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

Chen-Yu Wei CHENYU.WEI@USC.EDU
Chung-Wei Lee[*] LEECHUNG@USC.EDU
Mengxiao Zhang[*] MENGXIAO.ZHANG@USC.EDU
Haipeng Luo HAIPENGL@USC.EDU
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 with near-optimal sample complexity 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

[*] Equal contribution.

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.

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 know, 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,[2] 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) σ 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) $p : \mathcal{S}\times\mathcal{A}\times\mathcal{B} \to \Delta_{\mathcal{S}}$ 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 $\mathcal{S} \to \Delta_{\mathcal{A}}$ 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

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

The *minimax* game value on state s is then defined as

$$V^s_\star = \min_x \max_y V^s_{x,y} = \max_y \min_x V^s_{x,y}.$$

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

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

[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 $\operatorname{dist}_\star(\cdot)$ depends on its input.

We also define the *Q-function* on state s under policy pair (x, y) as

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

which can be compactly written as a matrix $Q^s_{x,y} \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{B}|}$ such that $V^s_{x,y} = {x^s}^\top Q^s_{x,y} y^s$. We write $Q^s_\star = Q^s_{x_\star,y_\star}$ for any minimax/maximin policy pair $(x_\star, y_\star)$ (which is unique even if $(x_\star, y_\star)$ is not). Finally, $\|Q\|$ for a matrix Q is defined as $\max_{i,j} |Q_{i,j}|$.

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 $Q \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{B}|}$. Specifically, OGDA maintains two sequences of action distributions $\widehat{x}_1, \widehat{x}_2, \ldots \in \Delta_\mathcal{A}$ and $x_1, x_2, \ldots \in \Delta_\mathcal{A}$ for Player 1, and similarly two sequences $\widehat{y}_1, \widehat{y}_2, \ldots \in \Delta_\mathcal{B}$ and $y_1, y_2, \ldots \in \Delta_\mathcal{B}$ for Player 2, following the updates below:

$$\begin{aligned} \widehat{x}_{t+1} &= \Pi_{\Delta_\mathcal{A}}\big\{\widehat{x}_t - \eta Q y_t\big\}, & x_{t+1} &= \Pi_{\Delta_\mathcal{A}}\big\{\widehat{x}_{t+1} - \eta Q y_t\big\}, \\ \widehat{y}_{t+1} &= \Pi_{\Delta_\mathcal{B}}\big\{\widehat{y}_t + \eta Q^\top x_t\big\}, & y_{t+1} &= \Pi_{\Delta_\mathcal{B}}\big\{\widehat{y}_{t+1} + \eta Q^\top x_t\big\}, \end{aligned} \tag{1}$$

where η is some learning rate. As one can see, unlike the standard Gradient Descent Ascent algorithm which simply sets $(x_t, y_t) = (\widehat{x}_t, \widehat{y}_t)$, 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 $(\widehat{x}_t, \widehat{y}_t)$ (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 Multiplicative 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.

