \max_{x,y} \max_{s,s'} T_{x,y}^{s \rightarrow s'}$, where $T_{x,y}^{s \rightarrow s'}$ is the expected time to reach $s'$ from $s$ following the policy pair $(x,y)$.

Under this assumption, the following theorem shows that taking $L \approx 1/\varepsilon^3$ is enough to ensure the accuracy of the estimators (see Appendix H for the proof).

Theorem 3 Suppose that Assumption 1 holds and $L = \tilde{\Omega}\left(\frac{|A|^3 + |B|^3}{(1-\gamma)\mu^2} \log^2(T/\delta)\right)$. Then the estimators Eq. (7), Eq. (8), and Eq. (9) ensure that with probability at least $1 - \delta$, $\|r_t^s - Q_t^s x_t^s\|$, $\|\rho_t^s - x_t^s Q_t^s y_t^s\|$ are all of order $O(\varepsilon)$ for all $t$.

Together with Theorem 1 and Theorem 2, given a fixed number of interactions between the players, we can now determine optimally how many iterations we should run our algorithm (and consequently how large we should set $\varepsilon$). Equivalently, we show below how many iterations or total interactions are needed to achieve a certain accuracy. (The choice of $\alpha_t$ is the same as in Theorem 1 and Theorem 2.)

Corollary 4 If Assumption 1 holds, then running Algorithm 1 with estimators Eq. (7), Eq. (8), Eq. (9) and $L = \tilde{\Omega}\left(\frac{|A|^3 + |B|^3}{(1-\gamma)\mu^2} \log^2(T/\delta)\right)$ for $T = \tilde{\Omega}\left(\frac{|S|^2}{\eta^2(1-\gamma)^2}\right)$ iterations ensures with probability at least $1 - \delta$, $\frac{1}{T} \sum_{t=1}^{T} \max_{s,x',y'}(V_{x_t,y_t}^s - V_{x_t,y_t}^{s'}) \leq \varepsilon$. Ignoring other dependence, this requires $\tilde{\Omega}(1/\varepsilon^8)$ interactions in total.

Corollary 5 If Assumption 1 holds, then running Algorithm 1 with estimators Eq. (7), Eq. (8), Eq. (9) and $L = \tilde{\Omega}\left(\frac{|A|^3 + |B|^3}{(1-\gamma)\mu^2} \log^2(T/\delta)\right)$ for $T = \tilde{\Omega}\left(\frac{|S|^2}{\eta^2(1-\gamma)^2}\right)$ iterations ensures with probability at least $1 - \delta$, $\frac{1}{|S|} \sum_{s \in S} \text{dist}_s^2(\varepsilon_T^s) \leq \varepsilon$. Ignoring other dependence, this requires $\tilde{\Omega}(1/\varepsilon^4)$ interactions in total.

3.2. Rationality

Finally, we argue that from the perspective of a single player (take Player 1 as an example), our algorithm is also rational, in the sense that it allows Player 1 to converge to the best response to her opponent if Player 2 is not applying our algorithm but instead uses an arbitrary stationary policy. We show this single-player-perspective version in Algorithm 2, where Player 1 still follows the updates Eq. (2), Eq. (3), and Eq. (6), while $y_1$ is fixed to a stationary policy $y$ used by Player 2.

In fact, thanks to the agnostic nature of our algorithm, rationality is essentially an implication of the convergence property. To see this, consider a modified two-player Markov game with the difference being that the opponent has only a single action (call it 1) on each state, the loss function is redefined as $\rho(s, a, 1) = \mathbb{E}_{b \sim y_1} \sigma(s, a, b)$, and the transition kernel is redefined as $p(s' | s, a, 1) = \mathbb{E}_{b \sim y_1} [p(s' | s, a, b)]$. It is straightforward to see that following our algorithm, Player 1's behaviors in the modified game are exactly the same. On the other hand, in the modified game, since Player 2 has only one action (and thus one strategy), she can also be seen as using our

6. We use $\tilde{\Omega}$ to hide logarithmic factors except for $\log(T)$ and $\log(1/\delta)$.
7. The rationality defined by Bowling and Veloso (2001) requires that the learner converges to the best response as long as the opponent converges to a stationary policy. While our algorithm does handle this case, as a proof of concept, we only consider the simpler scenario where the opponent simply uses a stationary policy.
algorithm. Therefore, we can apply our convergent guarantees to the modified game, and since the minimax policy in the modified game is exactly the best response in the original game, we know that Player 1 indeed converges to the best response. We summarize these rationality guarantees in the following theorem, with the formal proof deferred to Appendix I.

**Theorem 6** Algorithm 2 with the choice of \( \alpha_t = \frac{H+1}{H+t} \) where \( H = \frac{2}{1-\gamma} \) guarantees

\[
\frac{1}{T} \sum_{t=1}^{T} \max_{s,x'} (V_{s}^{x'} - V_{s}^{x}) = O \left( \frac{|S|}{\eta(1-\gamma)^2} \sqrt{\frac{\log T}{T}} + \frac{|S|\sqrt{\varepsilon}}{\sqrt\eta(1-\gamma)^2} \right),
\]

and for \( X_{BR} = \{ x : V_{x,y}^{s} = \min_{x'} V_{x',y}^{s}, \forall s \in S \} \) and some problem-dependent constant \( C' > 0 \),

\[
\frac{1}{|S|} \sum_{s \in S} \| \hat{x}_t^s - \Pi X_{BR} \{ \hat{x}_t^s \} \|^2 = O \left( \frac{|S|^2}{\eta^4C'^4(1-\gamma)^4T} + \frac{\varepsilon}{\eta C'^2(1-\gamma)^2} \right).
\]

4. Analysis Overview

In this section, we give an overview of how we analyze Algorithm 1 and prove Theorem 1 and Theorem 2. We start by giving a quick review of the analysis of (Wei et al., 2021) for matrix games, and then highlight how we overcome the challenges when generalizing it to Markov games.

**Review for matrix games** Recall the update in Eq. (1) for a fixed matrix \( Q \). Wei et al. (2021) show the following two convergence guarantees:

1. Average duality-gap convergence:

\[
\frac{1}{T} \sum_{t=1}^{T} \Delta(\hat{z}_t) = O \left( \frac{1}{\eta \sqrt{T}} \right)
\]

where \( \Delta(z) = \max_{x',y'} (x^T Q y' - x'^T Q y) \) is the duality gap of \( z = (x,y) \).

2. Last-iterate convergence:

\[
\text{dist}^2(\hat{z}_t) \leq C_1 \text{dist}^2(\hat{z}_1) (1 + \eta^2 C^2)^{-t}
\]

where \( \text{dist}_*(z) \) is the distance from \( z \) to the set of equilibria, \( C_1 \) is a universal constant, and \( C > 0 \) is a positive constant that depends on \( Q \).

The analysis of (Wei et al., 2021) starts from the following single-step inequality that follows the standard Online Mirror Descent analysis and describes the relation between \( \text{dist}^2(\hat{z}_{t+1}) \) and \( \text{dist}^2(\hat{z}_t) \):

\[
\text{dist}^2(\hat{z}_{t+1}) \leq \text{dist}^2(\hat{z}_t) + \eta^2 \| z_t - z_{t-1} \|^2 - \left( \| z_{t+1} - z_t \|^2 + \| z_t - \hat{z}_t \|^2 \right).
\]

---

8. This is not to be confused with the constant \( C \) in Theorem 2. We overload the notation because they indeed play the same role in the analysis.
The instability penalty term makes $\text{dist}^2_{s}(\hat{z}_{t+1})$ larger if $\|z_t - z_{t-1}\|$ is large, while the instability bonus term makes $\text{dist}^2_{s}(\hat{z}_{t+1})$ smaller if either $\|\hat{z}_{t+1} - z_t\|$ or $\|z_t - \hat{z}_t\|$ is large. To obtain Eq. (10), Wei et al. (2021) make the observation that the instability bonus term is lower bounded by a constant times the squared duality gap of $\hat{z}_{t+1}$, that is, $\|\hat{z}_{t+1} - z_t\|^2 + \|z_t - \hat{z}_t\|^2 \geq \eta^2 \Delta^2(\hat{z}_{t+1})$, and thus

$$\text{dist}^2_{s}(\hat{z}_{t+1}) \leq \text{dist}^2_{s}(\hat{z}_t) + \eta^2 \|z_t - z_{t-1}\|^2 - \frac{1}{2} \left(\|\hat{z}_{t+1} - z_t\|^2 + \|z_t - \hat{z}_t\|^2\right) - \Omega(\eta^2 \Delta^2(\hat{z}_{t+1})).$$

(13)

By taking $\eta \leq \frac{1}{2}$, summing over $t$, canceling the penalty term with the bonus term, telescoping and rearranging, we get $\sum_{t=1}^{T} \Delta^2(\hat{z}_t) \leq O(1/\eta^2)$. An application of Cauchy-Schwarz inequality then proves Eq. (10).

To further obtain Eq. (11), Wei et al. (2021) prove that there exists some problem-dependent constant $C > 0$ such that for all $z$, $\Delta(z) \geq C \text{dist}_s(z)$. This, when combined with Eq. (13), shows

$$\text{dist}^2_{s}(\hat{z}_{t+1}) \leq \frac{\text{dist}^2_{s}(\hat{z}_t)}{1 + \Omega(\eta^2 C^2)} + \eta^2 \|z_t - z_{t-1}\|^2 - \Omega(\|\hat{z}_{t+1} - z_t\|^2 + \|z_t - \hat{z}_t\|^2).$$

(14)

By upper bounding $\|z_t - z_{t-1}\|^2 \leq 2\|z_t - \hat{z}_t\|^2 + 2\|\hat{z}_t - z_{t-1}\|^2$ and rearranging, they further obtain:

$$\text{dist}^2_{s}(\hat{z}_{t+1}) + c' \|\hat{z}_{t+1} - z_t\|^2 + c' \|z_t - \hat{z}_t\|^2 \leq \frac{\text{dist}^2_{s}(\hat{z}_t) + c' \|\hat{z}_t - z_{t-1}\|^2 + c' \|z_{t-1} - \hat{z}_{t-1}\|^2}{1 + \Omega(\eta^2 C^2)}$$

(15)

for some universal constant $c'$, which clearly indicates the linear convergence of $\text{dist}_s(\hat{z}_t)$ and hence proves Eq. (11).

**Overview of our proofs** We are now ready to show the high-level ideas of our analysis. For simplicity, we consider the case with $\varepsilon = 0$ and also assume that there is a unique equilibrium $(x_*, y_*)$ (these assumptions are removed in the formal proofs). Our analysis follows the steps below.

**Step 1 (Appendix B)** Similar to Eq. (12), we conduct a single-step analysis for OGDA in Markov games (Lemma 24), which shows for all state $s$:

$$\text{dist}^2_{s}(\hat{z}_{t+1}) \leq \text{dist}^2_{s}(\hat{z}_t) + \eta^2 \|x_t^s - x_{t-1}^s\|^2 - \left(\|\hat{z}_{t+1}^s - x_t^s\|^2 + \|z_t^s - \hat{z}_t^s\|^2\right) + 8\eta^2 \|Q_{t+1}^s - Q_t^s\|^2 + 4\eta \|Q_t^s - Q_*^s\|.$$  

(16)

Comparing this with Eq. (12), we see that, importantly, since the game matrix $Q_t^s$ is changing over time, we have two extra instability penalty terms: $\eta^2 \|Q_{t}^s - Q_{t-1}^s\|^{2}$ and $\eta \|Q_t^s - Q_*^s\|$. Our hope is to further upper bound these two penalty terms by something related to $\|z_t^s - z_{t+1}^s\|^2$, so that they can again be canceled by the bonus term $-(\|\hat{z}_{t+1}^s - z_t^s\|^2 + \|z_t^s - \hat{z}_t^s\|^2)$. Indeed, in Steps 3-5, we show that part of them can be bounded by a weighted sum of $\{\|z_t^{s'} - z_{t+1}^{s'}\|^2\}_{s' \in S, \tau \leq t}$.  

10
Step 2 (Appendix C): Lower bounding $\|z_{t+1}^s - z_t^s\|^2 + \|z_t^s - \bar{z}_t^s\|^2$. As in Eq. (13), we aim to lower bound the instability bonus term by the duality gap. However, since the updates are based on $Q_t^s$ instead of $Q^s_\ast$, we can only relate the bonus term to the duality gap with respect to $Q_t^s$. To further relate this to the duality gap with respect to $Q^s_\ast$, we pay a quantity related to $\|Q_t^s - Q^s_\ast\|$. Formally, we show in Lemma 25:

$$\|z_{t+1}^s - z_t^s\|^2 + \|z_t^s - \bar{z}_t^s\|^2 \geq \Omega(\eta^2 \Delta^2(\bar{z}_{t+1})) - O(\eta\|Q_t^s - Q^s_\ast\|),$$

where $\Delta(z^s) \triangleq \max_{x^s,y^s}(x^sQ_t^s y^s - x^sQ_\ast y^s)$ is the duality gap on state $s$ with respect to $Q_\ast$.\(^9\)

Step 3 (Appendix D): Upper bounding $\|Q_{t+1}^s - Q_t^s\|^2$. $\|Q_{t+1}^s - Q_t^s\|^2$ is upper bounded by $\gamma^2 \max_{s'}(V_{t}^{s'} - V_{t-1}^{s'})^2$ by the definition of $Q_t^s$. Furthermore, $V_{t}^{s'} - V_{t-1}^{s'}$ is a weighted sum of $\{\rho_t^{s'} - \rho_{t-1}^{s'}\}_{t=1}^{T}$ by the definition of $V_{t}^{s'}$, and also $\rho_t^{s'} - \rho_{t-1}^{s'} = x_t^{s'}Q_t^{s'} y_{t}^{s'} - x_{t-1}^{s'}Q_{t-1}^{s'} y_{t-1}^{s'} = \mathcal{O}(\|z^s_t - z^{s'}_{t-1}\| + \|Q_t^{s'} - Q_{t-1}^{s'}\|)$. In sum, one can upper bound $\|Q_{t+1}^s - Q_t^s\|^2$ by a weighted sum of $\|z^s_t - z^{s'}_{t-1}\|^2$ and $\|Q_t^{s'} - Q_{t-1}^{s'}\|^2$. After formalizing the above relations, we obtain the following inequality (see Lemma 28):

$$\|Q_{t+1}^s - Q_t^s\|^2 \leq \max_{s'} \frac{8\gamma^2}{(1 - \gamma)^3} \sum_{\tau=1}^{T} \alpha_t^\tau \|z^s_{\tau} - z^{s'}_{\tau-1}\|^2 + \max_{s'} \frac{2\gamma^2}{1 + \gamma} \sum_{\tau=1}^{T} \alpha_t^\tau \|Q_t^{s'} - Q_{t-1}^{s'}\|^2$$

for some coefficient $\alpha_t^\tau$ defined in Appendix A.2. With recursive expansion, the above implies that $\|Q_{t+1}^s - Q_t^s\|^2$ can be upper bounded by a weighted sum of $\|z^s_t - z^{s'}_{t-1}\|^2$ for $s' \in S$ and $\tau \leq t$.

Step 4 (Appendix E): Upper bounding $\|Q_t^s - Q_\ast^s\|$ (Part 1). We first upper bound $\|Q_t^s - Q_\ast^s\|$ with respect to the following weighted-regret quantity

$$\text{Reg}_t \triangleq \max_s \left\{ \sum_{\tau=1}^{T} \alpha_t^\tau (x_t^{s'} - x_\ast^{s'})Q_t^{s'} y_{\tau}^{s'}, \sum_{\tau=1}^{T} \alpha_t^\tau x_t^{s'}Q_t^{s'} (y_t^{s'} - y_\ast^{s'}) \right\}.$$
Step 5 (Appendix E): Upper bounding $\|Q^s_{t} - Q^s_{t-1}\|$ (Part 2). In this step, we further relate $\text{Reg}_t$ to $\{\|z^s_{t'} - z^s_{t'-1}\|^2\}_{t \leq t', s' \in S}$. From a one-step regret analysis of OGDA, we have the following (for Player 1):

$$(x^s_t - x^s_{t-1})^\top Q^s_{t} y^s_{t} \leq \frac{1}{2\eta} \left( \text{dist}^2_s(\bar{x}^s_t) - \text{dist}^2_s(\bar{x}^s_{t+1}) \right) + \frac{4\eta}{(1-\gamma)^2} \|y^s_{t} - y^s_{t-1}\|^2 + 4\eta \|Q^s_{t} - Q^s_{t-1}\|^2.$$  

Recall that $\text{Reg}_t$ is defined via a weighted sum of the left-hand side above with weights $\alpha^*_t$. Therefore, we take the weighted sum of the above and bound $\sum_{\tau=1}^{t} \alpha^*_t (x^s_t - x^s_{t-1}) Q^s_{t} y^s_{t}$ by

\[
\frac{\alpha^*_t \text{dist}^2_s(\bar{x}^s_t)}{2\eta} + \sum_{\tau=1}^{t} \frac{\alpha^*_t}{2\eta} \left( \text{dist}^2_s(\bar{x}^s_{\tau}) - \text{dist}^2_s(\bar{x}^s_{\tau+1}) \right) + \frac{4\eta}{(1-\gamma)^2} \sum_{\tau=1}^{t} \alpha^*_s \|y^s_{\tau} - y^s_{\tau-1}\|^2 + 4\eta \sum_{\tau=1}^{t} \alpha^*_s \|Q^s_{\tau} - Q^s_{\tau-1}\|^2 \leq \frac{1}{2\eta} \sum_{\tau=1}^{t} \alpha^*_s \text{dist}^2_s(\bar{z}^s_{\tau}) + \frac{4\eta}{(1-\gamma)^2} \sum_{\tau=1}^{t} \alpha^*_s \|z^s_{\tau} - z^s_{\tau-1}\|^2 + 4\eta \sum_{\tau=1}^{t} \alpha^*_s \|Q^s_{\tau} - Q^s_{\tau-1}\|^2 \tag{19}
\]

where in the inequality we rearrange the first summation and use the fact $\alpha^*_t - \alpha^*_{t-1} = 1 - \alpha^*_{t-1} \alpha^*_t$ (see the formal proof in Lemma 30). Since the case for $\sum_{\tau=1}^{t} \alpha^*_s x^s_{\tau} Q^s_{\tau} (y^s_{\tau} - y^s_{\tau-1})$ is similar, by the definition of $\text{Reg}_t$, we conclude that $\text{Reg}_t$ is upper bounded by the maximum over $s$ of the sum of the three terms in Eq. (19). Note that, $\text{term}_2$ is itself a weighted sum of $\{\|z^s_{\tau} - z^s_{\tau-1}\|^2\}_{\tau \leq t}$, and $\text{term}_3$ can also be upper bounded by a weighted sum of $\{\|z^{s'}_{\tau} - z^{s'}_{\tau-1}\|^2\}_{\tau \leq t, s' \in S}$ as we already showed in Step 3.

Combining all steps. Summing up Eq. (16) over all $s$, and based on all earlier discussions, we have

\[
\sum_{s} \text{dist}^2_s(\bar{z}^s_{\tau+1}) \leq \sum_{s} \text{dist}^2_s(\bar{z}^s_{\tau}) + \sum_{\tau=1}^{t} \sum_{s} \mu^*_s \alpha^*_{t-1} \text{dist}^2_s(\bar{z}^s_{\tau}) + \sum_{\tau=1}^{t} \sum_{s} \nu^*_s \|z^s_{\tau} - z^s_{\tau-1}\|^2 \tag{20}
\]

for some weights $\mu^*_s$ and $\nu^*_s$ (a large part of the analysis is devoted to precisely calculating these weights). Here, the $-\Omega \left( \eta^2 \sum_{s} \Delta^2(\bar{z}^s_{\tau+1}) \right)$ term comes from Step 2; $\text{term}_1$ is a weighted sum of $\{\alpha^*_{t-1} \text{dist}^2_s(\bar{z}^s_{\tau})\}_{\tau \leq t, s' \in S}$ that comes from $\text{term}_1$ in Step 5; $\text{term}_5$ is a weighed sum of $\{\|z^{s'}_{\tau} - z^{s'}_{\tau-1}\|^2\}_{\tau \leq t, s' \in S}$ that comes from all other terms we discuss in Steps 3-5.

Obtaining average duality-gap bound To obtain the average duality-gap bound in Theorem 1, we sum Eq. (20) over $t$, and further argue that the sum of $\text{term}_5$ over $t$ is smaller than the sum of $\text{term}_6$ over $t$ (hence they are canceled with each other). Rearranging and telescoping leads to

\[
\eta^2 \sum_{t=1}^{T} \sum_{s} \Delta^2(\bar{z}^s_{t+1}) = O \left( \sum_{t=1}^{T} \sum_{\tau=1}^{t} \sum_{s} \mu^*_s \alpha^*_{t-1} \text{dist}^2_s(\bar{z}^s_{\tau}) \right) = O \left( \sum_{t=1}^{T} \sum_{\tau=1}^{t} \sum_{s} \mu^*_s \alpha^*_{t-1} \right).
\]
As long as $\alpha_t$ is decreasing and going to zero, the right-hand side above can be shown to be sub-linear in $T$. Further relating $\max_{x',y'} \left( V^s_{x_t,y'_t} - V^s_{x'_t,y_t} \right)$ to $\Delta(\hat{z}^s_t)$ (Lemma 32) proves Theorem 1.

**Obtaining last-iterate convergence bound** Following the matrix game case, there is a problem-dependent constant $C > 0$ such that $\Delta(\hat{z}^s_{t+1}) \geq C \text{dist}_s(\hat{z}^s_{t+1})$. Similarly to how Eq. (14) is obtained, we use this in Eq. (20) and arrive at

$$
\sum_s \text{dist}_s^2(\hat{z}^s_{t+1}) \leq \frac{1}{1 + \Omega(\eta^2 C^2)} \sum_s \text{dist}_s^2(\hat{z}^s_t) + \sum_{\tau=1}^t \sum_s \mu^s_\tau \alpha_{\tau-1} \text{dist}_s^2(\hat{z}^s_\tau)
$$

$$
\sum_{\tau=1}^t \sum_s \nu^s_\tau \left\| z^s_\tau - z^s_{\tau-1} \right\|^2 - \Omega \left( \sum_s \left( \left\| \hat{z}^s_{t+1} - z^s_t \right\|^2 + \left\| z^s_t - \hat{z}^s_t \right\|^2 \right) \right)
$$

(21)

Then ideally we would like to follow a similar argument from Eq. (14) to Eq. (15) to obtain a last-iterate convergence guarantee. However, we face two more challenges here. First, we have an extra term. Fortunately, this term vanishes when $t$ is large as long as $\alpha_t$ decreases and converges to zero. Second, in Eq. (14), the indices of the negative term $\left\| \hat{z}^s_{t+1} - z^s_t \right\|^2 + \left\| z^s_t - \hat{z}^s_t \right\|^2$ and the positive term $\eta^2 \left\| z^s_t - z^s_{t-1} \right\|^2$ are only offset by 1 so that a simple rearrangement is enough to get Eq. (15), while in Eq. (21), the indices in term and term are far from each other. To address this issue, we further introduce a set of weights and consider a weighted sum of Eq. (21) over $t$. We then show that the weighted sum of term can be canceled by the weighted sum of term. Combining the above proves Theorem 2. Note that due to these extra terms, our last-iterate convergence rate is only sublinear (while Eq. (11) shows a linear rate for matrix games).

### 5. Conclusion and Future Directions

In this work, we propose the first decentralized algorithm for two-player zero-sum Markov games that is rational, convergent, agnostic, symmetric, and having a finite-time convergence rate guarantee at the same time. The algorithm is based on running OGDA on each state, together with a slowly changing critic that stabilizes the game matrix on each state.

Our work studies the most basic tabular setting, and also requires a structural assumption when estimation is needed that sidesteps the difficulty of performing exploration over the state space. Important future directions include relaxing either of these assumptions, that is, extending our framework to allow function approximation and/or incorporating efficient exploration mechanisms. Studying OGDA-based algorithms beyond the two-player zero-sum setting is also an interesting future direction.

**Acknowledgments**

This work is supported by NSF Award IIS-1943607 and a Google Faculty Research Award.
References


Appendix A. Notations

A.1. Simplifications of the Notations

We define the following notations to simplify the proofs:

**Definition 7** \( \hat{x}_0 = x_0 = 0_{|A|} \) (zero vector with dimension \(|A|\)), \( \hat{y}_0 = y_0 = 0_{|B|} \), \( Q_0 = 0_{|A| \times |B|} \), \( \ell_0 = 0_{|A|} \), \( r_0 = 0_{|B|} \), \( \rho_0 = 0 \), \( \alpha_0 = 1 \).

Besides, for a matrix \( Q \), we define \( \|Q\| = \max_{i,j} |Q_{ij}| \). To avoid cluttered notation, a product of the form \( x^\top Q y \) is usually simply written as \( xQy \).

A.2. Auxiliary Coefficients

In this subsection, we define several coefficients that are related to the value learning rate \( \{\alpha_t\} \).

**Definition 8** \((\alpha^\tau_t)\) For non-negative integers \( \tau \) and \( t \) with \( \tau \leq t \), define \( \alpha^\tau_t = \alpha_\tau \prod_{i=\tau+1}^{t} (1 - \alpha_i) \).

**Definition 9** \((\delta^\tau_t)\) For non-negative integers \( \tau \) and \( t \) with \( \tau \leq t \), define \( \delta^\tau_t \triangleq \prod_{i=\tau+1}^{t} (1 - \alpha_i) \).

**Definition 10** \((\beta^\tau_t)\) For positive integers \( \tau \) and \( t \) with \( \tau < t \), define \( \beta^\tau_t = \alpha_\tau \prod_{i=\tau+1}^{t-1} (1 - \alpha_i + \alpha_i \gamma) \). Define \( \beta^t_t = 1 \).

**Definition 11** \((\lambda_t)\) For positive integers \( t \), define \( \lambda_t = \max \left\{ \frac{\alpha_{t+1}}{\alpha_t}, 1 - \frac{\alpha_t(1-\gamma)}{2} \right\} \).

**Definition 12** \((\lambda^\tau_t)\) For positive integers \( \tau \) and \( t \) with \( \tau < t \), define \( \lambda^\tau_t = \alpha_\tau \prod_{i=\tau+1}^{t-1} \lambda_i \). Define \( \lambda^t_t = 1 \).

A.3. Auxiliary Variables

In this subsection, we define several auxiliary variables to be used in the later analysis.

**Definition 13** \((J^s_t)\) For every state \( s \in S \), define the sequence \( \{J^s_t\}_{t=1,2,...} \) by

\[
J^s_1 = \| z^s_1 - z^s_0 \|^2 ,
\]

\[
J^s_t = (1 - \alpha_t) J^s_{t-1} + \alpha_t \| z^s_t - z^s_{t-1} \|^2 , \quad \forall t \geq 2 .
\]

Furthermore, define \( J_t \triangleq \max_s J^s_t \).

**Definition 14** \((K^s_t)\) For every state \( s \in S \), define the sequence \( \{K^s_t\}_{t=1,2,...} \) by

\[
K^s_1 = \| Q^s_1 - Q^s_0 \|^2 ,
\]

\[
K^s_t = (1 - \alpha_t) K^s_{t-1} + \alpha_t \| Q^s_t - Q^s_{t-1} \|^2 , \quad \forall t \geq 2 .
\]

Furthermore, define \( K_t \triangleq \max_s K^s_t \).
Definition 15 \((\tilde{x}_t, \tilde{y}_t, \tilde{z}_t)\) Define \(\tilde{x}_t = \Pi_{X_t} (\tilde{x}_t^s)\), i.e., the projection of \(\tilde{x}_t^s\) onto the set of optimal policy \(X_t^s\) on state \(s\). Similarly, \(\tilde{y}_t = \Pi_{Y_t} (\tilde{y}_t^s)\), and \(\tilde{z}_t = \Pi_{Z_t} (\tilde{z}_t^s)\) = \((\tilde{x}_t^s, \tilde{y}_t^s)\).

Definition 16 (\(\Delta_t^s\)) Define \(\Delta_t^s = \max_{x', y'} (\tilde{x}_t^s Q_t^s y' - x't^s Q_t^s y')\) for all \(t \geq 1\).

Definition 17 (\(\text{Reg}_t^s\)) Define
\[
\text{Reg}_t^s = \max \left\{ \sum_{\tau=1}^{t} \alpha_t^\tau (x_t^s - \tilde{x}_t^s) Q_t^s y_t^s, \sum_{\tau=1}^{t} \alpha_t^\tau x_t^s Q_t^s (\tilde{y}_t^s - y_t^s) \right\}
\]
and \(\text{Reg}_t = \max_s \text{Reg}_t^s\).

Definition 18 (\(\Gamma_t\)) Define \(\Gamma_t = \max_s \|Q_t^s - Q_t^s\|\).

Definition 19 (\(\theta_t^s\)) Define \(\theta_t^s = \frac{1}{16} \|z_t^s - z_{t-1}^s\|^2 + \frac{1}{16} \|z_{t-1}^s - z_{t-1}^s\|^2\)

Definition 20 (\(Z_t\)) Define \(Z_t = \max_s \sum_{\tau=1}^{t} \alpha_t^\tau \alpha_{t-1} \text{dist}_s (\tilde{z}_t^s)\).

A.4. Assumptions on \(\alpha_t\) and Simple Facts about \(\alpha_t^\tau\)

We require \(\alpha_t\) to satisfy the following:

- \(\alpha_1 = 1\)
- \(0 < \alpha_{t+1} \leq \alpha_t \leq 1\)
- \(\alpha_t \to 0\) as \(t \to \infty\)

Furthermore, \(\alpha_0 \triangleq 1\). Below is an useful lemma that is used in many places:

Lemma 21 If \(\{h_t\}_{t=0, 1, 2, \ldots}\) and \(\{k_t\}_{t=1, 2, \ldots}\) are non-negative sequences that satisfy \(h_t = (1 - \alpha_t) h_{t-1} + \alpha_t k_t\) for \(t \geq 1\), then \(h_t = \sum_{\tau=1}^{t} \alpha_t^\tau k_{t\tau}\).