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 $Q^s_{x_t,y_t}$. However, an important difference is that now the game matrix is *changing* over time. Indeed, if the polices are changing rapidly for subsequent states, the game matrix $Q^s_{x_t,y_t}$ 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 *critic* slowly learn the value function for each state. Specifically, for each state s, the critic maintains a sequence of values $V^s_0 = 0, V^s_1, V^s_2, \ldots$. During iteration t, instead of using $Q^s_{x_t,y_t}$ as the game matrix for state s, we use Q^s_t defined via $Q^s_t(a, b) = \sigma(s, a, b) + \gamma \mathbb{E}_{s'\sim p(\cdot|s,a,b)}[V^{s'}_{t-1}]$. Ideally, OGDA would then take the role of an *actor* and compute x^s_{t+1} and $\widehat{x}^s_{t+1}$ using the gradient $Q^s_t y^s_t$ (and similarly y^s_{t+1} and $\widehat{y}^s_{t+1}$ using the gradient ${Q^s_t}^\top x^s_t$). Since such exact gradient information is often unknown, we only require the algorithm to come up with estimations ℓ^s_t and r^s_t such that $\|\ell^s_t - Q^s_t y^s_t\| \le \varepsilon$ and $\|r^s_t - {Q^s_t}^\top x^s_t\| \le \varepsilon$ for some prespecified error ε (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: $\gamma \in [\frac{1}{2}, 1), \eta \le \frac{1}{10^4}\sqrt{\frac{(1-\gamma)^5}{S}}, \varepsilon \in \left[0, \frac{1}{1-\gamma}\right]$.
Parameters: a non-increasing sequence $\{\alpha_t\}_{t=1}^T$ that goes to zero.
Initialization: $\forall s \in \mathcal{S}$, arbitrarily initialize $\widehat{x}_1^s = x_1^s \in \Delta_{\mathcal{A}}$ and $\widehat{y}_1^s = y_1^s \in \Delta_{\mathcal{B}}$, and set $V_0^s \leftarrow 0$.
for $t = 1, \ldots, T$ do
    For all $s$, define $Q_t^s \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{B}|}$ 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
        $\widehat{x}_{t+1}^s = \Pi_{\Delta_{\mathcal{A}}}\left\{\widehat{x}_t^s - \eta \ell_t^s\right\}$,   (2)
        $x_{t+1}^s = \Pi_{\Delta_{\mathcal{A}}}\left\{\widehat{x}_{t+1}^s - \eta \ell_t^s\right\}$,   (3)
        $\widehat{y}_{t+1}^s = \Pi_{\Delta_{\mathcal{B}}}\left\{\widehat{y}_t^s + \eta r_t^s\right\}$,   (4)
        $y_{t+1}^s = \Pi_{\Delta_{\mathcal{B}}}\left\{\widehat{y}_{t+1}^s + \eta r_t^s\right\}$,   (5)
        $V_t^s = (1-\alpha_t)V_{t-1}^s + \alpha_t \rho_t^s$,   (6)
    where $\ell_t^s, r_t^s$, and $\rho_t^s$ are $\varepsilon$-approximations of $Q_t^s y_t^s$, ${Q_t^s}^\top x_t^s$, and ${x_t^s}^\top Q_t^s y_t^s$ respectively, such that
    $\|\ell_t^s - Q_t^s y_t^s\| \le \varepsilon$, $\|r_t^s - {Q_t^s}^\top x_t^s\| \le \varepsilon$, and $|\rho_t^s - {x_t^s}^\top Q_t^s y_t^s| \le \varepsilon$.
end
```

similar to (Wei et al., 2021), we adopt a constant learning rate η (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 ρ_t^s is an estimation of ${x_t^s}^\top Q_t^s y_t^s$ such that $|\rho_t^s - {x_t^s}^\top Q_t^s y_t^s| \le \varepsilon$.[5] To stabilize the game matrix, we require the learning rate α_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 α_t, which we set to $\alpha_t = \frac{H+1}{H+t}$ 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* $\alpha_t = \frac{H+1}{H+t}$ *where* $H = \frac{2}{1-\gamma}$ *guarantees*

$$\frac{1}{T}\sum_{t=1}^{T} \max_{s,x',y'} \left(V_{\widehat{x}_t, y'}^s - V_{x', \widehat{y}_t}^s\right) = \mathcal{O}\left(\frac{|\mathcal{S}|}{\eta(1-\gamma)^2}\sqrt{\frac{\log T}{T}} + \frac{|\mathcal{S}|\sqrt{\varepsilon}}{\sqrt{\eta}(1-\gamma)^2}\right).$$

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

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 $\widehat{z}_T^s = (\widehat{x}_T^s, \widehat{y}_T^s)$,*

$$\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} \operatorname{dist}_\star^2(\widehat{z}_T^s) = \mathcal{O}\left(\frac{|\mathcal{S}|^2}{\eta^4 C^4 (1-\gamma)^4 T} + \frac{\varepsilon}{\eta C^2 (1-\gamma)^3}\right),$$

where $C > 0$ is a problem-dependent constant (that always exists) satisfying: for all state s and all policy pair $z = (x, y)$, $\max_{x',y'}\left(x^s Q_\star^s y'^s - x'^s Q_\star^s y^s\right) \geq C\operatorname{dist}_\star(z^s)$.

Theorem 1 shows that the average duality-gap for each state s goes to zero when both $1/T$ and ε 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 $\widehat{z}_T^s$ from the equilibrium set, which goes down at the rate of $1/T$ up to the estimation error ε. 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 $Q_t^s y_t^s$, ${Q_t^s}^\top x_t^s$, and ${x_t^s}^\top Q_t^s y_t^s$, making $\varepsilon = 0$. 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: $\widetilde{x}_t^s(a) = \left(1 - \frac{\varepsilon'}{2}\right) x_t^s(a) + \frac{\varepsilon'}{2|\mathcal{A}|}$ and $\widetilde{y}_t^s(b) = \left(1 - \frac{\varepsilon'}{2}\right) y_t^s(b) + \frac{\varepsilon'}{2|\mathcal{B}|}$, where $\varepsilon' = (1-\gamma)\varepsilon$. 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 $a_i \sim \widetilde{x}_t^{s_i}$, $b_i \sim \widetilde{y}_t^{s_i}$, and $s_{i+1} \sim p(\cdot|s_i, a_i, b_i)$. Then we construct the estimators as follows:

$$\ell_t^s(a) = \frac{\sum_{i=1}^L \mathbb{1}[s_i = s, a_i = a]\left(\sigma(s, a, b_i) + \gamma V_{t-1}^{s_{i+1}}\right)}{\sum_{i=1}^L \mathbb{1}[s_i = s, a_i = a]}, \tag{7}$$