Proof We prove it by induction. When \(t = 1\), since \(\alpha_1 = 1\), \(h_1 = k_1 = \alpha_1^t k_1\). Assume that the formula is correct for \(h_t\). Then
\[
h_{t+1} = (1 - \alpha_{t+1}) h_t + \alpha_{t+1} k_{t+1}
\]
\[
= (1 - \alpha_{t+1}) \sum_{\tau=1}^{t} \alpha_t^\tau k_{t\tau} + \alpha_{t+1}^t k_{t+1}
\]
\[
= \sum_{\tau=1}^{t} \alpha_{t+1}^\tau k_{t\tau} + \alpha_{t+1}^t k_{t+1} = \sum_{\tau=1}^{t+1} \alpha_{t+1}^\tau k_{t\tau}.
\]

\[\square\]

Corollary 22 The following hold:

- \(V_t^s = \sum_{\tau=1}^{t} \alpha_t^\tau \rho_t^s\)
- \(J_t^s = \sum_{\tau=1}^{t} \alpha_t^\tau \|z_t^s - z_{t-1}^s\|^2\)
- \(K_t^s = \sum_{\tau=1}^{t} \alpha_t^\tau \|Q_t^s - Q_{t-1}^s\|^2\)

Proof They immediately follow from Lemma 21 and the definition of \(J_t^s, K_t^s, V_t^s\).
Appendix B. Proof for Step 1: Single-Step Inequality

Lemma 23  For any state \( s \) and \( t \),

\[
(x_t^s - \tilde{x}_{t^*}^s) Q_t^s y_t^s \leq \frac{1}{2\eta} \left( \text{dist}_s^2(\tilde{x}_t^s) - \text{dist}_s^2(\tilde{x}_{t+1}^s) - \|\tilde{x}_{t+1}^s - x_t^s\|^2 - \|x_t^s - \tilde{x}_t^s\|^2 \right) \\
+ \frac{4\eta}{(1 - \gamma)^2} \|y_t^s - y_{t-1}^s\|^2 + 4\eta \|Q_t^s - Q_{t-1}^s\|^2 + 3\varepsilon,
\]

\[
x_t^s Q_t^s(\tilde{g}_{t^*}^s - y_t^s) \leq \frac{1}{2\eta} \left( \text{dist}_s^2(\tilde{y}_t^s) - \text{dist}_s^2(\tilde{y}_{t+1}^s) - \|\tilde{y}_{t+1}^s - y_t^s\|^2 - \|y_t^s - \tilde{y}_t^s\|^2 \right) \\
+ \frac{4\eta}{(1 - \gamma)^2} \|x_t^s - x_{t-1}^s\|^2 + 4\eta \|Q_t^s - Q_{t-1}^s\|^2 + 3\varepsilon.
\]

Proof  By standard proof of OGDA (see, e.g., the proof of Lemma 1 in (Wei et al., 2021) or Lemma 1 in (Rakhlin and Sridharan, 2013)), we have

\[
(x_t^s - \tilde{x}_{t^*}^s) ^T \ell_t^s \leq \frac{1}{2\eta} \left( \|\tilde{x}_t^s - \tilde{x}_{t^*}^s\|^2 - \|\tilde{x}_{t+1}^s - \tilde{x}_{t^*}^s\|^2 - \|\tilde{x}_{t+1}^s - x_t^s\|^2 - \|x_t^s - \tilde{x}_t^s\|^2 \right) + \eta \|\ell_t^s - \ell_{t-1}^s\|^2.
\]

Since \( \|\tilde{x}_t^s - \tilde{x}_{t^*}^s\|^2 = \text{dist}_s^2(\tilde{x}_t^s) \) and \( \|\tilde{x}_{t+1}^s - \tilde{x}_{t^*}^s\|^2 \geq \text{dist}_s^2(\tilde{x}_{t+1}^s) \) by the definition of \( \text{dist}_s(\cdot) \), we further have

\[
(x_t^s - \tilde{x}_{t^*}^s) ^T \ell_t^s \leq \frac{1}{2\eta} \left( \text{dist}_s^2(\tilde{x}_t^s) - \text{dist}_s^2(\tilde{x}_{t+1}^s) - \|\tilde{x}_{t+1}^s - x_t^s\|^2 - \|x_t^s - \tilde{x}_t^s\|^2 \right) + \eta \|\ell_t^s - \ell_{t-1}^s\|^2.
\]

By the definition of \( \ell_t^s \), we have

\[
\eta \|\ell_t^s - \ell_{t-1}^s\|^2 \\
\leq \eta \|\ell_t^s - Q_t^s y_t^s + (Q_t^s - Q_{t-1}^s) y_t^s + Q_{t-1}^s (y_t^s - y_{t-1}^s) + Q_{t-1}^s y_{t-1}^s - \ell_{t-1}^s\|^2 \\
\leq 4\eta \|\ell_t^s - Q_t^s y_t^s\|^2 + 4\eta \|Q_t^s - Q_{t-1}^s\|^2 + 4\eta \|Q_{t-1}^s (y_t^s - y_{t-1}^s)\|^2 + 4\eta \|Q_{t-1}^s y_{t-1}^s - \ell_{t-1}^s\|^2 \\
\leq 4\eta \|Q_t^s - Q_{t-1}^s\|^2 + \frac{4\eta}{(1 - \gamma)^2} \|y_t^s - y_{t-1}^s\|^2 + 8\eta\varepsilon^2
\]

and

\[
(x_t^s - \tilde{x}_{t^*}^s) Q_t^s y_t^s \leq (x_t^s - \tilde{x}_{t^*}^s) \ell_t^s + 2\varepsilon.
\]

Combining them with Eq. (22) and the fact that \( \eta \leq \frac{\gamma}{1 - \gamma} \leq \frac{1}{5} \), we get the first inequality that we want to prove. The other inequality is similar.

\[\blacksquare\]

Lemma 24  For all \( t \geq 1 \),

\[
\text{dist}_s^2(\tilde{z}_{t+1}^s) \leq \text{dist}_s^2(\tilde{z}_t^s) - 15\theta_{t+1}^s + \theta_t^s + 4\eta \Gamma_t + 8\eta^2 \|Q_t^s - Q_{t-1}^s\|^2 + 6\eta\varepsilon.
\]
Proof. Summing up the two inequalities in Lemma 23, we get

\[
2\eta(x_s^t - \widehat{x}_t^s)Q_t^s y_t^s + 2\eta x_t^s Q_t^s (\widehat{y}_t^s - y_t^s)
\]

\[
\leq \text{dist}_s^2(\mathcal{z}_t^s) - \text{dist}_s^2(\mathcal{z}_{t+1}^s) + \frac{4\eta^2}{(1 - \gamma)^2} \|z_t^s - z_{t-1}^s\|^2 + 8\eta^2\|Q_t^s - Q_{t-1}^s\|^2 - \|\mathcal{z}_{t+1}^s - z_{t-1}^s\|^2 - \|z_t^s - \mathcal{z}_t^s\|^2 + 6\eta \varepsilon
\]

\[
\leq \text{dist}_s^2(\mathcal{z}_t^s) - \text{dist}_s^2(\mathcal{z}_{t+1}^s) + \frac{1}{32}\|z_t^s - z_{t-1}^s\|^2 + 8\eta^2\|Q_t^s - Q_{t-1}^s\|^2 - \|\mathcal{z}_{t+1}^s - z_{t-1}^s\|^2 - \|z_t^s - \mathcal{z}_t^s\|^2 + 6\eta \varepsilon
\]

\[
\leq \text{dist}_s^2(\mathcal{z}_t^s) - \text{dist}_s^2(\mathcal{z}_{t+1}^s) + \frac{1}{16}\|z_t^s - \mathcal{z}_t^s\|^2 + \|z_t^s - z_{t-1}^s\|^2
\]

\[
+ 8\eta^2\|Q_t^s - Q_{t-1}^s\|^2 - \|\mathcal{z}_{t+1}^s - z_{t-1}^s\|^2 - \|z_t^s - \mathcal{z}_t^s\|^2 + 6\eta \varepsilon
\]

\[
= \text{dist}_s^2(\mathcal{z}_t^s) - \text{dist}_s^2(\mathcal{z}_{t+1}^s) + 8\eta^2\|Q_t^s - Q_{t-1}^s\|^2 - \frac{15}{16}\|z_t^s - \mathcal{z}_t^s\|^2 - \|\mathcal{z}_{t+1}^s - z_{t-1}^s\|^2 + \frac{1}{16}\|z_t^s - \mathcal{z}_t^s\|^2 + 6\eta \varepsilon
\]

The left-hand side above, can be lower bounded by

\[
2\eta(x_s^t - \widehat{x}_t^s)Q_t^s y_t^s + 2\eta x_t^s Q_t^s (\widehat{y}_t^s - y_t^s) = 2\eta x_t^s Q_t^s \widehat{y}_t^s - 2\eta \widehat{x}_t^s Q_t^s y_t^s
\]

\[
\geq 2\eta x_t^s Q_t^s \widehat{y}_t^s - 2\eta \widehat{x}_t^s Q_t^s y_t^s - 4\eta \Gamma_t
\]

\[
\geq -4\eta \Gamma_t. \quad \text{ (by the optimality of } \widehat{x}_t^s \text{ and } \widehat{y}_t^s \text{)}
\]

Combining the inequalities and using the definition of $\theta_t^s$ finish the proof.

Appendix C. Proof for Step 2: Lower Bounding $\|\mathcal{z}_{t+1}^s - z_t^s\|^2 + \|z_t^s - \mathcal{z}_t^s\|^2$

Lemma 25. For all $t \geq 1$, we have $20\theta_{t+1}^s + \eta \Gamma_t + 2\eta \varepsilon^2 \geq \frac{\eta^2}{64} (\Delta_{t+1}^s)^2$.

Proof. By Eq. (2) and the optimality condition for $\widehat{x}_{t+1}^s$, we have

\[
(\mathcal{z}_{t+1}^s - \widehat{x}_t^s + \eta \ell_t^s) \cdot (x^s - \widehat{x}_{t+1}^s) \geq 0 \quad \text{(23)}
\]

for any $x^s \in \Delta_A$. Then by the definition of $\ell_t^s$,

\[
(\mathcal{z}_{t+1}^s - \widehat{x}_t^s + \eta Q_t^s y_t^s) \cdot (x^s - \widehat{x}_{t+1}^s) \geq (\mathcal{z}_{t+1}^s - \widehat{x}_t^s + \eta \ell_t^s) \cdot (x^s - \widehat{x}_{t+1}^s) - 2\eta \varepsilon \geq -2\eta \varepsilon \quad \text{(24)}
\]

where in the last inequality we use Eq. (23). Thus we have for any $x^s \in \Delta_A$,

\[
\sqrt{2}(\|\mathcal{z}_{t+1}^s - x_t^s\| + \|x_t^s - \widehat{x}_t^s\|)
\]

\[
\geq \sqrt{2}\|\mathcal{z}_{t+1}^s - \widehat{x}_t^s\|
\]

\[
\geq \|\mathcal{z}_{t+1}^s - \widehat{x}_t^s\|^2
\]

\[
\geq (\mathcal{z}_{t+1}^s - \widehat{x}_t^s) \cdot (x^s - \widehat{x}_{t+1}^s)
\]

\[
\geq \eta (\mathcal{z}_{t+1}^s - x^s) Q_t^s y_t^s - 2\eta \varepsilon \quad \text{(by Eq. (24))}
\]

\[
= \eta (x_t^s - x^s) Q_t^s y_t^s + \eta (\mathcal{z}_{t+1}^s - x_t^s) Q_t^s y_t^s - 2\eta \varepsilon
\]

\[
\geq \eta (x_t^s - x^s) Q_t^s y_t^s - \frac{\eta \|\mathcal{z}_{t+1}^s - \widehat{x}_t^s\|}{1 - \gamma} - 2\eta \varepsilon.
\]
Using the fact that \( \frac{\eta}{1-\gamma} \leq \frac{1}{10} \), we get

\[
\|\tilde{x}_{t+1}^s - x_t^s\| + \|x_t^s - \tilde{x}_t^s\| + \sqrt{2}\eta\varepsilon \geq \frac{1}{\sqrt{2} + 1/10} \left( \eta \max_{x'} (x_t^s - x'^s) Q_t^s y_t^s \right) \geq \frac{\eta}{2} \max_{x'} (x_t^s - x'^s) Q_t^s y_t^s.
\]

Similarly, we have \( \|\tilde{y}_{t+1}^s - y_t^s\| + \|y_t^s - \tilde{y}_t^s\| + \sqrt{2}\eta\varepsilon \geq \frac{\eta}{2} \max_{y'} x_t^s Q_t^s (y'^s - y_t^s) \). Combining them and using \( \|z - z'\| \geq \frac{1}{2} \|x - x'\| + \frac{1}{2} \|y - y'\| \), we get

\[
\|\tilde{z}_{t+1}^s - z_t^s\| + \|z_t^s - \tilde{z}_t^s\| + \sqrt{2}\eta\varepsilon \\
\geq \frac{\eta}{4} \left( \max_{x'} (x_t^s - x'^s) Q_t^s y_t^s + \max_{y'} x_t^s Q_t^s (y'^s - y_t^s) \right) \\
\geq \frac{\eta}{4} \left( \max_{x'} x_t^s Q_t^s y_t^s - \min_{x'} x'^s Q_t^s y_t^s \right) \\
= \frac{\eta}{4} \left( \tilde{x}_{t+1}^s Q_t^s y_t^s + x_t^s (Q_t^s - Q_t^s) y_t^s + (x_t^s - \tilde{x}_{t+1}^s) Q_t^s y_t^s \right) \\
\geq \frac{\eta}{4} \left( \tilde{x}_{t+1}^s Q_t^s y_t^s - x_t^s Q_t^s \tilde{y}_{t+1}^s \right) - \frac{\eta}{4} \left( \max_{x'} Q_t^s \tilde{x}_{t+1}^s - \max_{y'} Q_t^s \tilde{y}_{t+1}^s \right) \\
\geq \frac{\eta}{4} \left( \tilde{x}_{t+1}^s Q_t^s y_t^s - x_t^s Q_t^s \tilde{y}_{t+1}^s \right) - \frac{\eta \Gamma_t}{2} - \frac{\eta}{4(1 - \gamma)} \left( \|\tilde{x}_{t+1}^s - x_t^s\| + \|\tilde{y}_{t+1}^s - y_t^s\| \right) \\
\geq \frac{\eta}{4} \Delta_{t+1}^s - \frac{\eta \Gamma_t}{2} - \frac{\eta}{4(1 - \gamma)} \|\tilde{x}_{t+1}^s - z_t^s\| \\
\geq \frac{\eta}{4} \Delta_{t+1}^s - \frac{\eta \Gamma_t}{2} - \frac{1}{16} \|\tilde{x}_{t+1}^s - z_t^s\|. \tag{25}
\]

Then notice that we have

\[
20 \theta_{t+1}^s + \eta \Gamma_t + 2\eta^2 \varepsilon^2 \\
\geq \frac{289}{256} \|\tilde{x}_{t+1}^s - z_t^s\|^2 + \|z_t^s - \tilde{z}_t^s\|^2 + \frac{\eta^2 \Gamma_t^2}{4} + 2\eta^2 \varepsilon^2 \\
\geq \frac{1}{4} \left( \frac{17}{16} \|\tilde{x}_{t+1}^s - z_t^s\|^2 + \|z_t^s - \tilde{z}_t^s\|^2 + \frac{\eta \Gamma_t}{2} + \sqrt{2}\eta\varepsilon \right)^2 \\
\geq \frac{\eta^2}{64} \left( \Delta_{t+1}^s \right)^2. \tag{by Eq. (25) and notice that \( \Delta_{t+1}^s \geq 0 \)}
\]

**Lemma 26 (Key Lemma for Average Duality-gap Bounds)** For all \( t \geq 1 \), we have

\[
\dist_t^s(\tilde{x}_{t+1}^s) \leq \dist_t^s(\tilde{x}_t^s) - 5\theta_{t+1}^s + \theta_t^s - \frac{\eta^2}{128} \left( \Delta_{t+1}^s \right)^2 + 5\eta \Gamma_t + 8\eta^2 \|Q_t^s - Q_{t-1}^s\|^2 + 7\eta\varepsilon.
\]
Proof Combining Lemma 25 with Lemma 24, we get
\[
\text{dist}_s^2(\tilde{z}_{t+1}^s) \leq \text{dist}_s^2(\hat{z}_t^s) - 5\theta_t^s + 4\eta \Gamma_t + 8\eta^2 \|Q_t^s - Q_{t-1}^s\|^2 + 6\eta \varepsilon
\]
\[
\leq \text{dist}_s^2(\hat{z}_t^s) - 5\theta_t^s - \left(\frac{\eta^2}{128} (\Delta_t^s + 1)^2 - \frac{1}{2}\eta \Gamma_t - \eta^2 \varepsilon^2\right) + \theta_t^s + 4\eta \Gamma_t + 8\eta^2 \|Q_t^s - Q_{t-1}^s\|^2 + 6\eta \varepsilon
\]
\[
\leq \text{dist}_s^2(\hat{z}_t^s) - 5\theta_t^s + \theta_t^s - \frac{\eta^2}{128} (\Delta_t^s + 1)^2 + 5\eta \Gamma_t + 8\eta^2 \|Q_t^s - Q_{t-1}^s\|^2 + 7\eta \varepsilon.
\]
\[\eta \varepsilon \leq 1\]