$$r_t^s(b) = \frac{\sum_{i=1}^L \mathbb{1}[s_i = s, b_i = b]\left(\sigma(s, a_i, b) + \gamma V_{t-1}^{s_{i+1}}\right)}{\sum_{i=1}^L \mathbb{1}[s_i = s, b_i = b]}, \tag{8}$$

$$\rho_t^s = \frac{\sum_{i=1}^L \mathbb{1}[s_i = s]\left(\sigma(s, a_i, b_i) + \gamma V_{t-1}^{s_{i+1}}\right)}{\sum_{i=1}^L \mathbb{1}[s_i = s]}. \tag{9}$$

(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 $\frac{1}{\mu} = \max_{x,y} \max_{s,s'} T^{s \to s'}_{x,y}$, where $T^{s \to s'}_{x,y}$ 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 $L = \widetilde{\Omega}\left(\frac{|\mathcal{A}|^3+|\mathcal{B}|^3}{(1-\gamma)\mu\varepsilon^3} \log^2(T/\delta)\right)$.[6] Then the estimators Eq. (7), Eq. (8), and Eq. (9) ensure that with probability at least $1 - \delta$, $\|\ell^s_t - Q^s_t y^s_t\|$, $\|r^s_t - {Q^s_t}^\top x^s_t\|$, and $|\rho^s_t - {x^s_t}^\top Q^s_t y^s_t|$ are all of order $\mathcal{O}(\varepsilon)$ 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 ε). Equivalently, we show below how many iterations or total interactions are need to achieve a certain accuracy. (The choice of α_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 $L = \widetilde{\Omega}\left(\frac{(|\mathcal{A}|^3+|\mathcal{B}|^3)|\mathcal{S}|^6}{(1-\gamma)^{13}\mu\eta^3\xi^6} \log^2(T/\delta)\right)$ for $T = \widetilde{\Omega}\left(\frac{|\mathcal{S}|^2}{\eta^2(1-\gamma)^4\xi^2}\right)$ iterations ensures with probability at least $1 - \delta$, $\frac{1}{T}\sum_{t=1}^T \max_{s,x',y'}(V^s_{\widehat{x}_t,y'} - V^s_{x',\widehat{y}_t}) \le \xi$. Ignoring other dependence, this requires $\widetilde{\Omega}(1/\xi^8)$ interactions in total.*

Corollary 5 *If Assumption 1 holds, then running Algorithm 1 with estimators Eq. (7), Eq. (8), Eq. (9) and $L = \widetilde{\Omega}\left(\frac{|\mathcal{A}|^3+|\mathcal{B}|^3}{(1-\gamma)^{10}\mu\eta^3C^6\xi^3} \log^2(T/\delta)\right)$ for $T = \Omega\left(\frac{|\mathcal{S}|^2}{\eta^4C^4(1-\gamma)^4\xi}\right)$ iterations ensures with probability at least $1 - \delta$, $\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} \mathrm{dist}^2_\star(\widehat{z}^s_T) \le \xi$. Ignoring other dependence, this requires $\widetilde{\Omega}(1/\xi^4)$ 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.[7] 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_t 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 $\underline{\sigma}(s, a, 1) = \mathbb{E}_{b\sim y^s}[\sigma(s, a, b)]$, and the transition kernel is redefined as $\underline{p}(s'|s, a, 1) = \mathbb{E}_{b\sim y^s}[p(s'|s, a, b)]$. It is straightforward to see that following our algorithm, Player 1's behaviors in the original game and 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 $\widetilde{\Omega}$ 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*

$$\frac{1}{T}\sum_{t=1}^{T}\max_{s,x'}\left(V^s_{\widehat{x}_t,y^s} - V^s_{x',y^s}\right) = \mathcal{O}\left(\frac{|\mathcal{S}|}{\eta(1-\gamma)^2}\sqrt{\frac{\log T}{T}} + \frac{|\mathcal{S}|\sqrt{\varepsilon}}{\sqrt{\eta}(1-\gamma)^2}\right),$$

and for $\mathcal{X}_{BR} = \left\{x : V^s_{x,y} = \min_{x'} V^s_{x',y}, \forall s \in \mathcal{S}\right\}$ and some problem-dependent constant $C' > 0$,

$$\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\|\widehat{x}^s_T - \Pi_{\mathcal{X}_{BR}}\{\widehat{x}^s_T\}\|^2 = \mathcal{O}\left(\frac{|\mathcal{S}|^2}{\eta^4 C'^4(1-\gamma)^4 T} + \frac{\varepsilon}{\eta C'^2(1-\gamma)^3}\right).$$

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:

$$\frac{1}{T}\sum_{t=1}^{T}\Delta(\widehat{z}_t) = \mathcal{O}\left(\frac{1}{\eta\sqrt{T}}\right) \tag{10}$$

   where $\Delta(z) = \max_{x',y'}\left(x^\top Q y' - x'^\top Q y\right)$ is the duality gap of $z = (x, y)$.

2. Last-iterate convergence:

$$\text{dist}^2_\star(\widehat{z}_t) \le C_1 \text{dist}^2_\star(\widehat{z}_1)\left(1 + \eta^2 C^2\right)^{-t} \tag{11}$$

   where $\text{dist}_\star(z)$ 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.[8]

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 $\text{dist}^2_\star(\widehat{z}_{t+1})$ and $\text{dist}^2_\star(\widehat{z}_t)$:

$$\text{dist}^2_\star(\widehat{z}_{t+1}) \le \text{dist}^2_\star(\widehat{z}_t) + \underbrace{\eta^2\|z_t - z_{t-1}\|^2}_{\text{instability penalty}} - \underbrace{\left(\|\widehat{z}_{t+1} - z_t\|^2 + \|z_t - \widehat{z}_t\|^2\right)}_{\text{instability bonus}}. \tag{12}$$

[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 $\mathrm{dist}_\star^2(\widehat{z}_{t+1})$ larger if $\|z_t - z_{t-1}\|$ is large, while the instability bonus term makes $\mathrm{dist}_\star^2(\widehat{z}_{t+1})$ smaller if either $\|\widehat{z}_{t+1} - z_t\|$ or $\|z_t - \widehat{z}_t\|$ 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 $\widehat{z}_{t+1}$, that is, $\|\widehat{z}_{t+1} - z_t\|^2 + \|z_t - \widehat{z}_t\|^2 \gtrsim \eta^2\Delta^2(\widehat{z}_{t+1})$, and thus

$$\mathrm{dist}_\star^2(\widehat{z}_{t+1}) \le \mathrm{dist}_\star^2(\widehat{z}_t) + \underbrace{\eta^2 \|z_t - z_{t-1}\|^2}_{\text{instability penalty}} - \underbrace{\frac{1}{2}\left(\|\widehat{z}_{t+1} - z_t\|^2 + \|z_t - \widehat{z}_t\|^2\right)}_{\frac{1}{2}\text{ instability bonus}} - \Omega(\eta^2\Delta^2(\widehat{z}_{t+1})). \tag{13}$$