Lemma 27 (Key Lemma for Point-wise Convergence Bounds) There exists a constant $C' > 0$ (which depends on the transition and the loss/payoff functions) such that for all $t \geq 1$,
\[
\text{dist}_s^*(\tilde{z}_{t+1}^s) + 4.5\theta_t^s \leq \frac{1}{1 + \eta^2 C'^2} (\text{dist}_s^*(\hat{z}_t^s) + 4.5\theta_t^s) + 5\eta \Gamma_t + 8\eta^2 \|Q_t^s - Q_{t-1}^s\|^2 + 7\eta \varepsilon.
\]

Proof By Theorem 5 of (Wei et al., 2021) or Lemma 3 of (Gilpin et al., 2012), we have
\[
\Delta_t^s \geq C \text{dist}_s^*(\hat{z}_{t+1}^s)
\]
for some problem-dependent constant $0 < C \leq \frac{1}{1 - \gamma}$ ($C$ depends on $\{Q_t^s\}_s$). Thus Theorem 26 implies
\[
\text{dist}_s^2(\tilde{z}_{t+1}^s) + 5\theta_t^s \leq \text{dist}_s^2(\hat{z}_t^s) + \theta_t^s - \frac{\eta^2 C'^2}{128} \text{dist}_s^2(\hat{z}_{t+1}^s) + 5\eta \Gamma_t + 8\eta^2 \|Q_t^s - Q_{t-1}^s\|^2 + 7\eta \varepsilon.
\]
By defining $C'^2 = \frac{C'^2}{128}$, we further get
\[
\text{dist}_s^2(\tilde{z}_{t+1}^s) + \frac{5}{1 + \eta^2 C'^2} \theta_t^s \leq \frac{1}{1 + \eta^2 C'^2} (\text{dist}_s^2(\hat{z}_t^s) + \theta_t^s + 5\eta \Gamma_t + 8\eta^2 \|Q_t^s - Q_{t-1}^s\|^2 + 7\eta \varepsilon)
\]
\[
\leq \frac{1}{1 + \eta^2 C'^2} (\text{dist}_s^2(\hat{z}_t^s) + \theta_t^s) + 5\eta \Gamma_t + 8\eta^2 \|Q_t^s - Q_{t-1}^s\|^2 + 7\eta \varepsilon.
\]
Notice that $\frac{5}{1 + \eta^2 C'^2} \geq \frac{5}{1 + \eta^2} \geq 4.5$. Thus we further have
\[
\text{dist}_s^2(\tilde{z}_{t+1}^s) + 4.5\theta_t^s \leq \frac{1}{1 + \eta^2 C'^2} (\text{dist}_s^2(\hat{z}_t^s) + 4.5\theta_t^s) + 5\eta \Gamma_t + 8\eta^2 \|Q_t^s - Q_{t-1}^s\|^2 - 3\theta_t^s + 7\eta \varepsilon
\]
where in the last inequality we use $\frac{1}{1 + \eta^2 C'^2} \leq \frac{4.5}{1 + \eta^2 C'^2}$ because $\eta^2 C'^2 \leq \frac{1}{128}$. ■

Appendix D. Proof for Step 3: Bounding $\|Q_t^s - Q_{t-1}^s\|^2$

Lemma 28 We have for $t \geq 2$ and all $s \in S$,
\[
\|Q_t^s - Q_{t-1}^s\|^2 \leq \frac{8\gamma^2}{(1 - \gamma)^2} J_{t-1} + \frac{2\gamma^2}{1 + \gamma} K_{t-1} + \frac{16\gamma^2 \varepsilon^2}{1 - \gamma}.
\]
Proof. It is equivalent to prove that for all $t \geq 1$,
\[
\|Q_{t+1}^* - Q_t^*\|^2 \leq \frac{8\gamma^2}{(1-\gamma)^2}J_t + \frac{2\gamma^2}{1+\gamma}K_t + \frac{16\gamma^2\varepsilon^2}{1-\gamma}.
\]

By definition,
\[
Q_t^*(a, b) = \sigma(s, a, b) + \gamma \mathbb{E}_{s' \sim p(\cdot|s, a, b)} \left[ V_{s'}^t \right],
\]
we have
\[
\|Q_{t+1}^* - Q_t^*\|^2 = \max_{a, b} (Q_{t+1}^*(a, b) - Q_t^*(a, b))^2 \leq \gamma^2 \max_{s'} (V_{s'}^t - V_{s'}^{t-1})^2
\]
(26)

Now it suffices to upper bound $(V_s^t - V_{s-1}^t)^2$ for any $s$. By Corollary 22, we have $V_{t-1}^s = \sum_{\tau=1}^{t-1} \alpha_{t-1}^\tau \rho_t^\tau$. Therefore,
\[
V_s^t - V_{s-1}^t = \alpha_t \left( \rho_t^s - V_{t-1}^s \right)
\]
\[
= \alpha_t \left( \rho_t^s - \sum_{\tau=0}^{t-1} \alpha_{t-1}^\tau \rho_t^\tau \right)
\]
\[
= \alpha_t \left( \sum_{\tau=0}^{t-1} \alpha_{t-1}^\tau (\rho_t^s - \rho_t^\tau) \right) \quad \text{(because } \sum_{\tau=0}^{t-1} \alpha_{t-1}^\tau = 1)\]

In the following calculation, we omit the superscript $s$ for simplicity. By defining $\text{diff}_h \triangleq |\rho_h - \rho_{h-1}|$, we have
\[
(V_t - V_{t-1})^2 \leq (\alpha_t)^2 \left( \sum_{\tau=0}^{t-1} \alpha_{t-1}^\tau (\rho_t - \rho_{\tau}) \right)^2
\]
\[
\leq (\alpha_t)^2 \left( \sum_{\tau=0}^{t-1} \alpha_{t-1}^\tau \sum_{h=\tau+1}^{t} (\rho_h - \rho_{h-1}) \right)^2
\]
\[
\leq (\alpha_t)^2 \left( \sum_{\tau=0}^{t-1} \alpha_{t-1}^\tau \sum_{h=\tau+1}^{t} \text{diff}_h \right)^2
\]
\[
= (\alpha_t)^2 \left( \sum_{h=1}^{t-1} \sum_{\tau=0}^{h-1} \alpha_{t-1}^\tau \text{diff}_h \right)^2
\]
\[
\leq (\alpha_t)^2 \left( \sum_{h=1}^{t-1} \sum_{\tau=0}^{h-1} \delta_{t-1}^{h-1} \text{diff}_h \right)^2.
\]
(by Lemma 35)

Then we continue:
\[
(V_t - V_{t-1})^2 \leq (\alpha_t)^2 \left( \sum_{h=1}^{t} \delta_{t-1}^{h-1} \text{diff}_h \right)^2
\]

23
\[
\leq (\alpha_t)^2 \left( \sum_{h=1}^{t-1} \delta_{h-1} h \right) \left( \sum_{h=1}^{t} \delta_{h-1} h \right) (\text{Cauchy-Schwarz inequality})
\]
\[
\leq \left( \sum_{h=1}^{t} \alpha_h \delta_{h-1} h \right) \left( \sum_{h=1}^{t} \alpha_h \delta_{h-1} h \right) (\alpha_t \leq \alpha_h \text{ for } h \leq t)
\]
\[
\leq \sum_{\tau=1}^{t} \alpha_\tau \text{diff}_\tau^2 \quad \text{(note that } \alpha_h \delta_{h-1} = \alpha_h \prod_{\tau=h}^{t-1} (1-\alpha_\tau) \leq \alpha_h \prod_{\tau=h+1}^{t} (1-\alpha_\tau) = \alpha_h^t)\]
\[
= \sum_{\tau=1}^{t} \alpha_\tau \left( \rho_\tau - x_\tau Q_\tau y_\tau + x_\tau (Q_\tau - Q_{\tau-1}) y_\tau + (x_\tau - x_{\tau-1}) Q_{\tau-1} y_\tau 
\quad + x_{\tau-1} Q_{\tau-1} (y_\tau - y_{\tau-1}) + x_{\tau-1} Q_{\tau-1} y_{\tau-1} - \rho_{t-1} \right)^2
\leq \sum_{\tau=1}^{t} \alpha_\tau \left( \frac{8 \varepsilon^2}{1-\gamma} + \frac{2}{1+\gamma} \|Q_\tau - Q_{\tau-1}\|^2 + \frac{8}{(1-\gamma)^3} \|x_\tau - x_{\tau-1}\|^2 + \frac{8}{(1-\gamma)^3} \|y_\tau - y_{\tau-1}\|^2 + \frac{8 \varepsilon^2}{1-\gamma} \right),
\]
where we use \((a + b + c + d + e)^2 \leq \frac{8}{1-\gamma} a^2 + \frac{2}{1+\gamma} b^2 + \frac{8}{1-\gamma} c^2 + \frac{8}{1-\gamma} d^2 + \frac{8}{1-\gamma} e^2\) which is due to Cauchy-Schwarz inequality. By Lemma 21 and the definitions of \(J_t^s\), \(K_t^s\), \(J_t\), \(K_t\) in Definition 13 and Definition 14,
\[
\sum_{\tau=1}^{t} \alpha_\tau \|Q_\tau - Q_{\tau-1}\|^2 = K_t^s \leq K_t,
\]
\[
\sum_{\tau=1}^{t} \alpha_\tau \|z_\tau - z_{\tau-1}\|^2 \leq J_t^s \leq J_t.
\]
Combining them with the previous upper bound for \((V_t^s - V_{t-1}^s)^2\), we get
\[
(V_t^s - V_{t-1}^s)^2 \leq \frac{8}{(1-\gamma)^3} J_t + \frac{2}{1+\gamma} K_t + \frac{16 \varepsilon^2}{1-\gamma}
\]
for all \(s\). Further combining this with Eq. (26), we get
\[
\|Q_t^s - Q_{t-1}^s\|^2 \leq \frac{8 \gamma^2}{(1-\gamma)^3} J_t + \frac{2 \gamma^2}{1+\gamma} K_t + \frac{16 \gamma^2 \varepsilon^2}{1-\gamma}.
\]

**Appendix E. Proof for Steps 4 and 5: Bounding \(\|Q_t^s - Q_{t-1}^s\|\)**

**Lemma 29** For all \(t \geq 2\),
\[
\Gamma_t \leq \gamma \left( \sum_{\tau=1}^{t-1} \alpha_{\tau-1} \Gamma_\tau + \text{Reg}_{t-1} + \varepsilon \right).
\]
**Proof** We proceed with

\[ Q^s_{t+1}(a, b) = \sigma(s, a, b) + \gamma \mathbb{E}_{s' \sim p(\cdot | s, a, b)} \left[ V^s_t \right] \]

\[ = \sigma(s, a, b) + \gamma \mathbb{E}_{s' \sim p(\cdot | s, a, b)} \left[ \sum_{\tau=1}^{t} \alpha^t_{\tau} \rho^s_{\tau} \right] \quad \text{(Corollary 22)} \]

\[ \leq \sigma(s, a, b) + \gamma \mathbb{E}_{s' \sim p(\cdot | s, a, b)} \left[ \sum_{\tau=1}^{t} \alpha^t_{\tau} x^s_{t, \tau} Q^{s'}_{t, \tau} y^{s'}_{t, \tau} + \varepsilon \right] \]

(by the definition of \( \rho^s_{\tau} \) and that \( \sum_{\tau=1}^{t} \alpha^t_{\tau} = 1 \) for \( t \geq 1 \))

\[ \leq \sigma(s, a, b) + \gamma \mathbb{E}_{s' \sim p(\cdot | s, a, b)} \left[ \sum_{\tau=1}^{t} \alpha^t_{\tau} x^s_{t, \tau} Q^{s'}_{t, \tau} y^{s'}_{t, \tau} + \sum_{\tau=1}^{t} \alpha^t_{\tau} \Gamma_{t, \tau} + \text{Reg}_t + \varepsilon \right] \]

(by the definition of \( \Gamma_{t, \tau} \))

\[ \leq \sigma(s, a, b) + \gamma \mathbb{E}_{s' \sim p(\cdot | s, a, b)} \left[ \sum_{\tau=1}^{t} \alpha^t_{\tau} x^s_{t, \tau} Q^{s'}_{t, \tau} y^{s'}_{t, \tau} + \sum_{\tau=1}^{t} \alpha^t_{\tau} \Gamma_{t, \tau} \right] + \varepsilon \]

(by definition of \( y^{s'}_{t, \tau} \))

\[ = \sigma(s, a, b) + \gamma \mathbb{E}_{s' \sim p(\cdot | s, a, b)} \left[ V^s_t \right] + \gamma \left( \sum_{\tau=1}^{t} \alpha^t_{\tau} \Gamma_{t, \tau} \right) \quad \text{(by definition of } y^{s'}_{t, \tau}) \]

\[ = Q^s_t(a, b) + \gamma \left( \sum_{\tau=1}^{t} \alpha^t_{\tau} \Gamma_{t, \tau} \right) + \text{Reg}_t + \varepsilon \].

Similarly,

\[ Q^s_{t+1}(a, b) \geq Q^s_t(a, b) - \gamma \left( \sum_{\tau=1}^{t} \alpha^t_{\tau} \Gamma_{t, \tau} \right) \]

They jointly imply

\[ \Gamma_{t+1} \leq \gamma \left( \sum_{\tau=1}^{t} \alpha^t_{\tau} \Gamma_{t, \tau} \right) + \text{Reg}_t + \varepsilon \].

\[ \text{Lemma 30} \quad \text{For any state } s \text{ and time } t \geq 1, \]

\[ \text{Reg}_t \leq \frac{1}{2\eta} Z_t + \frac{4\eta}{(1-\gamma)^2} J_t + 4\eta K_t + 3\varepsilon. \]
Proof Summing the first bound in Lemma 23 over $\tau = 1, \ldots, t$ with weights $\alpha_t^\tau$, and dropping negative terms $-\|\bar{x}_{t+1}^s - x_t^s\|^2 - \|x_t^s - \bar{x}_t^s\|^2$, we get

$$\sum_{\tau=1}^{t} \alpha_t^\tau (x_t^s - \bar{x}_t^s) Q_t^s y_t^s$$

$$\leq \sum_{\tau=1}^{t} \frac{\alpha_t^\tau}{2\eta} \left(\operatorname{dist}_1^s(\bar{x}_t^s) - \operatorname{dist}_1^s(\bar{x}_{t+1}^s)\right) + \frac{4\eta}{(1-\gamma)^2} \sum_{\tau=1}^{t} \alpha_t^\tau \|y_t^s - y_{t-1}^s\|^2 + 4\eta \sum_{\tau=1}^{t} \alpha_t^\tau \|Q_t^s - Q_{t-1}^s\|^2 + 3\varepsilon$$

$$\leq \frac{\alpha_t^1}{2\eta} \operatorname{dist}_1^s(\bar{x}_1^s) + \sum_{\tau=2}^{t} \frac{\alpha_t^\tau - \alpha_t^\tau-1}{2\eta} \operatorname{dist}_1^s(\bar{x}_t^s) + \frac{4\eta}{(1-\gamma)^2} \sum_{\tau=1}^{t} \alpha_t^\tau \|y_t^s - y_{t-1}^s\|^2 + 4\eta \sum_{\tau=1}^{t} \alpha_t^\tau \|Q_t^s - Q_{t-1}^s\|^2 + 3\varepsilon$$

Observe that by definition, we have for $\tau \geq 2$,

$$\alpha_t^\tau - \alpha_t^\tau-1 = \alpha_t^\tau \left(1 - \frac{\alpha_{\tau-1}(1 - \alpha_{\tau})}{\alpha_{\tau}}\right) = \alpha_t^\tau \times \frac{\alpha_{\tau} - \alpha_{\tau-1} + \alpha_{\tau-1}\alpha_{\tau}}{\alpha_{\tau}} \leq \alpha_{\tau-1}\alpha_t^\tau$$

where in the inequality we use $\alpha_{\tau} \leq \alpha_{\tau-1}$. Using this in Eq. (27), we get

$$\sum_{\tau=1}^{t} \alpha_t^\tau (x_t^s - \bar{x}_t^s) Q_t^s y_t^s \leq \frac{\alpha_t^1}{2\eta} \operatorname{dist}_1^s(\bar{x}_1^s) + \frac{1}{2\eta} \sum_{\tau=2}^{t} \alpha_t^\tau \alpha_{\tau-1} \operatorname{dist}_1^s(\bar{x}_t^s) + \frac{4\eta}{(1-\gamma)^2} J_t^s + 4\eta K_t^s + 3\varepsilon$$

$$= \frac{1}{2\eta} \sum_{\tau=1}^{t} \alpha_t^\tau \alpha_{\tau-1} \operatorname{dist}_1^s(\bar{x}_t^s) + \frac{4\eta}{(1-\gamma)^2} J_t^s + 4\eta K_t^s + 3\varepsilon$$

(recall that $\alpha_0 = 1$)

Using $J_n^s \leq J, K_n^s \leq K$, and the definition of $Z_t, \overline{\operatorname{Reg}_t}$, finishes the proof. □

Appendix F. Combining Lemmas to Show Last-iterate Convergence

In this section, we provide proofs for Theorem 1 and Theorem 2. To achieve so, we first prove Lemma 31 by combining the results in Appendix D and Appendix E. Then we combine Theorem 26, Lemma 27, and Lemma 31 to prove Theorem 1 and Theorem 2.

Lemma 31 For any $s$ and $t \geq 1$,

$$5\eta \Gamma_t + 8\eta^2 \|Q_t^s - Q_{t-1}^s\|^2$$

$$\leq \max_{s'} \left(\frac{C_1\eta^2}{(1-\gamma)^4} \sum_{\tau=1}^{t-1} \beta_t^\tau \left(\theta_{s'}^\tau + \theta_{s'+1}^\tau\right) + C_2 \sum_{\tau=1}^{t-1} \beta_t^\tau \alpha_{\tau-1} \operatorname{dist}_2^s(\bar{x}_{s'})\right) + 80\beta_t^1 + \frac{80\eta^2}{(1-\gamma)^2}.$$
Proof By Lemma 28, for all \( t \geq 2 \),
\[
\eta^2 \| Q^s_t - Q^s_{t-1} \|^2 \leq \frac{8\eta^2 \gamma^2}{(1-\gamma)^3} J_{t-1} + \frac{2\eta^2 \gamma^2}{1+\gamma} K_{t-1} + \frac{16\eta^2 \varepsilon^2}{1-\gamma}.
\] (28)

By Lemma 29 and Lemma 30, for all \( t \geq 2 \),
\[
\eta \Gamma_t \leq \gamma \sum_{\tau=1}^{t-1} \alpha^\tau_{t-1} \eta \Gamma_{\tau} + \frac{4\eta^2}{(1-\gamma)^2} J_{t-1} + \frac{4\eta^2 K_{t-1}}{2} + \frac{1}{2} Z_{t-1} + 4\eta \varepsilon.
\] (29)

Now, multiply Eq. (29) with \( \frac{1-\gamma}{16} \), and then add it to Eq. (28). Then we get that for \( t \geq 2 \),
\[
\eta^2 \| Q^s_t - Q^s_{t-1} \|^2 + \frac{1-\gamma}{16} \eta \Gamma_t \\
\leq \gamma \sum_{\tau=1}^{t-1} \alpha^\tau_{t-1} \left( \frac{1-\gamma}{16} \eta \Gamma_{\tau} \right) + \left( \frac{8\gamma^2}{(1-\gamma)^3} + \frac{1}{4(1-\gamma)} \right) \eta^2 J_{t-1} + \\
\quad \left( \frac{2\gamma^2}{1+\gamma} + \frac{1-\gamma}{4} \right) \eta^2 K_{t-1} + \frac{(1-\gamma)Z_{t-1}}{32} + \frac{\eta \varepsilon}{2}
\] (see explanation below)
\[
\leq \gamma \sum_{\tau=1}^{t-1} \alpha^\tau_{t-1} \left( \frac{1-\gamma}{16} \eta \Gamma_{\tau} \right) + \frac{9}{(1-\gamma)^3} \eta^2 J_{t-1} + \frac{\gamma \eta^2 K_{t-1}}{2} + \frac{(1-\gamma)Z_{t-1}}{32} + \frac{\eta \varepsilon}{2},
\]

where in the second inequality we use that \( \frac{2\gamma^2}{1+\gamma} + \frac{1-\gamma}{4} - \gamma = (1-\gamma) \left( \frac{1}{4} - \frac{\gamma}{1+\gamma} \right) \leq 0 \) since \( \gamma \geq \frac{1}{2} \),
and in the last inequality, we use \( K^s_{t-1} = \sum_{\tau=1}^{t-1} \alpha^\tau_{t-1} \| Q^s_\tau - Q^s_{\tau-1} \|^2 \). Define the new variable
\[
u_t = \eta^2 \max_s \| Q^s_t - Q^s_{t-1} \|^2 + \frac{1-\gamma}{16} \eta \Gamma_t.
\]

Then the above implies that for all \( t \geq 2 \),
\[
u_t \leq \gamma \sum_{\tau=1}^{t-1} \alpha^\tau_{t-1} \nu_{\tau} + \frac{9}{(1-\gamma)^3} \eta^2 J_{t-1} + \frac{(1-\gamma)Z_{t-1}}{32} + \frac{\eta \varepsilon}{2}.
\] (30)

Observe that Eq. (30) is in the form of Lemma 33 with the following choices:
\[
g_t = \nu_t, \quad \forall t \geq 1,
\]
\[
h_t = \begin{cases} 
\nu_t + \frac{\eta \varepsilon}{2} & \text{for } t = 1 \\
\frac{9\eta^2}{(1-\gamma)^3} J_{t-1} + \frac{(1-\gamma)Z_{t-1}}{32} + \frac{\eta \varepsilon}{2} & \text{for } t \geq 2
\end{cases}
\]
and get that for \( t \geq 2 \),
\[
u_t \leq \frac{9\eta^2}{(1-\gamma)^3} \sum_{\tau=2}^{t} \beta^\tau_{t-1} J_{\tau-1} + \frac{1-\gamma}{32} \sum_{\tau=2}^{t} \beta^\tau_{t-1} Z_{\tau-1} + \frac{\eta \varepsilon}{2} \sum_{\tau=1}^{t} \beta^\tau_{t-1} + \frac{\eta \varepsilon}{1-\gamma}
\] (by Lemma 38)
because \( u_t \leq \frac{\eta^2}{(1 - \gamma)^2} + \frac{1 - \gamma}{16} \leq \frac{1 - \gamma}{2} \). Further using Lemma 34 on the first two terms on the right-hand side, and noticing that \( \frac{1 - \gamma}{2} \leq 4 \), we further get that for \( t \geq 2 \),

\[
\begin{aligned}
\quad u_t & \leq \max_s \frac{36\eta^2}{(1 - \gamma)^3} \sum_{\tau=1}^{t-1} \beta_t^\tau \| z_{\tau}^s - z_{\tau-1}^s \|^2 + 1 - \frac{\gamma}{8} \sum_{\tau=1}^{t-1} \beta_t^\tau \alpha_{\tau-1} \text{dist}^2(z_{\tau}^s) + (1 - \gamma)\beta_t^1 + \frac{\eta\varepsilon}{1 - \gamma} \\
& \leq \max_s \frac{72\eta^2}{(1 - \gamma)^3} \sum_{\tau=1}^{t-1} \beta_t^\tau \left( \| z_{\tau}^s - z_{\tau-1}^s \|^2 + \| z_{\tau}^s - z_{\tau-1}^s \|^2 \right) + 1 - \frac{\gamma}{8} \sum_{\tau=1}^{t-1} \beta_t^\tau \alpha_{\tau-1} \text{dist}^2(z_{\tau}^s) + (1 - \gamma)\beta_t^1 + \frac{\eta\varepsilon}{1 - \gamma} \\
& \leq \max_s \frac{1152\eta^2}{(1 - \gamma)^3} \sum_{\tau=1}^{t-1} \beta_t^\tau (\theta_{\tau+1}^s + \theta_{\tau}^s) + 1 - \frac{\gamma}{8} \sum_{\tau=1}^{t-1} \beta_t^\tau \alpha_{\tau-1} \text{dist}^2(z_{\tau}^s) + (1 - \gamma)\beta_t^1 + \frac{\eta\varepsilon}{1 - \gamma}.
\end{aligned}
\]

(31)

Finally, notice that according to the definition of \( u_t \), we have \( 5\eta \Gamma_t + 8\eta^2 \| Q_t^s - Q_{t-1}^s \| \leq \frac{80}{1 - \gamma} u_t \). Combining Eq. (31), we finish the proof for case for \( t \geq 2 \). The case for \( t = 1 \) is trivial since \( 5\eta \Gamma_1 + 8\eta^2 \| Q_1^s - Q_{1-1}^s \| \leq 1 \leq 80 = 80\beta_1 \).

**Proof of Theorem 1.** Define \( C_\alpha(T) \triangleq 1 + \sum_{t=1}^{T} \alpha_t \) and let \( C_\beta \) be an upper bound of \( \sum_{t=1}^{\infty} \beta_t^\tau \) for any \( \tau \). With the choice of \( \alpha_t \) specified in the theorem, we have \( C_\alpha(T) = 1 + \sum_{t=1}^{T} \frac{H + 1}{H + t} = \mathcal{O}(H \log T) = \mathcal{O}\left(\frac{\log T}{1 - \gamma}\right) \). By Lemma 40, we have \( C_\beta \leq \frac{2}{1 - \gamma} + 3 \). Define \( S = |S| \).

Combining Lemma 31 and Theorem 26, we get that for \( t \geq 1 \),

\[
\left( \frac{\eta^2}{128} \right) (\Delta_{t+1}^s)^2 \leq \text{dist}^2(z_t^s) - \text{dist}^2(z_{t+1}^s) - 5\theta_{t+1}^s + \theta_t^s \\
+ \max_{s'} \left( \frac{C_1 \eta^2}{(1 - \gamma)^4} \sum_{\tau=1}^{t-1} \beta_t^\tau (\theta_{\tau+1}^{s'} + \theta_{\tau}^{s'}) + \frac{C_2}{(1 - \gamma)^4} \sum_{\tau=1}^{t-1} \beta_t^\tau \alpha_{\tau-1} \text{dist}^2(z_{\tau}^{s'}) \right) \right) + 80\beta_t^1 + \frac{87\eta\varepsilon}{(1 - \gamma)^2}.
\]

Summing the above over \( s \in S \) and \( t \in [T - 1] \), and denoting \( \Theta_t = \sum_{s} \theta_t^s \), we get

\[
\left( \frac{\eta^2}{128} \right) \sum_{t=1}^{T} \sum_{s} (\Delta_t^s)^2 \leq \mathcal{O}(S) - 4\Theta_t + \frac{C_1 S \eta^2}{(1 - \gamma)^4} \sum_{t=1}^{T} \sum_{\tau=1}^{t-1} \beta_t^\tau \text{dist}^2(z_{\tau}^{s'}) + \Theta_{t+1} \right) \\
+ \mathcal{O}\left( S \sum_{t=1}^{T} \sum_{\tau=1}^{t-1} \beta_t^\tau \alpha_{\tau-1} + S \sum_{t=1}^{T} \beta_t^1 + \frac{S \eta\varepsilon T}{(1 - \gamma)^2} \right) \right)
\]

(32)

since \( \text{dist}^2(z_t^s) = \mathcal{O}(1) \) and \( \Theta_1 = \mathcal{O}(S) \). Notice that the following hold

\[
\sum_{t=1}^{T} \sum_{\tau=1}^{t-1} \beta_t^\tau (\Theta_{\tau} + \Theta_{\tau+1}) \leq \sum_{t=1}^{T} \sum_{\tau=1}^{t-1} \beta_t^\tau (\Theta_{\tau} + \Theta_{\tau+1}) \leq 2C_\beta \sum_{\tau=1}^{T} \Theta_{\tau}.
\]

\[
\sum_{t=1}^{T} \sum_{\tau=1}^{t-1} \beta_t^\tau \alpha_{\tau-1} \leq \sum_{t=1}^{T} \sum_{\tau=1}^{t-1} \beta_t^\tau \alpha_{\tau-1} \leq C_\beta \sum_{\tau=1}^{T} \alpha_{\tau-1} = C_\alpha(T) C_\beta
\]

28
By Cauchy-Schwarz inequality, we further have

\[ \sum_{t=1}^{T} \sum_{s} (\Delta_t^s)^2 \leq \frac{128}{\eta^2} \sum_{t=1}^{T} \left( -4 + 2C_\beta C_1 S \eta^2 (1-\gamma)^4 \right) \Theta_t + O \left( \frac{SC_\alpha(T)C_\beta}{\eta^2} + \frac{S\varepsilon T}{\eta(1-\gamma)^2} \right) \]

\[ = O \left( \frac{SC_\alpha(T)C_\beta}{\eta^2} + \frac{S\varepsilon T}{\eta(1-\gamma)^2} \right). \]

(by our choice of \( \eta \), we have \( 2C_\beta C_1 S \eta^2 \leq \frac{10C_1 S \eta^2}{(1-\gamma)} \leq 4 \))

By Cauchy-Schwarz inequality, we further have

\[ \sum_{t=1}^{T} \sum_{s} \Delta_t^s \leq \sqrt{ST} \left( \sum_{t=1}^{T} \sum_{s} (\Delta_t^s)^2 \right)^{1/2} = O \left( \frac{S\sqrt{C_\alpha(T)C_\beta T}}{\eta} + \frac{ST \sqrt{\varepsilon}}{\sqrt{\eta}(1-\gamma)} \right). \]

Finally, by Lemma 32, we get

\[ \frac{1}{T} \max_{s,t} \left( V_{s,t}^s - V_{s,t}^t_{\hat{g}_t} \right) \leq \frac{2}{1-\gamma} \frac{1}{T} \max_{s} \sum_{t=1}^{T} \Delta_t^s = O \left( \frac{S\sqrt{C_\alpha(T)C_\beta}}{\eta(1-\gamma)\sqrt{T}} + \frac{S\sqrt{\varepsilon}}{\sqrt{\eta}(1-\gamma)^2} \right) \]

\[ = O \left( \frac{S\sqrt{\log T}}{\eta(1-\gamma)^2\sqrt{T}} + \frac{S\sqrt{\varepsilon}}{\sqrt{\eta}(1-\gamma)^2} \right). \]