By taking $\eta \le \frac{1}{8}$, summing over t, canceling the penalty term with the bonus term, telescoping and rearranging, we get $\sum_{t=1}^T \Delta^2(\widehat{z}_t) \le \mathcal{O}(1/\eta^2)$. 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, $\Delta(z) \ge C\mathrm{dist}_\star(z)$. This, when combined with Eq. (13), shows

$$\mathrm{dist}_\star^2(\widehat{z}_{t+1}) \le \frac{\mathrm{dist}_\star^2(\widehat{z}_t)}{1+\Omega(\eta^2 C^2)} + \eta^2\|z_t - z_{t-1}\|^2 - \Omega\left(\|\widehat{z}_{t+1} - z_t\|^2 + \|z_t - \widehat{z}_t\|^2\right). \tag{14}$$

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

$$\mathrm{dist}_\star^2(\widehat{z}_{t+1}) + c'\|\widehat{z}_{t+1} - z_t\|^2 + c'\|z_t - \widehat{z}_t\|^2 \le \frac{\mathrm{dist}_\star^2(\widehat{z}_t) + c'\|\widehat{z}_t - z_{t-1}\|^2 + c'\|z_{t-1} - \widehat{z}_{t-1}\|^2}{1+\Omega(\eta^2C^2)} \tag{15}$$

for some universal constant c', which clearly indicates the linear convergence of $\mathrm{dist}_\star^2(\widehat{z}_t)$ 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 $(x_\star, y_\star)$ (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:

$$\begin{aligned}\mathrm{dist}_\star^2(\widehat{z}_{t+1}^s) \le \mathrm{dist}_\star^2(\widehat{z}_t^s) + \eta^2\|z_t^s - z_{t-1}^s\|^2 - \left(\|\widehat{z}_{t+1}^s - z_t^s\|^2 + \|z_t^s - \widehat{z}_t^s\|^2\right) \\ + 8\eta^2\|Q_t^s - Q_{t+1}^s\|^2 + 4\eta\|Q_t^s - Q_\star^s\|.\end{aligned} \tag{16}$$

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 $\eta\|Q_t^s - Q_\star^s\|$. 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 $-(\|\widehat{z}_{t+1}^s - z_t^s\|^2 + \|z_t^s - \widehat{z}_t^s\|^2)$. Indeed, in Steps 3-5, we show that part of them can be bounded by a weighted sum of $\{\|z_\tau^{s'} - z_{\tau+1}^{s'}\|^2\}_{s'\in\mathcal{S},\tau\le t}$.

Step 2 (Appendix C): Lower bounding $\left\|\widehat{z}^s_{t+1} - z^s_t\right\|^2 + \|z^s_t - \widehat{z}^s_t\|^2$. 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^s_t instead of $Q^s_\star$, we can only relate the bonus term to the duality gap with respect to Q^s_t. To further relate this to the duality gap with respect to $Q^s_\star$, we pay a quantity related to $\|Q^s_t - Q^s_\star\|$. Formally, we show in Lemma 25:

$$\left\|\widehat{z}^s_{t+1} - z^s_t\right\|^2 + \|z^s_t - \widehat{z}^s_t\|^2 \gtrsim \Omega(\eta^2\Delta^2(\widehat{z}^s_{t+1})) - \mathcal{O}(\eta\|Q^s_t - Q^s_\star\|),$$

where $\Delta(z^s) \triangleq \max_{x'^s, y'^s}\left(x^s Q^s_\star y'^s - x'^s Q^s_\star y^s\right)$ is the duality gap on state s with respect to $Q^s_\star$.[9]

Step 3 (Appendix D): Upper bounding $\left\|Q^s_{t+1} - Q^s_t\right\|^2$. $\left\|Q^s_{t+1} - Q^s_t\right\|^2$ is upper bounded by $\gamma^2\max_{s'}(V^{s'}_t - V^{s'}_{t-1})^2$ by the definition of Q^s_t. Furthermore, $V^{s'}_t - V^{s'}_{t-1}$ is a weighted sum of $\{\rho^{s'}_\tau - \rho^{s'}_{\tau-1}\}_{\tau=1}^{t-1}$ by the definition of $V^{s'}_t$, and also $\rho^{s'}_\tau - \rho^{s'}_{\tau-1} = x^{s'}_\tau Q^{s'}_\tau y^{s'}_\tau - x^{s'}_{\tau-1}Q^{s'}_{\tau-1}y^{s'}_{\tau-1} = \mathcal{O}(\|z^{s'}_\tau - z^{s'}_{\tau-1}\| + \|Q^{s'}_\tau - Q^{s'}_{\tau-1}\|)$. In sum, one can upper bound $\|Q^s_{t+1} - Q^s_t\|^2$ by a weighted sum of $\|z^{s'}_\tau - z^{s'}_{\tau-1}\|^2$ and $\|Q^{s'}_\tau - Q^{s'}_{\tau-1}\|^2$. After formalizing the above relations, we obtain the following inequality (see Lemma 28):

$$\left\|Q^s_{t+1} - Q^s_t\right\|^2 \leq \max_{s'} \frac{8\gamma^2}{(1-\gamma)^3}\sum_{\tau=1}^{t}\alpha^\tau_t\|z^{s'}_\tau - z^{s'}_{\tau-1}\|^2 + \max_{s'}\frac{2\gamma^2}{1+\gamma}\sum_{\tau=1}^{t}\alpha^\tau_t\|Q^{s'}_\tau - Q^{s'}_{\tau-1}\|^2 \quad (17)$$

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

Step 4 (Appendix E): Upper bounding $\|Q^s_t - Q^s_\star\|$ (Part 1). We first upper bound $\|Q^s_t - Q^s_\star\|$ with respect to the following weighted-regret quantity

$$\overline{\text{Reg}}_t \triangleq \max_s \max\left\{\sum_{\tau=1}^{t}\alpha^\tau_t(x^s_\tau - x^s_\star)Q^s_\tau y^s_\tau,\ \ \sum_{\tau=1}^{t}\alpha^\tau_t x^s_\tau Q^s_\tau(y^s_\star - y^s_\tau)\right\}.$$

To do so, we define $\Gamma_t = \max_s\|Q^s_t - Q^s_\star\|$ and show for the same coefficient α^τ_t mentioned earlier,

$$\begin{aligned}V^s_t &= \sum_{\tau=1}^{t}\alpha^\tau_t\rho^s_\tau = \sum_{\tau=1}^{t}\alpha^\tau_t x^s_\tau Q^s_\tau y^s_\tau \leq \sum_{\tau=1}^{t}\alpha^\tau_t x^s_\star Q^s_\tau y^s_\tau + \overline{\text{Reg}}_t \leq \sum_{\tau=1}^{t}\alpha^\tau_t x^s_\star Q^s_\star y^s_\tau + \sum_{\tau=1}^{t}\alpha^\tau_t\Gamma_\tau + \overline{\text{Reg}}_t\\ &\leq \sum_{\tau=1}^{t}\alpha^\tau_t x^s_\star Q^s_\star y^s_\star + \sum_{\tau=1}^{t}\alpha^\tau_t\Gamma_\tau + \overline{\text{Reg}}_t = V^s_\star + \sum_{\tau=1}^{t}\alpha^\tau_t\Gamma_\tau + \overline{\text{Reg}}_t\end{aligned}$$

where the last inequality is by the fact $\sum_{\tau=1}^t\alpha^\tau_t = 1$. Using the definition of Q^s_t again, we then have $Q^s_{t+1}(a,b) - Q^s_\star(a,b) = \gamma\mathbb{E}_{s'\sim p(\cdot|s,a,b)}\left[V^{s'}_t - V^{s'}_\star\right] \leq \gamma(\sum_{\tau=1}^t\alpha^\tau_t\Gamma_\tau + \overline{\text{Reg}}_t)$. By the same reasoning, we can also show $Q^s_{t+1}(a,b) - Q^s_\star(a,b) \geq -\gamma(\sum_{\tau=1}^t\alpha^\tau_t\Gamma_\tau + \overline{\text{Reg}}_t)$, and therefore we obtain the following recursive relation (Lemma 29)

$$\Gamma_{t+1} = \max_s\|Q^s_{t+1} - Q^s_\star\| \leq \gamma\left(\sum_{\tau=1}^{t}\alpha^\tau_t\Gamma_\tau + \overline{\text{Reg}}_t\right). \quad (18)$$