**Proof of Theorem 2.** Combining Lemma 27 and Lemma 31, we get that for all \( t \geq 1 \),

\[ \text{dist}_s^2(\hat{z}_{t+1}^s) + 4.5\theta_{t+1}^s \]

\[ \leq \frac{1}{1 + \eta^2 C_\beta^2} \left( \text{dist}_s^2(\hat{z}_t^s) + 4.5\theta_t^s \right) \]

\[ + \max_{s'} \left( \frac{C_1 \eta^2}{(1-\gamma)^4} \sum_{\tau=1}^{t-1} \beta_{\tau}^s (\theta_{\tau}^{s'} + \theta_{\tau+1}^{s'}) + C_2 \sum_{\tau=1}^{t-1} \beta_{\tau}^s \alpha_{\tau-1} \text{dist}_s^2(\hat{z}_{\tau}^{s'}) \right) + 80\beta_1^s - 3\theta_t^s + \frac{87\eta \varepsilon}{(1-\gamma)^2}. \]

Summing the above inequality over \( s \in \mathcal{S} \), and denoting \( L_t = \sum_s \text{dist}_s^2(\hat{z}_t^s) \), \( \Theta_t = \sum_s \theta_t^s \), we get that for all \( t \),

\[ L_{t+1} + 4.5\Theta_{t+1} \leq \frac{1}{1 + \eta^2 C_\beta^2} (L_t + 4.5\Theta_t) + \frac{C_1 S \eta^2}{(1-\gamma)^4} \sum_{\tau=1}^{t-1} \beta_{\tau}^s (\Theta_{\tau} + \Theta_{\tau+1}) + \]

\[ C_2 S \sum_{\tau=1}^{t-1} \beta_{\tau}^s \alpha_{\tau-1} L_{\tau} + 80S\beta_1^s - 3\Theta_t + \frac{87S\eta \varepsilon}{(1-\gamma)^2}. \]  

(33)

The key idea of the following analysis is to use the negative (bonus) term \(-3\Theta_t\) to cancel the positive (penalty) term \( \frac{C_1 S \eta^2}{(1-\gamma)^4} \sum_{\tau=1}^{t} \beta_{\tau}^s \Theta_{\tau} \). Since the time indices do not match, we perform *smoothing over time* to help. Consider the following weighted sum of \( L_{\tau} + 4.5\Theta_{\tau} \) with weights \( \lambda^\tau_{t+1} = \sum_{\tau=2}^{t+1} \lambda^\tau_{t+1} (L_{\tau} + 4.5\Theta_{\tau}) \)
\[
\sum_{\tau=1}^{t} \lambda_{\tau+1}^{t}(L_{\tau+1} + 4.5\Theta_{\tau+1}) \quad \text{(re-indexing)}
\]
\[
\leq \sum_{\tau=1}^{t} \lambda_{\tau}^{t}(L_{\tau} + 4.5\Theta_{\tau+1}) \quad \text{(Lemma 37)}
\]
\[
\leq \frac{1}{1 + \eta^{2}C''^{2}} \sum_{\tau=1}^{t} \lambda_{\tau}^{t}(L_{\tau} + 4.5\Theta_{\tau}) + \frac{C_{1}S\eta^{2}}{(1-\gamma)^4} \sum_{\tau=1}^{t} \lambda_{\tau}^{t} \sum_{i=1}^{t-1} \beta_{i} \alpha \sum_{i=1}^{t-1} \beta_{i} \Theta_{\tau} + 8\frac{C\eta}{(1-\gamma)^2} \sum_{\tau=1}^{t} \lambda_{\tau}^{t}
\]
\[
\leq \frac{1}{1 + \eta^{2}C''^{2}} \sum_{\tau=1}^{t} \lambda_{\tau}^{t}(L_{\tau} + 4.5\Theta_{\tau}) + \frac{C_{1}S\eta^{2}}{(1-\gamma)^4} \sum_{\tau=1}^{t-1} \left( \sum_{i=1}^{t} \lambda_{i}^{t} \beta_{i} \right) \left( \Theta_{\tau} + \Theta_{\tau+1} \right)
\]
\[
\leq \frac{1}{1 + \eta^{2}C''^{2}} \sum_{\tau=1}^{t} \lambda_{\tau}^{t}(L_{\tau} + 4.5\Theta_{\tau}) + \frac{3C_{1}S\eta^{2}}{(1-\gamma)^5} \sum_{\tau=1}^{t-1} \lambda_{\tau}^{t} \Theta_{\tau} + \frac{8\frac{C\eta}{(1-\gamma)^2}}{(1-\gamma)^2} \sum_{\tau=1}^{t} \lambda_{\tau}^{t} \quad \text{(by Lemma 36)}
\]
\[
\leq \frac{1}{1 + \eta^{2}C''^{2}} \sum_{\tau=1}^{t} \lambda_{\tau}^{t}(L_{\tau} + 4.5\Theta_{\tau}) + \frac{6C_{1}S\eta^{2}}{(1-\gamma)^5} \sum_{\tau=1}^{t} \lambda_{\tau}^{t} \Theta_{\tau} + \frac{8\frac{C\eta}{(1-\gamma)^2}}{(1-\gamma)^2} \sum_{\tau=1}^{t} \lambda_{\tau}^{t}
\]
\[
\leq \frac{1}{1 + \eta^{2}C''^{2}} \sum_{\tau=1}^{t} \lambda_{\tau}^{t}(L_{\tau} + 4.5\Theta_{\tau}) + \frac{3C_{1}S\eta^{2}}{(1-\gamma)^4} \sum_{\tau=1}^{t} \lambda_{\tau}^{t} \Theta_{\tau} + \frac{8\frac{C\eta}{(1-\gamma)^2}}{(1-\gamma)^2} \sum_{\tau=1}^{t} \lambda_{\tau}^{t}
\]
\[
\leq \frac{1}{1 + \eta^{2}C''^{2}} \sum_{\tau=1}^{t} \lambda_{\tau}^{t}(L_{\tau} + 4.5\Theta_{\tau}) + \frac{12C_{1}S\eta}{1-\gamma} \sum_{\tau=1}^{t} \lambda_{\tau}^{t} \Theta_{\tau} + \frac{8\frac{C\eta}{(1-\gamma)^2}}{(1-\gamma)^2} \sum_{\tau=1}^{t} \lambda_{\tau}^{t}
\]
\[
\leq \frac{1}{1 + 0.1\eta^{2}C''^{2}} \sum_{\tau=3}^{t} \lambda_{\tau}^{t}(L_{\tau} + 4.5\Theta_{\tau}) + \frac{12C_{1}S\eta}{1-\gamma} \sum_{\tau=3}^{t} \lambda_{\tau}^{t} \Theta_{\tau} + \frac{8\frac{C\eta}{(1-\gamma)^2}}{(1-\gamma)^2} \sum_{\tau=3}^{t} \lambda_{\tau}^{t}. \quad (\eta C'' \leq 2^{-15} \text{ according to Lemma 27, } L_{\tau} \leq S \cdot \max_{z,z'} ||z - z'||^2 \leq 4S)
\]
Finally, we add $\lambda_{t+1}^1(L_1 + 4.5\Theta_1)$ to both sides, and note that

$$\lambda_{t+1}^1(L_1 + 4.5\Theta_1) = \alpha_{t+1}(L_1 + 4.5\Theta_1) \leq \alpha_t(L_1 + 4.5\Theta_1) \leq \alpha_t \cdot 22S = 22S\lambda_t^1,$$

where the first and second equality is by Lemma 41. Then we get

$$\sum_{\tau=1}^{t+1} \lambda_{t+1}^\tau(L_\tau + 4.5\Theta_\tau) \leq \frac{1}{1 + 0.1\eta^2C^2} \sum_{\tau=1}^{t} \lambda_t^\tau(L_\tau + 4.5\Theta_\tau) + \frac{12C_2S^2}{1 - \gamma} \sum_{\tau=1}^{t} \lambda_t^\tau \alpha_{t-1} + \frac{240S}{1 - \gamma} \lambda_t^1 + 22S\lambda_t^1 + \frac{87S\eta \varepsilon}{(1 - \gamma)^2} \sum_{\tau=1}^{t} \lambda_t^\tau.$$

Define

$$Y_t \triangleq \sum_{\tau=1}^{t} \lambda_t^\tau(L_\tau + 4.5\Theta_\tau).$$

Then we can further write that for $t \geq 1$,

$$Y_{t+1} \leq \frac{1}{1 + 0.1\eta^2C^2} Y_t + \frac{274C_2S^2t_0}{1 - \gamma} \lambda_t^\min\{t_0, t\} + \frac{87S\eta \varepsilon}{(1 - \gamma)^2} \sum_{\tau=1}^{t} \lambda_t^\tau. \quad \text{(upper bounding $\alpha_{t-1}$ by 1)}$$

Applying Lemma 39 with $c = \frac{0.1\eta^2C^2}{1 + 0.1\eta^2C^2}$, $g_t = Y_{t+1}$, $h_t = \frac{274C_2S^2t_0}{1 - \gamma} \lambda_t^\min\{t_0, t\} + \frac{87S\eta \varepsilon}{(1 - \gamma)^2} \sum_{\tau=1}^{t} \lambda_t^\tau$, we get

$$Y_t \leq Y_1(1 + 0.1\eta^2C^2)^{-t} + \frac{2}{0.1\eta^2C^2} \left( \frac{274C_2S^2t_0}{1 - \gamma} + \frac{87S\eta \varepsilon}{(1 - \gamma)^2} \sup_{t' \in \left[\frac{1}{2}, t\right]} \sum_{\tau=1}^{t'} \lambda_{t'}^\tau \right) \left(1 + 0.1\eta^2C^2\right)^{-\frac{t}{2}}$$

$$+ \frac{2}{0.1\eta^2C^2} \left( \frac{274C_2S^2t_0}{1 - \gamma} \sup_{t' \in \left[\frac{1}{2}, t\right]} \lambda_{t'}^\min\{t_0, t'\} + \frac{87S\eta \varepsilon}{(1 - \gamma)^2} \sup_{t' \in \left[\frac{1}{2}, t\right]} \sum_{\tau=1}^{t'} \lambda_{t'}^t \right). \quad (\lambda_t^\tau \leq 1)$$

With the choice of $\alpha_t = \frac{H+1}{H+t}$ where $H = \frac{2}{1 - \gamma}$, we have

$$t_0 = \Theta\left(\frac{6C_2S}{(1 - \gamma)\eta^2C^2}(H + 1)\right) = \Theta\left(\frac{S}{(1 - \gamma)^2\eta^2C^2}\right)$$

$$\sup_{t' \in \left[\frac{1}{2}, t\right]} \sum_{\tau=1}^{t'} \lambda_{t'}^\tau \leq \sup_{t' \in \left[\frac{1}{2}, t\right]} t'\alpha_{t'} \leq \frac{t(H + 1)}{H + t} \leq H + 1 = O\left(\frac{1}{1 - \gamma}\right) \quad \text{(Lemma 41)}$$

$$\sup_{t' \in \left[\frac{1}{2}, t\right]} \lambda_{t'}^\min\{t_0, t'\} = \begin{cases} 1 & \text{if } \frac{t}{2} \leq t_0 \\ \alpha_{\frac{t}{2}} & \text{else} \end{cases} \quad \text{(Lemma 41)}$$
Combining them and noticing that \((1 + 0.1\eta^2C'^2)^{-\frac{1}{2}} = \mathcal{O}(\frac{1}{t})\) when \(t \geq \frac{20}{\eta^2C'^2}\), we get that for \(t \geq 2t_0 = \Theta\left(\frac{S}{(1-\gamma)^3\eta^4C'^4}\right)\),

\[
Y_t = \mathcal{O}\left(\frac{S^3}{\eta^4C'^4(1-\gamma)^3\alpha_t} + \frac{S\varepsilon}{\eta C'^2(1-\gamma)^3}\right) = \mathcal{O}\left(\frac{S^3}{\eta^4C'^4(1-\gamma)^4t} + \frac{S\varepsilon}{\eta C'^2(1-\gamma)^3}\right).
\]

Since \(Y_t \leq 22S + 22S(t - 1)\alpha_t \leq \mathcal{O}\left(\frac{S\log t}{1-\gamma}\right)\), the above bound also trivially holds for \(t \leq 2t_0\). Then noticing that \(L_t \leq Y_t\) finishes the proof.

**Lemma 32** For any policy pair \(x, y\), the duality gap on the game can be related to the duality gap on individual states as follows:

\[
\max_{s,x',y'} \left( V_{x,y}^s - V_{x',y'}^s \right) \leq \frac{2}{1 - \gamma} \max_{s,x',y'} \left( x^s Q^s_y y'^s - x'^s Q^s_x y^s \right).
\]

**Proof** Notice that for any policy \(x\) and state \(s\),

\[
\max_{y'} \left( V_{x,y'}^s - V_*^s \right) = \sum_{a,b} x^s(a)y^s(b)Q_{x,y'}^s(a, b) - \sum_{a,b} x_*^s(a)y_*^s(b)Q_*^s(a, b)
\]

\[
= \sum_{a,b} x^s(a)y^s(b) \left( Q_{x,y'}^s(a, b) - Q_*^s(a, b) \right) + \sum_{a,b} \left( x^s(a)y^s(b) - x_*^s(a)y_*^s(b) \right) Q_*^s(a, b)
\]

\[
= \gamma \sum_{a,b} x^s(a)y^s(b)p(s'|s, a, b) \left( V_{x,y'}^s - V_*^s \right) + x^s Q^s_y y^s - x_*^s Q_*^s y_*^s
\]

\[
\leq \gamma \max_{s',y'} \left( V_{s',y'}^s - V_*^s \right) + x^s Q^s_y y^s - x_*^s Q_*^s y_*^s.
\]

Taking max over \(s\) on two sides and rearranging, we get

\[
\max_{s,y'} \left( V_{x,y'}^s - V_*^s \right) \leq \frac{1}{1 - \gamma} \max_{s,y'} \left( x^s Q^s_y y^s - x_*^s Q^s_y y_*^s \right) \leq \frac{1}{1 - \gamma} \max_{s,x',y'} \left( x^s Q^s_y y'^s - x'^s Q^s_y y^s \right). \]

Similarly,

\[
\max_{s,x'} \left( V_*^s - V_{x,y}^s \right) \leq \frac{1}{1 - \gamma} \max_{s,x',y'} \left( x^s Q^s_y y'^s - x'^s Q^s_y y^s \right).
\]

Combining the two inequalities, we get

\[
\max_{s,x',y'} \left( V_{x,y'}^s - V_{x',y}^s \right) \leq \frac{2}{1 - \gamma} \max_{s,x',y'} \left( x^s Q^s_y y'^s - x'^s Q^s_y y^s \right).
\]
Appendix G. Auxiliary Lemmas

G.1. Interactions between the Auxiliary Coefficients

Lemma 33  Let \( \{g_t\}_{t=1,2,...}, \{h_t\}_{t=1,2,...} \) be non-negative sequences that satisfy \( g_t \leq \gamma \sum_{\tau=1}^{t-1} \alpha^\tau_{t-1} g_{\tau} + h_t \) for all \( t \geq 1 \). Then \( g_t \leq \sum_{\tau=1}^{t} \beta^\tau_t h_{\tau} \).