[9] Similar to the notation $\text{dist}_\star(\cdot)$, we also omit writing the s dependence for the function $\Delta(\cdot)$.

Step 5 (Appendix E): Upper bounding $\|Q^s_t - Q^s_\star\|$ (Part 2). In this step, we further relate $\overline{\text{Reg}}_t$ to $\{\|z^{s'}_\tau - z^{s'}_{\tau-1}\|^2\}_{\tau\le t, s'\in\mathcal{S}}$. From a one-step regret analysis of OGDA, we have the following (for Player 1):

$$(x^s_t - x^s_\star)\, Q^s_t y^s_t \le \frac{1}{2\eta}\Big(\text{dist}^2_\star(\widehat{x}^s_t) - \text{dist}^2_\star(\widehat{x}^s_{t+1})\Big) + \frac{4\eta}{(1-\gamma)^2}\|y^s_t - y^s_{t-1}\|^2 + 4\eta\|Q^s_t - Q^s_{t-1}\|^2.$$

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

$$\begin{aligned}
&\frac{\alpha^1_t \text{dist}^2_\star(\widehat{x}^s_1)}{2\eta} + \sum_{\tau=1}^t \frac{\alpha^\tau_t}{2\eta}\left(\text{dist}^2_\star(\widehat{x}^s_\tau) - \text{dist}^2_\star(\widehat{x}^s_{\tau+1})\right) \\
&\qquad + \frac{4\eta}{(1-\gamma)^2}\sum_{\tau=1}^t \alpha^\tau_t\|y^s_\tau - y^s_{\tau-1}\|^2 + 4\eta\sum_{\tau=1}^t \alpha^\tau_t\|Q^s_\tau - Q^s_{\tau-1}\|^2 \\
&\le \underbrace{\frac{1}{2\eta}\sum_{\tau=1}^t \alpha^\tau_t\alpha_{\tau-1}\text{dist}^2_\star(\widehat{z}^s_\tau)}_{\textbf{term}_1} + \underbrace{\frac{4\eta}{(1-\gamma)^2}\sum_{\tau=1}^t \alpha^\tau_t\|z^s_\tau - z^s_{\tau-1}\|^2}_{\textbf{term}_2} + \underbrace{4\eta\sum_{\tau=1}^t \alpha^\tau_t\|Q^s_\tau - Q^s_{\tau-1}\|^2}_{\textbf{term}_3}
\end{aligned} \tag{19}$$

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

$$\begin{aligned}
\sum_s \text{dist}^2_\star(\widehat{z}^s_{t+1}) \le \sum_s \text{dist}^2_\star(\widehat{z}^s_t) + \underbrace{\sum_{\tau=1}^t\sum_s \mu^s_\tau\alpha_{\tau-1}\text{dist}^2_\star(\widehat{z}^s_\tau)}_{\textbf{term}_4} + \underbrace{\sum_{\tau=1}^t\sum_s \nu^s_\tau\|z^s_\tau - z^s_{\tau-1}\|^2}_{\textbf{term}_5} \\
\underbrace{-\frac{1}{2}\sum_s\left(\left\|\widehat{z}^s_{t+1} - z^s_t\right\|^2 + \|z^s_t - \widehat{z}^s_t\|^2\right)}_{\textbf{term}_6} - \Omega\left(\eta^2\sum_s \Delta^2(\widehat{z}^s_{t+1})\right)
\end{aligned} \tag{20}$$

for some weights μ^s_τ and ν^s_τ (a large part of the analysis is devoted to precisely calculating these weights). Here, the $-\Omega\left(\eta^2\sum_s \Delta^2(\widehat{z}^s_{t+1})\right)$ term comes from Step 2; $\textbf{term}_4$ is a weighted sum of $\{\alpha_{\tau-1}\text{dist}^2_\star(\widehat{z}^{s'}_\tau)\}_{\tau\le t,s'\in\mathcal{S}}$ that comes from $\textbf{term}_1$ in Step 5; $\textbf{term}_5$ is a weighed sum of $\{\|z^{s'}_\tau - z^{s'}_{\tau-1}\|^2\}_{\tau\le t,s'\in\mathcal{S}}$ 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 $\textbf{term}_5$ over t is smaller than the sum of $\textbf{term}_6$ over t (hence they are canceled with each other). Rearranging and telescoping leads to

$$\eta^2\sum_{t=1}^T\sum_s \Delta^2(\widehat{z}^s_{t+1}) = \mathcal{O}\left(\sum_{t=1}^T\sum_{\tau=1}^t\sum_s \mu^s_\tau\alpha_{\tau-1}\text{dist}^2_\star(\widehat{z}^s_\tau)\right) = \mathcal{O}\left(\sum_{t=1}^T\sum_{\tau=1}^t\sum_s \mu^s_\tau\alpha_{\tau-1}\right).$$

As long as α_t is decreasing and going to zero, the right-hand side above can be shown to be sublinear in T. Further relating $\max_{x',y'}\left(V^s_{\widehat{x}_t,y'} - V^s_{x',\widehat{y}_t}\right)$ to $\Delta(\widehat{z}^s_t)$ (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 $\Delta(\widehat{z}^s_{t+1}) \geq C\mathrm{dist}_\star(\widehat{z}^s_{t+1})$. Similarly to how Eq. (14) is obtained, we use this in Eq. (20) and arrive at

$$\begin{aligned}\sum_s \mathrm{dist}^2_\star(\widehat{z}^s_{t+1}) \leq \frac{1}{1+\Omega(\eta^2C^2)}\sum_s \mathrm{dist}^2_\star(\widehat{z}^s_t) + \underbrace{\sum_{\tau=1}^{t}\sum_s \mu^s_\tau \alpha_{\tau-1}\mathrm{dist}^2_\star(\widehat{z}^s_\tau)}_{\textbf{term}_4} \\ + \underbrace{\sum_{\tau=1}^{t}\sum_s \nu^s_\tau \|z^s_\tau - z^s_{\tau-1}\|^2}_{\textbf{term}_5} - \Omega\left(\underbrace{\sum_s\left(\left\|\widehat{z}^s_{t+1} - z^s_t\right\|^2 + \|z^s_t - \widehat{z}^s_t\|^2\right)}_{\textbf{term}_6}\right)\end{aligned} \tag{21}$$

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 $\textbf{term}_4$. Fortunately, this term vanishes when t is large as long as α_t decreases and converges to zero. Second, in Eq. (14), the indices of the negative term $\|\widehat{z}_{t+1} - z_t\|^2 + \|z_t - \widehat{z}_t\|^2$ and the positive term $\eta^2\|z_t - z_{t-1}\|^2$ are only offset by 1 so that a simple rearrangement is enough to get Eq. (15), while in Eq. (21), the indices in $\textbf{term}_6$ and $\textbf{term}_5$ 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 $\textbf{term}_5$ can be canceled by the weighted sum of $\textbf{term}_6$. 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

Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, 2019.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. In *Conference on Learning Theory*, 2020.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, 2007.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. *International Conference on Machine Learning*, 2020.

Yu Bai, Chi Jin, and Tiancheng Yu. Near-optimal reinforcement learning with self-play. *Advances in Neural Information Processing Systems*, 2020.

Michael Bowling and Manuela Veloso. Rational and convergent learning in stochastic games. In *Proceedings of the 17th international joint conference on Artificial intelligence*, 2001.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 2002.

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Conference on Learning Theory*, 2012.

Vincent Conitzer and Tuomas Sandholm. Awesome: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. *Machine Learning*, 2007.

Constantinos Daskalakis, Dylan J. Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. In *Advances in neural information processing systems*, 2020.

Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

Jerzy A Filar and Boleslaw Tolwinski. On the algorithm of pollatschek and avi-ltzhak. 1991.

Andrew Gilpin, Javier Pena, and Tuomas Sandholm. First-order algorithm with $\mathcal{O}(\ln(1/\epsilon))$ convergence for ϵ-equilibrium in two-person zero-sum games. *Mathematical programming*, 2012.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. *Advances in neural information processing systems*, 2020.

Alan J Hoffman and Richard M Karp. On nonterminating stochastic games. *Management Science*, 1966.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems*, 2019.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, 2018.

Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei, and Mengxiao Zhang. Linear last-iterate convergence for matrix games and stochastic games. *arXiv preprint arXiv:2006.09517v1*, 2020.

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings*. 1994.

Boyi Liu, Zhuoran Yang, and Zhaoran Wang. Policy optimization in zero-sum markov games: Fictitious self-play provably attains nash equilibria, 2020. URL https://openreview.net/forum?id=c3MWGN_cTf.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. *International Conference on Machine Learning*, 2021.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *International Conference on Artificial Intelligence and Statistics*, 2020.

Julien Perolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. Approximate dynamic programming for two-player zero-sum markov games. In *International Conference on Machine Learning*, 2015.

Julien Pérolat, Bilal Piot, Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. Softened approximate policy iteration for markov games. In *International Conference on Machine Learning*, 2016.

Julien Perolat, Bilal Piot, and Olivier Pietquin. Actor-critic fictitious play in simultaneous move multistage games. In *International Conference on Artificial Intelligence and Statistics*, 2018.

MA Pollatschek and B Avi-Itzhak. Algorithms for stochastic games with geometrical interpretation. *Management Science*, 1969.

Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 1980.

Goran Radanovic, Rati Devidze, David Parkes, and Adish Singla. Learning to collaborate in markov decision processes. In *International Conference on Machine Learning*, 2019.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, 2013.

Muhammed O Sayin, Francesca Parise, and Asuman Ozdaglar. Fictitious play in zero-sum stochastic games. *arXiv preprint arXiv:2010.04223*, 2020.

Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 1953.

Aaron Sidford, Mengdi Wang, Lin Yang, and Yinyu Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, 2020.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 2017.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, 2015.

Csaba Szepesvári and Michael L Littman. A unified analysis of value-function-based reinforcement-learning algorithms. *Neural computation*, 1999.

Yi Tian, Yuanhao Wang, Tiancheng Yu, and Suvrit Sra. Online learning in unknown markov games. *International Conference on Machine Learning*, 2021.

J Van Der Wal. Discounted markov games: Generalized policy iteration method. *Journal of Optimization Theory and Applications*, 1978.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 2019.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems*, 2017.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. *International Conference on Learning Representations*, 2021.

Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. *Conference on Learning Theory*, 2020.

Kaiqing Zhang, Zhuoran Yang, and Tamer Basar. Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, 2019.

Kaiqing Zhang, Sham M Kakade, Tamer Başar, and Lin F Yang. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, 2020a.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Springer Studies in Systems, Decision and Control, Handbook on RL and Control*, 2020b.

Appendix A. Notations

A.1. Simplifications of the Notations

We define the following notations to simplify the proofs:

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

Besides, for a matrix Q, we define $\|Q\| = \max_{i,j} |Q_{ij}|$. To avoid cluttered notation, a product of the form $x^\top Qy$ 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 (α_t^τ) *For non-negative integers τ and t with $\tau \le t$, define* $\alpha_t^\tau = \alpha_\tau \prod_{i=\tau+1}^t (1-\alpha_i)$.

Definition 9 (δ_t^τ) *For non-negative integers τ and t with $\tau \le t$, define* $\delta_t^\tau \triangleq \prod_{i=\tau+1}^t (1-\alpha_i)$.

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