Proof  We prove it by induction. When \( t = 1 \), the condition guarantees \( g_1 \leq h_1 = \beta^1_1 h_1 \). Suppose that it holds for \( 1, \ldots, t-1 \). Then

\[
g_t \leq \gamma \sum_{\tau=1}^{t-1} \alpha^\tau_{t-1} g_{\tau} + h_t
\]

\[
\leq \gamma \sum_{\tau=1}^{t-1} \alpha^\tau_{t-1} \left( \sum_{i=1}^{\tau} \beta^i_\tau h_i \right) + h_t
\]

\[
= \sum_{i=1}^{t} \left( \sum_{\tau=i}^{t-1} \gamma \alpha^\tau_{t-1} \beta^i_\tau \right) h_i + h_t
\]

It remains to prove that \( \sum_{\tau=i}^{t-1} \gamma \alpha^\tau_{t-1} \beta^i_\tau \leq \beta^i_t \) for all \( i \leq t-1 \). We use another induction to show this. Fix \( i \) and \( t \), and define the partial sum \( \zeta_r = \sum_{\tau=i}^{r} \gamma \alpha^\tau_{t-1} \beta^i_\tau \) for \( r \in [i, t-1] \). Below we show that

\[
\zeta_r \leq \alpha_i \prod_{\tau=i}^{r} (1 - \alpha_\tau + \alpha_\tau \gamma) \prod_{\tau=r+1}^{t-1} (1 - \alpha_\tau) \tag{34}
\]

Notice that the right-hand side above is \( \beta^i_t \) when \( r = t-1 \), which is exactly what we want to prove.

When \( r = i \), \( \zeta_i = \gamma \alpha^i_{t-1} = \gamma \alpha_i \prod_{\tau=i+1}^{t-1} (1 - \alpha_\tau) \leq \alpha_i (1 - \alpha_i + \alpha_i \gamma) \prod_{\tau=i+1}^{t-1} (1 - \alpha_\tau) \) where the inequality is because \( 1 - \alpha_i + \alpha_i \gamma - \gamma = (1 - \alpha_i)(1 - \gamma) \geq 0 \). Now assume that Eq. (34) holds up to \( r \) for some \( r \geq i \). Then

\[
\zeta_{r+1} = \zeta_r + \beta^i_{r+1} \gamma \alpha^r_{t-1}
\]

\[
\leq \alpha_i \prod_{\tau=i}^{r} (1 - \alpha_\tau + \alpha_\tau \gamma) \prod_{\tau=r+1}^{t-1} (1 - \alpha_\tau) + \left( \alpha_i \prod_{\tau=i}^{r} (1 - \alpha_\tau + \alpha_\tau \gamma) \right) \gamma \alpha_{r+1} \prod_{\tau=r+2}^{t-1} (1 - \alpha_\tau)
\]

\[
= \alpha_i \prod_{\tau=i}^{r+1} (1 - \alpha_\tau + \alpha_\tau \gamma) \prod_{\tau=r+2}^{t-1} (1 - \alpha_\tau).
\]

This finishes the induction. \( \blacksquare \)

Lemma 34  Let \( \{h_t\}_{t=1,2,...} \) and \( \{k_i\}_{t=1,2,...} \) be non-negative sequences that satisfy \( h_t = \sum_{\tau=1}^{t} \alpha^\tau_t k_{\tau} \). Then \( \sum_{\tau=2}^{t} \beta^\tau_t h_{\tau-1} \leq \frac{1}{\tau^2} \sum_{\tau=1}^{t-1} \beta^\tau_t k_{\tau} \).

Proof  By the assumption on \( h_{\tau} \), we have

\[
\sum_{\tau=2}^{t} \beta^\tau_t h_{\tau-1} \leq \sum_{\tau=2}^{t} \beta^\tau_t \left( \sum_{i=1}^{\tau-1} \alpha^i_{\tau-1} k_i \right) = \sum_{i=1}^{t} \left( \sum_{\tau=i+1}^{t} \beta^\tau_t \alpha^i_{\tau-1} \right) k_i
\]

33
It remains to prove that for \( i < t \), \( \sum_{\tau=i+1}^{t} \beta^\tau_i \alpha^i_{\tau-1} \leq \frac{1}{\gamma} \beta^i_i \), or equivalently, \( \gamma \sum_{\tau=i+1}^{t} \beta^\tau_i \alpha^i_{\tau-1} \leq \frac{1}{\gamma} \beta^i_i \). Below we use another induction to prove this. Fix \( i \) and \( t \), and define the partial sum \( \zeta_r = \gamma \sum_{\tau=\tau}^{t} \beta^\tau_i \alpha^i_{\tau-1} \) for \( r \in [i+1, t] \). We will show that

\[
\zeta_r \leq \alpha_i \prod_{\tau=i+1}^{r-1} (1 - \alpha_\tau) \prod_{\tau=r}^{t-1} (1 - \alpha_\tau + \alpha_\tau \gamma).
\]  

(35)

For the base case \( r = t \), we have

\[
\zeta_r = \gamma \beta^t_i \alpha^i_{t-1} = \gamma \alpha_i \prod_{\tau=i+1}^{t-1} (1 - \alpha_\tau) \leq \alpha_i \prod_{\tau=i+1}^{t-1} (1 - \alpha_\tau).
\]

Suppose that Eq. (35) holds up to \( r \) for some \( r \leq t \). Then

\[
\zeta_{r-1} = \zeta_r + \alpha_i \beta^r_i \beta^{r-1}_i
\]

\[
\leq \alpha_i \prod_{\tau=i+1}^{r-1} (1 - \alpha_\tau) \prod_{\tau=r}^{t-1} (1 - \alpha_\tau + \alpha_\tau \gamma) + \alpha_i \left( \prod_{\tau=r}^{t-2} (1 - \alpha_\tau) \right) \gamma \alpha_{r-1} \prod_{\tau=r-1}^{t-1} (1 - \alpha_\tau + \alpha_\tau \gamma)
\]

\[
\leq \left( \alpha_i \prod_{\tau=i+1}^{r-2} (1 - \alpha_\tau) \prod_{\tau=r}^{t-1} (1 - \alpha_\tau + \alpha_\tau \gamma) \right) \times (1 - \alpha_{r-1} + \gamma \alpha_{r-1} (1 - \alpha_{r-1} + \alpha_{r-1} \gamma))
\]

\[
\leq \left( \alpha_i \prod_{\tau=i+1}^{r-2} (1 - \alpha_\tau) \prod_{\tau=r}^{t-1} (1 - \alpha_\tau + \alpha_\tau \gamma) \right) \times (1 - \alpha_{r-1} + \alpha_{r-1} \gamma)
\]

\[
= \alpha_i \prod_{\tau=i+1}^{r-2} (1 - \alpha_\tau) \prod_{\tau=r-1}^{t-1} (1 - \alpha_\tau + \alpha_\tau \gamma).
\]

This finishes the induction. Applying the result with \( r = i + 1 \), we get

\[
\zeta_{i+1} \leq \alpha_i \prod_{\tau=i+1}^{t-1} (1 - \alpha_\tau + \alpha_\tau \gamma) = \frac{\beta^i_i}{1 - \alpha_i + \alpha_i \gamma} \leq \frac{\beta^i_i}{\gamma}
\]

where the last inequality is by \( 1 - \alpha_i + \alpha_i \gamma - \gamma = (1 - \alpha_i)(1 - \gamma) \geq 0 \). This finishes the proof. \( \blacksquare \)

**Lemma 35** For \( 0 \leq h \leq t \), \( \sum_{\tau=0}^{h} \alpha^\tau_t = \delta^h_t \).

**Proof** We prove it by induction on \( h \). When \( h = 0 \), \( \sum_{\tau=0}^{h} \alpha^\tau_t = \alpha^0_t = \prod_{\tau=1}^{t} (1 - \alpha_\tau) = \delta^0_t \) since \( \alpha_0 = 1 \). Suppose that the formula holds for \( h \). Then \( \sum_{\tau=0}^{h+1} \alpha^\tau_t = \sum_{\tau=0}^{h} \alpha^\tau_t + \alpha^{h+1}_t = \prod_{\tau=h+1}^{t} (1 - \alpha_\tau) + \alpha_{h+1} \prod_{\tau=h+2}^{t} (1 - \alpha_\tau) = \prod_{\tau=h+2}^{t} (1 - \alpha_\tau) = \delta^{h+1}_t \), which finishes the induction. \( \blacksquare \)

**Lemma 36** For any positive integers \( i, t \) with \( i \leq t \), \( \sum_{\tau=i}^{t} \lambda^\tau_t \beta^i_t \leq \frac{3}{\gamma} \lambda^i_t \).
Proof} Notice that

\[ \sum_{\tau = i}^{t} \lambda_i^\tau \beta_\tau^i = \lambda_i^i + \sum_{\tau = i+1}^{t} \lambda_i^\tau \beta_\tau^i. \] (36)

Below we use induction to prove

\[ \sum_{\tau = r}^{t} \lambda_i^\tau \beta_\tau^i \leq \frac{2}{1 - \gamma} \alpha_i \prod_{\tau = i}^{r-1} (1 - \alpha_\tau(1 - \gamma)) \prod_{\tau = r}^{t-1} \lambda_\tau + \beta_i^{r-1} \lambda_i \]

for \( r \in [i+1, t] \). When \( r = t \), \( \sum_{\tau = r}^{t} \lambda_i^\tau \beta_\tau^i = \lambda_i^t \beta_i^t = \alpha_i \prod_{\tau = i}^{t-1} (1 - \alpha_\tau(1 - \gamma)) \). Suppose this holds for some \( r \leq t \). Then

\[
\begin{align*}
\sum_{\tau = r-1}^{t} \lambda_i^\tau \beta_\tau^i &\leq \frac{2}{1 - \gamma} \alpha_i \prod_{\tau = i}^{r-2} (1 - \alpha_\tau(1 - \gamma)) \prod_{\tau = r-1}^{t-1} \lambda_\tau + \left( \alpha_i \prod_{\tau = i}^{r-2} (1 - \alpha_\tau(1 - \gamma)) \right) \alpha_{r-1} \prod_{\tau = r-1}^{t-1} \lambda_\tau \\
&\leq \left[ \alpha_i \prod_{\tau = i}^{r-2} (1 - \alpha_\tau(1 - \gamma)) \prod_{\tau = r}^{t-1} \lambda_\tau \right] \left( \frac{2}{1 - \gamma} (1 - \alpha_{r-1}(1 - \gamma)) + \alpha_{r-1} \right) \\
&\leq \left[ \alpha_i \prod_{\tau = i}^{r-2} (1 - \alpha_\tau(1 - \gamma)) \prod_{\tau = r}^{t-1} \lambda_\tau \right] \left( \frac{2}{1 - \gamma} \left( 1 - \frac{1}{2} \alpha_{r-1}(1 - \gamma) \right) \right) \\
&\leq \frac{2}{1 - \gamma} \alpha_i \prod_{\tau = i}^{r-2} (1 - \alpha_\tau(1 - \gamma)) \prod_{\tau = r-1}^{t-1} \lambda_\tau, \\
\end{align*}
\]

(\( \lambda_{r-1} \leq 1 \))

which finishes the induction. Notice that this implies

\[ \sum_{\tau = i+1}^{t} \lambda_i^\tau \beta_\tau^i \leq \frac{2}{1 - \gamma} \alpha_i (1 - \alpha_i(1 - \gamma)) \prod_{\tau = i+1}^{t-1} \lambda_\tau \leq \frac{2}{1 - \gamma} \alpha_i \prod_{\tau = i}^{t-1} \lambda_\tau = \frac{2}{1 - \gamma} \lambda_i 
\]

where the second inequality is by the definition of \( \lambda_i \). Thus,

\[ \sum_{\tau = 1}^{t} \lambda_i^\tau \beta_\tau^i \leq \frac{3}{1 - \gamma} \lambda_i^i. \] (37)

Lemma 37 \( \lambda_i^{\tau+1} \leq \lambda_i^\tau \).
Proof When \( \tau < t \), we have

\[
\frac{\lambda_{\tau+1}^t}{\lambda_t^\tau} = \frac{\alpha_{\tau+1} \prod_{i=\tau}^{\tau+1} \lambda_i}{\alpha_{\tau} \prod_{i=\tau}^{\tau+1} \lambda_i} \leq \frac{\alpha_{\tau+1}}{\alpha_{\tau} \lambda_\tau} \leq 1
\]

where in the first inequality we use \( \lambda_\tau \leq 1 \) and in the second inequality we use the definition of \( \lambda_\tau \).

When \( \tau = t \), we have \( \frac{\lambda_{t+1}^t}{\lambda_t^t} = \frac{1}{1} = 1 \). \( \blacksquare \)

Lemma 38 \( \sum_{\tau=1}^{t} \beta_{r}^{\tau} \leq \frac{2}{1 - \gamma} \).

Proof Below we use induction to prove that for all \( r = 1, 2, \ldots, t - 1 \),

\[
\sum_{\tau=1}^{r} \beta_{r}^{\tau} \leq \frac{1}{1 - \gamma} \prod_{i=r}^{t-1} (1 - \alpha_i + \alpha_i \gamma).
\]

When \( r = 1 \), the left-hand side is \( \beta_1^1 = \alpha_1 \prod_{i=1}^{t-1} (1 - \alpha_i + \alpha_i \gamma) \leq \frac{1}{1 - \gamma} \prod_{i=1}^{t-1} (1 - \alpha_i + \alpha_i \gamma) \), which is the right-hand side.

Suppose that this holds for \( r \), then

\[
\sum_{\tau=1}^{r+1} \beta_{r+1}^{\tau} = \beta_{r+1}^{\tau+1} + \sum_{\tau=1}^{r} \beta_{r}^{\tau}
\]

\[
\leq \alpha_{r+1} \prod_{i=r+1}^{t-1} (1 - \alpha_i + \alpha_i \gamma) + \frac{1}{1 - \gamma} \prod_{i=r}^{t-1} (1 - \alpha_i + \alpha_i \gamma)
\]

\[
\leq \left( \frac{1}{1 - \gamma} \prod_{i=r+1}^{t-1} (1 - \alpha_i + \alpha_i \gamma) \right) (\alpha_{r+1}(1 - \gamma) + 1 - \alpha_r(1 - \gamma))
\]

\[
\leq \frac{1}{1 - \gamma} \prod_{i=r+1}^{t-1} (1 - \alpha_i + \alpha_i \gamma)
\]

(because \( \alpha_{r+1} \leq \alpha_r \))

which finishes the induction.

Therefore, \( \sum_{\tau=1}^{t} \beta_{r}^{\tau} = 1 + \sum_{\tau=1}^{t-1} \beta_{r}^{\tau} \leq 1 + \frac{1}{1 - \gamma} \leq \frac{2}{1 - \gamma} \). \( \blacksquare \)

Lemma 39 Let \( \{g_t\}_{t=0,1,2,\ldots} \), \( \{h_t\}_{t=1,2,\ldots} \) be non-negative sequences that satisfy \( g_t \leq (1-c)g_{t-1} + h_t \) for some \( c \in (0, 1) \) for all \( t \geq 1 \). Then

\[
g_t \leq g_0 (1-c)^t + \frac{\max_{\tau \in [1,t]} h_{\tau}}{c} (1-c)^{\frac{t}{2}} + \frac{\max_{\tau \in [t/2,t]} h_{\tau}}{c}.
\]

Proof We first show that

\[
g_t \leq g_0 (1-c)^t + \sum_{\tau=1}^{t} (1-c)^{t-\tau} h_{\tau}.
\]
The case of \( t = 1 \) is clear. Suppose that this holds for \( g_t \). Then

\[
g_{t+1} \leq (1 - c) \left( g_0(1 - c)^t + \sum_{\tau=1}^{t} (1 - c)^{t-\tau} h_\tau \right) + h_{t+1} = g_0(1 - c)^{t+1} + \sum_{\tau=1}^{t+1} (1 - c)^{t+1-\tau} h_\tau,
\]

which finishes the induction. Therefore,

\[
g_t \leq g_0(1 - c)^t + \sum_{\tau=1}^{t/2} (1 - c)^{t-\tau} \max_{\tau \in [1,t/2]} h_\tau + \sum_{\tau=t/2+1}^{t} (1 - c)^{t-\tau} \max_{\tau \in [t/2,t]} h_\tau
\]

\[
\leq g_0(1 - c)^t + \frac{\max_{\tau \in [1,t/2]} h_\tau}{c} (1 - c)^{\frac{t}{2}} + \frac{\max_{\tau \in [t/2,t]} h_\tau}{c}.
\]

G.2. Some Properties for the choice of \( \alpha_t = \frac{H+1}{H+t} \)

**Lemma 40** For the choice \( \alpha_t = \frac{H+1}{H+t} \) with \( H \geq \frac{2}{1-\gamma} \), we have \( \sum_{t=\tau}^{\infty} \beta_t^\tau \leq H + 3 \).