Definition 11 (λ_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 (λ_t^τ) *For positive integers τ and t with $\tau < t$, define* $\lambda_t^\tau = \alpha_\tau \prod_{i=\tau}^{t-1}\lambda_i$. *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_t^s) *For every state $s \in \mathcal{S}$, define the sequence* $\{J_t^s\}_{t=1,2,\ldots}$ *by*

$$\begin{aligned} J_1^s &= \|z_1^s - z_0^s\|^2, \\ J_t^s &= (1-\alpha_t)J_{t-1}^s + \alpha_t \left\|z_t^s - z_{t-1}^s\right\|^2, \qquad \forall t \ge 2. \end{aligned}$$

Furthermore, define $J_t \triangleq \max_s J_t^s$.

Definition 14 (K_t^s) *For every state $s \in \mathcal{S}$, define the sequence* $\{K_t^s\}_{t=1,2,\ldots}$ *by*

$$\begin{aligned} K_1^s &= \|Q_1^s - Q_0^s\|^2, \\ K_t^s &= (1-\alpha_t)K_{t-1}^s + \alpha_t \left\|Q_t^s - Q_{t-1}^s\right\|^2, \qquad \forall t \ge 2. \end{aligned}$$

Furthermore, define $K_t \triangleq \max_s K_t^s$.

Definition 15 $(\widehat{x}_{t\star}^s, \widehat{y}_{t\star}^s, \widehat{z}_{t\star}^s)$ *Define* $\widehat{x}_{t\star}^s = \Pi_{\mathcal{X}_\star^s}(\widehat{x}_t^s)$*, i.e., the projection of* $\widehat{x}_t^s$ *onto the set of optimal policy* $\mathcal{X}_\star^s$ *on state* s*. Similarly,* $\widehat{y}_{t\star}^s = \Pi_{\mathcal{Y}_\star^s}(\widehat{y}_t^s)$*, and* $\widehat{z}_{t\star}^s = \Pi_{\mathcal{Z}_\star^s}(\widehat{z}_t^s) = (\widehat{x}_{t\star}^s, \widehat{y}_{t\star}^s)$*.*

Definition 16 (Δ_t^s) *Define* $\Delta_t^s = \max_{x',y'} \left(\widehat{x}_t^s Q_\star^s y'^s - x'^s Q_\star^s \widehat{y}_t^s\right)$ *for all* $t \geq 1$*.*

Definition 17 $(\overline{\mathrm{Reg}}_t^s)$ *Define*

$$\overline{\mathrm{Reg}}_t^s = \max\left\{\sum_{\tau=1}^{t} \alpha_t^\tau (x_\tau^s - \widehat{x}_{t\star}^s) Q_\tau^s y_\tau^s, \quad \sum_{\tau=1}^{t} \alpha_t^\tau x_\tau^s Q_\tau^s (\widehat{y}_{t\star}^s - y_\tau^s)\right\}$$

and $\overline{\mathrm{Reg}}_t = \max_s \overline{\mathrm{Reg}}_t^s$*.*

Definition 18 (Γ_t) *Define* $\Gamma_t = \max_s \|Q_t^s - Q_\star^s\|$*.*

Definition 19 (θ_t^s) *Define* $\theta_t^s = \frac{1}{16}\|\widehat{z}_t^s - z_{t-1}^s\|^2 + \frac{1}{16}\|z_{t-1}^s - \widehat{z}_{t-1}^s\|^2$

Definition 20 (Z_t) *Define* $Z_t = \max_s \sum_{\tau=1}^t \alpha_t^\tau \alpha_{\tau-1} \mathrm{dist}_\star(\widehat{z}_\tau^s)$*.*

A.4. Assumptions on α_t and Simple Facts about α_t^τ

We require α_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* $h_t = \sum_{\tau=1}^t \alpha_t^\tau k_\tau$*.*

Proof We prove it by induction. When $t = 1$, since $\alpha_1 = 1$, $h_1 = k_1 = \alpha_1^1 k_1$. Assume that the formula is correct for h_t. Then

$$\begin{aligned} h_{t+1} &= (1 - \alpha_{t+1})h_t + \alpha_{t+1}k_{t+1} \\ &= (1 - \alpha_{t+1})\sum_{\tau=1}^{t} \alpha_t^\tau k_\tau + \alpha_{t+1}^{t+1} k_{t+1} \\ &= \sum_{\tau=1}^{t} \alpha_{t+1}^\tau k_\tau + \alpha_{t+1}^{t+1} k_{t+1} = \sum_{\tau=1}^{t+1} \alpha_{t+1}^\tau k_\tau. \end{aligned}$$

■

Corollary 22 *The following hold:*

- $V_t^s = \sum_{\tau=1}^t \alpha_t^\tau \rho_\tau^s$
- $J_t^s = \sum_{\tau=1}^t \alpha_t^\tau \left\|z_\tau^s - z_{\tau-1}^s\right\|^2$
- $K_t^s = \sum_{\tau=1}^t \alpha_t^\tau \left\|Q_\tau^s - Q_{\tau-1}^s\right\|^2$

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,*

$$\begin{aligned}(x_t^s - \widehat{x}_{t\star}^s)\, Q_t^s y_t^s \le \frac{1}{2\eta}\Big(&\mathrm{dist}_\star^2(\widehat{x}_t^s) - \mathrm{dist}_\star^2(\widehat{x}_{t+1}^s) - \|\widehat{x}_{t+1}^s - x_t^s\|^2 - \|x_t^s - \widehat{x}_t^s\|^2\Big)\\ &+ \frac{4\eta}{(1-\gamma)^2}\|y_t^s - y_{t-1}^s\|^2 + 4\eta\|Q_t^s - Q_{t-1}^s\|^2 + 3\varepsilon,\\ x_t^s Q_t^s(\widehat{y}_{t\star}^s - y_t^s) \le \frac{1}{2\eta}\Big(&\mathrm{dist}_\star^2(\widehat{y}_t^s) - \mathrm{dist}_\star^2(\widehat{y}_{t+1}^s) - \|\widehat{y}_{t+1}^s - y_t^s\|^2 - \|y_t^s - \widehat{y}_t^s\|^2\Big)\\ &+ \frac{4\eta}{(1-\gamma)^2}\|x_t^s - x_{t-1}^s\|^2 + 4\eta\|Q_t^s - Q_{t-1}^s\|^2 + 3\varepsilon.\end{aligned}$$

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

$$(x_t^s - \widehat{x}_{t\star}^s)^\top \ell_t^s \le \frac{1}{2\eta}\Big(\|\widehat{x}_t^s - \widehat{x}_{t\star}^s\|^2 - \left\|\widehat{x}_{t+1}^s - \widehat{x}_{t\star}^s\right\|^2 - \|\widehat{x}_{t+1}^s - x_t^s\|^2 - \|x_t^s - \widehat{x}_t^s\|^2\Big) + \eta\|\ell_t^s - \ell_{t-1}^s\|^2.$$

Since $\|\widehat{x}_t^s - \widehat{x}_{t\star}^s\|^2 = \mathrm{dist}_\star^2(\widehat{x}_t^s)$ and $\left\|\widehat{x}_{t+1}^s - \widehat{x}_{t\star}^s\right\|^2 \ge \mathrm{dist}_\star^2(\widehat{x}_{t+1}^s)$ by the definition of $\mathrm{dist}_\star(\cdot)$, we further have

$$(x_t^s - \widehat{x}_{t\star}^s)^\top \ell_t^s \le \frac{1}{2\eta}\Big(\mathrm{dist}_\star^2(\widehat{x}_t^s) - \mathrm{dist}_\star^2(\widehat{x}_{t+1}^s) - \|\widehat{x}_{t+1}^s - x_t^s\|^2 - \|x_t^s - \widehat{x}_t^s\|^2\Big) + \eta\|\ell_t^s - \ell_{t-1}^s\|^2. \tag{22}$$

By the definition of ℓ_t^s, we have

$$\begin{aligned}&\eta\left\|\ell_t^s - \ell_{t-1}^s\right\|^2\\ &\le \eta\left\|\ell_t^s - Q_t^s y_t^s + (Q_t^s - Q_{t-1}^s)y_t^s + Q_{t-1}^s(y_t^s - y_{t-1}^s) + Q_{t-1}^s y_{t-1}^s - \ell_{t-1}^s\right\|^2\\ &\le 4\eta\left\|\ell_t^s - Q_t^s y_t^s\right\|^2 + 4\eta\left\|(Q_t^s - Q_{t-1}^s)y_t^s\right\|^2 + 4\eta\left\|Q_{t-1}^s(y_t^s - y_{t-1}^s)\right\|^2 + 4\eta\left\|Q_{t-1}^s y_{t-1}^s - \ell_{t-1}^s\right\|^2\\ &\le 4\eta\left\|Q_t^s - Q_{t-1}^s\right\|^2 + \frac{4\eta}{(1-\gamma)^2}\left\|y_t^s - y_{t-1}^s\right\|^2 + 8\eta\varepsilon^2\end{aligned}$$