**Proof** When \( t \geq \tau + 2 \),

\[
\beta_t^\tau = \alpha_\tau \prod_{i=\tau}^{t-1} (1 - \alpha_i (1 - \gamma))
\]

\[
\leq \alpha_\tau \prod_{i=\tau}^{t-1} \left( 1 - \alpha_i \times \frac{2}{H + 1} \right)
\]

\[
= \alpha_\tau \prod_{i=\tau}^{t-1} \left( 1 - \frac{2}{H + i} \right)
\]

\[
= \frac{H + 1}{H + \tau} \times \frac{H + \tau - 2}{H + \tau} \times \frac{H + \tau - 1}{H + \tau + 1} \times \cdots \times \frac{H + t - 3}{H + t - 1}
\]

\[
= \frac{H + 1}{H + \tau} \times \frac{(H + \tau - 2)(H + \tau - 1)}{(H + t - 2)(H + t - 1)}
\]

\[
= \frac{H + 1}{H + \tau} (H + \tau - 2)(H + \tau - 1) \left( \frac{1}{H + t - 2} - \frac{1}{H + t - 1} \right)
\]

\[
\leq (H + 1)(H + \tau - 2) \left( \frac{1}{H + t - 2} - \frac{1}{H + t - 1} \right).
\]

Therefore,

\[
\sum_{t=\tau+2}^{\infty} \beta_t^\tau \leq (H + 1)(H + \tau - 2) \times \frac{1}{H + \tau} \leq H + 1,
\]

and thus \( \sum_{t=\tau}^{\infty} \beta_t^\tau \leq H + 3 \).
Lemma 41 For the choice $\alpha_t = \frac{H + 1}{H + t}$ with $H \geq \frac{2}{1 - \gamma}$, we have $\lambda_t^\tau = \alpha_t$ for $\tau < t$.

Proof With this choice of $\alpha_t$,

$$
\lambda_t = \max \left\{ \frac{H + t}{H + t + 1}, 1 - \frac{1 - \gamma}{2} \times \frac{H + 1}{H + t} \right\}
= \max \left\{ 1 - \frac{1}{H + t + 1}, 1 - \frac{1 - \gamma}{2} \times \frac{H + 1}{H + t} \right\}.
$$

By the condition, we have $\frac{1 - \gamma}{2} \times \frac{H + 1}{H + t} \geq \frac{1}{H + t + 1}$. Therefore, $\lambda_t = \frac{H + t}{H + t + 1} = \frac{\alpha_t + 1}{\alpha_t}$. Thus for $\tau < t$,

$$
\lambda_t^\tau = \alpha_t \prod_{i=\tau}^{t-1} \frac{\alpha_{i+1}}{\alpha_i} = \alpha_t.
$$

\[\blacksquare\]

Appendix H. Analysis on Sample Complexity

H.1. Proof of Theorem 3

Proof As long as we can make

$$
|\ell_t^a(a) - e_a^\top Q_t^s \tilde{y}_t^s| \leq \frac{\varepsilon}{|A|}
$$

hold with high probability, then $|\ell_t^a(a) - e_a^\top Q_t^s \tilde{y}_t^s| \leq |\ell_t^a(a) - e_a^\top Q_t^s \tilde{y}_t^s| + |e_a^\top Q_t^s \tilde{y}_t^s - e_a^\top Q_t^s \tilde{y}_t^s| \leq \frac{\varepsilon}{2|A|} + \frac{1 - \gamma}{2} \times \frac{1}{|A|}$, which implies $|\ell_t^a - Q_t^s \tilde{y}_t^s| \leq \varepsilon$. We can similarly ensure $|x_t^a - Q_t^s \tilde{y}_t^s| \leq \varepsilon$ and $|\rho_t^s - x_t^a Q_t^s \tilde{y}_t^s| \leq \varepsilon$ by the same way. Let $N_{s,a} = \sum_{i=1}^L \mathbb{1}[s_i = s, a_i = a]$ be the number of times the $(s,a)$ pair is visited. For a deterministic $N$, we can use Azuma-Hoeffding’s inequality and know that Eq. (38) holds with probability $1 - \delta$ if $N = \Omega(\frac{|A|^2 \log(1/\delta)}{\varepsilon^2})$. However, $N_{s,a}$ is random, so we cannot use Azuma-Hoeffding’s inequality directly. Let $(b^{(1)}, s^{(1)}), (b^{(2)}, s^{(2)}), \ldots$ be a sequence of independent random variables where $b^{(i)} \sim \tilde{y}_t^s, s^{(i)} \sim p(\cdot | s, a)$, $i = 1, 2, \ldots$ and define $\tilde{e}_{t,m}^a = \frac{1}{m} \sum_{i=1}^m (\sigma(s, a, b^{(i)}) + \gamma V_{t-1}^{(i)})$. It is direct to see that $\tilde{e}_{t,m}^a$ is an unbiased estimator of $e_a^\top Q_t^s \tilde{y}_t^s$. Then by Azuma-Hoeffding’s inequality, we have

$$
\text{Prob} \left[ |\ell_t^a(a) - e_a^\top Q_t^s \tilde{y}_t^s| \leq \mathcal{O} \left( \sqrt{\frac{\log(L/\delta)}{N_{s,a}}} \right) \right] \\
\leq \text{Prob} \left[ \exists m \in [L], \tilde{e}_{t,m}^a - e_a^\top Q_t^s \tilde{y}_t^s \leq \mathcal{O} \left( \sqrt{\frac{\log(L/\delta)}{m}} \right) \right] \\
\leq \sum_{m=1}^L \text{Prob} \left[ |\tilde{e}_{t,m}^a| \leq \mathcal{O} \left( \sqrt{\frac{\log(L/\delta)}{m}} \right) \right] \leq \delta.
$$
Therefore, with probability at least $1 - \delta$, Eq. (38) holds if

$$N_{s,a} = \Omega \left( \frac{|A|^2}{\varepsilon^2} \log(L/\delta) \right).$$

Now it remains to determine $L$ to make Eq. (39) hold with high probability. Note that by Assumption 1, we know that $T^{s,s} \leq \frac{1}{\mu}$ for any $s'$. Let $T_{s,a}$ be the distribution of random variable which is the number of rounds between the current state-action pair $(s', a')$ and the next occurrence of $(s, a)$ under strategy $\tilde{x}^a_t$ and $\tilde{y}^a_t$. The mean of this distribution is $t_{s,a} = 1 + \frac{2|A|}{\varepsilon \mu}$, within $1 - \delta$. Solving $L$ gives $L = \Omega \left( \frac{|A|^3 (1 - \gamma) \mu^2}{\varepsilon^2} \log^2(L/\delta) \right)$ with probability $1 - \delta$. Solving $L$ gives $L = \Omega \left( \frac{|A|^3 (1 - \gamma) \mu^2}{\varepsilon^2} \log^2(T/\delta) \right)$. The cases for $r^s(b)$ and $\tilde{p}^s$ are similar. Finally, using a union bound on all $A, B, S$, and all iterations, we know that with probability $1 - \delta$, the $\varepsilon$-approximations are always guaranteed if we use the estimation above and take $L = \Omega \left( \frac{|A|^3 + |B|^3}{(1 - \gamma) \mu^2} \log^2(T/\delta) \right)$.

### H.2. Proof of Corollary 4

**Proof** From Theorem 1, we know that in order to show $\frac{1}{T} \sum_{t=1}^{T} \max_{x', y'} \left( V^s_{x, y} - V^s_{x', y'} \right) \leq \xi$, it is sufficient to show $\frac{|S|}{\eta(1 - \gamma)^2} \sqrt{\frac{\log T}{T}} \leq \xi$ and $\frac{|S|}{\sqrt{\eta(1 - \gamma)^2}} \leq \xi$. Solving these two inequalities, we get $T = \Omega \left( \frac{|S|^2}{\eta(1 - \gamma)^2 \varepsilon^2 \log |S|} \right)$ and $\varepsilon = \mathcal{O} \left( \frac{1}{|S|^2} \right)$. Plugging $\varepsilon$ into $L$ in Theorem 3 gives the required $L$.

### H.3. Proof of Corollary 5

**Proof** From Theorem 2, we know that in order to show $\frac{1}{|S|} \sum_{s \in S} \text{dist}^2_{s} (\mathbb{E}^s_{T}) \leq \xi$, it is sufficient to show $\frac{|S|^2}{\eta^2 C^2 (1-\gamma)^2 T} \leq \xi$ and $\frac{|S|^2}{\eta^2 C^2 (1-\gamma)^2} \leq \xi$. Solving these two inequalities, we get $T = \Omega \left( \frac{|S|^2}{\eta^2 C^2 (1-\gamma)^2 \varepsilon} \right)$ and $\varepsilon = \mathcal{O} \left( \eta C^2 (1 - \gamma)^3 \xi \right)$. Plugging $\varepsilon$ into $L$ in Theorem 3 gives the required $L$.

### Appendix I. Analysis on Rationality

In this section, we analyze the rationality of our algorithm. First, we present the full pseudocode of Algorithm 2, which is the single-player-perspective version of Algorithm 1, and then prove that Algorithm 2 achieves rationality.
**Algorithm 2** Single-Player-Perspective Optimistic Gradient Descent/Ascent for Markov Games

**Parameter:** $\gamma \in [\frac{1}{2}, 1], \eta \leq \frac{1}{10^4} \sqrt{\frac{(1-\gamma)^3}{S}}, \varepsilon \in [0, \frac{1}{1-\gamma}]$

**Parameters:** a non-increasing sequence $\{\alpha_t\}_{t=1}^{T}$ that goes to zero.

**Initialization:** arbitrarily initialize $\hat{x}_1^s = x_1^s \in \Delta_A$ for all $s \in S$.

$V_t^s \leftarrow 0$ for all $s \in S$.

For $t = 1, \ldots, T$ do

For all $s$, define $Q_t^s \in \mathbb{R}^{|A| \times |B|}$ as

$$Q_t^s(a, b) \triangleq \sigma(s, a, b) + \gamma \mathbb{E}_{s' \sim p(\cdot | s, a, b)} \left[ V_{t-1}^{s'} \right],$$

and update

$$\hat{x}_{t+1} = \Pi_{\Delta_A} \left\{ \hat{x}_t - \eta \ell_t^s \right\},$$

$$x_{t+1}^s = \Pi_{\Delta_A} \left\{ \hat{x}_{t+1} - \eta \rho_t^s \right\},$$

$$V_t^s = (1 - \alpha_t) V_{t-1}^s + \alpha_t \rho_t^s,$$

where $\ell_t^s$ and $\rho_t^s$ are $\varepsilon$-approximations of $Q_t^s y^s$ and $x_t^s \top Q_t^s y^s$, respectively.

**I.1. Single-Player-Perspective Version of Algorithm 1**

**I.2. Analysis of Algorithm 2**

In this section, we prove Theorem 6, which shows the rationality of Algorithm 2. We call the original game Game 1 and construct a two-player Markov game Game 2 with the difference being that Player 2 has only one single action (call it $1$) on each state, the loss function is redefined as $\sigma(s, a, 1) = \mathbb{E}_{b \sim y^s}[\sigma(s, a, b)]$, and the transition kernel is redefined as $p(s' | s, a, 1) = \mathbb{E}_{b \sim y^s}[p(s' | s, a, b)]$. Correspondingly, we define

$$Q_t^s(a, 1) = \sigma(s, a, 1) + \gamma \mathbb{E}_{s' \sim p(\cdot | s, a, 1)} \left[ V_{t-1}^{s'} \right],$$

$$\hat{x}_{t+1} = \Pi_{\Delta_A} \left\{ \hat{x}_t - \eta \ell_t^s \right\},$$

$$x_{t+1}^s = \Pi_{\Delta_A} \left\{ \hat{x}_{t+1} - \eta \rho_t^s \right\},$$

$$V_t^s = (1 - \alpha_t) V_{t-1}^s + \alpha_t \rho_t^s,$$

where $V_t^s = 0$ for all $s$, and $\ell_t$ and $\rho_t^s$ are the same as in Algorithm 2. Clearly, the sequences $\{\hat{x}_t, x_t\}_{t \in [T]}$ and $\{\hat{x}_t, x_t\}_{t \in [T]}$ are exactly the same (assuming their initializations are the same, that is, $\hat{x}_1 = \hat{x}_1$ and $x_1 = x_1$). In the following lemma, we show that $\ell_t$ and $\rho_t^s$ are indeed $\varepsilon$-approximation of $Q_t^s(\cdot, 1)$ and $x_t^s \top Q_t^s(\cdot, \cdot)$, which then implies that the sequence $\{\hat{x}_t\}_{t \in [T]}$ is indeed the output of our Algorithm 1 for Game 2 (note that we can think of Player 2 executing Algorithm 1 in Game 2 as well since she only has one unique strategy). Realizing that $X^s$ for Game 2 is exactly the set of best responses of $y^s$, we can thus conclude that Theorem 6 is a direct corollary of Theorem 1 and Theorem 2.

**Lemma 42** For all $t$ and $s$, $\ell_t$ and $\rho_t^s$ are $\varepsilon$-approximation of $Q_t^s(\cdot, 1)$ and $x_t^s \top Q_t^s(\cdot, \cdot)$ respectively.
Proof We prove the result together with $V_s^t = V_s^t$, $Q_s^t(\cdot, 1) = Q_s^t y^s$ for all $t \in [T]$ by induction. When $t = 1$, $V_s^t = V_s^1$ clearly holds. In addition, $Q_1^t(a, \cdot) y^s = \mathbb{E}_{b \sim y^s}[\sigma(s, a, b)] = Q_s^1(a, 1)$. Therefore $\ell_s^t$ and $\rho_s^t$ are indeed $\varepsilon$-approximation of $Q_s^1(a, \cdot)$ and $x_s^T Q_s^t(a, \cdot)$.

Suppose that the claim holds at $t$. By definition and the inductive assumption, we have

$$Q_{t+1}^s(a, 1) = \sigma(s, a, 1) + \gamma \mathbb{E}_{s' \sim y^s}[V^s_{s'} - V^s_t]$$

$$= \mathbb{E}_{b \sim y^s}[\sigma(s, a, b) + \gamma \mathbb{E}_{s' \sim p(s'|s, a, b)}[V^s_{s'}]]$$

$$= Q_{t+1}^s(a, \cdot) y^s,$$

which also shows that $\ell_{t+1}^s$ and $\rho_{t+1}^s$ are indeed $\varepsilon$-approximation of $Q_{t+1}^s(\cdot, 1)$ and $x_{t+1}^s Q_s^t(a, \cdot)$ (recall $x_{t+1}^s = x_{t+1}^s$). By definition of $V_{t+1}^s$, we also have $V_{t+1}^s = (1 - \alpha_{t+1}) V_t^s + \alpha_t \rho_t^s \sigma(s, a, 1) + \gamma \mathbb{E}_{s' \sim y^s}[V^s_{s'}]$. By definition of $V_{t+1}^s$, we also have $V_{t+1}^s = (1 - \alpha_{t+1}) V_t^s + \alpha_t \rho_t^s = (1 - \alpha_{t+1}) V_t^s + \alpha_t \rho_t^s = V_{t+1}^s$, which finishes the induction.