and

$$(x_t^s - \widehat{x}_{t\star}^s)\, Q_t^s y_t^s \le (x_t^s - \widehat{x}_{t\star}^s)\, \ell_t^s + 2\varepsilon.$$

Combining them with Eq. (22) and the fact that $\eta\varepsilon \le \frac{\eta}{1-\gamma} \le \frac{1}{8}$, we get the first inequality that we want to prove. The other inequality is similar. ■

Lemma 24 *For all $t \ge 1$,*

$$\mathrm{dist}_\star^2(\widehat{z}_{t+1}^s) \le \mathrm{dist}_\star^2(\widehat{z}_t^s) - 15\theta_{t+1}^s + \theta_t^s + 4\eta\Gamma_t + 8\eta^2\|Q_t^s - Q_{t-1}^s\|^2 + 6\eta\varepsilon.$$

Proof Summing up the two inequalities in Lemma 23, we get

$$\begin{aligned}
&2\eta(x_t^s - \widehat{x}_{t\star}^s)Q_t^s y_t^s + 2\eta x_t^s Q_t^s(\widehat{y}_{t\star}^s - y_t^s) \\
&\le \mathrm{dist}_\star^2(\widehat{z}_t^s) - \mathrm{dist}_\star^2(\widehat{z}_{t+1}^s) + \frac{4\eta^2}{(1-\gamma)^2}\|z_t^s - z_{t-1}^s\|^2 + 8\eta^2\|Q_t^s - Q_{t-1}^s\|^2 - \|\widehat{z}_{t+1}^s - z_t^s\|^2 - \|z_t^s - \widehat{z}_t^s\|^2 + 6\eta\varepsilon \\
&\le \mathrm{dist}_\star^2(\widehat{z}_t^s) - \mathrm{dist}_\star^2(\widehat{z}_{t+1}^s) + \frac{1}{32}\|z_t^s - z_{t-1}^s\|^2 + 8\eta^2\|Q_t^s - Q_{t-1}^s\|^2 - \|\widehat{z}_{t+1}^s - z_t^s\|^2 - \|z_t^s - \widehat{z}_t^s\|^2 + 6\eta\varepsilon \\
&\le \mathrm{dist}_\star^2(\widehat{z}_t^s) - \mathrm{dist}_\star^2(\widehat{z}_{t+1}^s) + \frac{1}{16}\left(\|z_t^s - \widehat{z}_t^s\|^2 + \|\widehat{z}_t^s - z_{t-1}^s\|^2\right) \\
&\qquad\qquad + 8\eta^2\|Q_t^s - Q_{t-1}^s\|^2 - \|\widehat{z}_{t+1}^s - z_t^s\|^2 - \|z_t^s - \widehat{z}_t^s\|^2 + 6\eta\varepsilon \\
&= \mathrm{dist}_\star^2(\widehat{z}_t^s) - \mathrm{dist}_\star^2(\widehat{z}_{t+1}^s) + 8\eta^2\|Q_t^s - Q_{t-1}^s\|^2 - \frac{15}{16}\|z_t^s - \widehat{z}_t^s\|^2 - \|\widehat{z}_{t+1}^s - z_t^s\|^2 + \frac{1}{16}\|\widehat{z}_t^s - z_{t-1}^s\|^2 + 6\eta\varepsilon
\end{aligned}$$

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

$$\begin{aligned}
2\eta(x_t^s - \widehat{x}_{t\star}^s)Q_t^s y_t^s + 2\eta x_t^s Q_t^s(\widehat{y}_{t\star}^s - y_t^s) &= 2\eta x_t^s Q_t^s \widehat{y}_{t\star}^s - 2\eta \widehat{x}_{t\star}^s Q_t^s y_t^s \\
&\ge 2\eta x_t^s Q_\star^s \widehat{y}_{t\star}^s - 2\eta \widehat{x}_{t\star}^s Q_\star^s y_t^s - 4\eta\Gamma_t \\
&\ge -4\eta\Gamma_t. \qquad \text{(by the optimality of } \widehat{x}_{t\star}^s \text{ and } \widehat{y}_{t\star}^s\text{)}
\end{aligned}$$

Combining the inequalities and using the definition of θ_t^s finish the proof. ■

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

Lemma 25 *For all $t \ge 1$, we have $20\theta_{t+1}^s + \eta\Gamma_t + 2\eta^2\epsilon^2 \ge \frac{\eta^2}{64}\left(\Delta_{t+1}^s\right)^2$.*

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

$$(\widehat{x}_{t+1}^s - \widehat{x}_t^s + \eta\ell_t^s)\cdot(x'^s - \widehat{x}_{t+1}^s) \ge 0 \tag{23}$$

for any $x'^s \in \Delta_{\mathcal{A}}$. Then by the definition of ℓ_t^s,

$$(\widehat{x}_{t+1}^s - \widehat{x}_t^s + \eta Q_t^s y_t^s)\cdot(x'^s - \widehat{x}_{t+1}^s) \ge (\widehat{x}_{t+1}^s - \widehat{x}_t^s + \eta\ell_t^s)\cdot(x'^s - \widehat{x}_{t+1}^s) - 2\eta\varepsilon \ge -2\eta\varepsilon \tag{24}$$

where in the last inequality we use Eq. (23). Thus we have for any $x'^s \in \Delta_{\mathcal{A}}$,

$$\begin{aligned}
&\sqrt{2}(\|\widehat{x}_{t+1}^s - x_t^s\| + \|x_t^s - \widehat{x}_t^s\|) \\
&\ge \sqrt{2}\|\widehat{x}_{t+1}^s - \widehat{x}_t^s\| \\
&\ge \|\widehat{x}_{t+1}^s - \widehat{x}_t^s\|^2 \\
&\ge (\widehat{x}_{t+1}^s - \widehat{x}_t^s)\cdot(x'^s - \widehat{x}_{t+1}^s) \\
&\ge \eta(\widehat{x}_{t+1}^s - x'^s)Q_t^s y_t^s - 2\eta\varepsilon && \text{(by Eq. (24))} \\
&= \eta(x_t^s - x'^s)Q_t^s y_t^s + \eta(\widehat{x}_{t+1}^s - x_t^s)Q_t^s y_t^s - 2\eta\varepsilon \\
&\ge \eta(x_t^s - x'^s)Q_t^s y_t^s - \frac{\eta\|\widehat{x}_{t+1}^s - x_t^s\|}{1-\gamma} - 2\eta\varepsilon.
\end{aligned}$$

Using the fact that $\frac{\eta}{1-\gamma} \le \frac{1}{16}$, we get

$$\|\widehat{x}^s_{t+1} - x^s_t\| + \|x^s_t - \widehat{x}^s_t\| + \sqrt{2}\eta\varepsilon \ge \frac{1}{\sqrt{2}+\frac{1}{16}}\left(\eta \max_{x'}(x^s_t - x'^s)Q^s_t y^s_t\right) \ge \frac{\eta}{2}\max_{x'}(x^s_t - x'^s)Q^s_t y^s_t.$$

Similarly, we have $\|\widehat{y}^s_{t+1} - y^s_t\| + \|y^s_t - \widehat{y}^s_t\| + \sqrt{2}\eta\varepsilon \ge \frac{\eta}{2}\max_{y'} x^s_t Q^s_t(y'^s - y^s_t)$. Combining them and using $\|z - z'\| \ge \frac{1}{2}\|x - x'\| + \frac{1}{2}\|y - y'\|$, we get

$$\begin{aligned}
&\|\widehat{z}^s_{t+1} - z^s_t\| + \|z^s_t - \widehat{z}^s_t\| + \sqrt{2}\eta\varepsilon \\
&\ge \frac{\eta}{4}\left(\max_{x'}(x^s_t - x'^s)Q^s_t y^s_t + \max_{y'} x^s_t Q^s_t(y'^s - y^s_t)\right) \\
&\ge \frac{\eta}{4}\Big(\max_{y'} x^s_t Q^s_t y'^s - \min_{x'} x'^s Q^s_t y^s_t\Big) \\
&= \frac{\eta}{4}\max_{y'}\left(\widehat{x}^s_{t+1}Q^s_\star y'^s + x^s_t(Q^s_t - Q^s_\star)y'^s + (x^s_t - \widehat{x}^s_{t+1})Q^s_\star y'^s\right) \\
&\qquad - \frac{\eta}{4}\min_{x'}\left(x'^s Q^s_\star \widehat{y}^s_{t+1} + x'^s(Q^s_t - Q^s_\star)y^s_t + x' Q^s_\star(y^s_t - \widehat{y}^s_{t+1})\right) \\
&\ge \frac{\eta}{4}\max_{x',y'}\left(\widehat{x}^s_{t+1}Q^s_\star y'^s - x'^s Q^s_\star \widehat{y}^s_{t+1}\right) - \frac{\eta\Gamma_t}{2} - \frac{\eta}{4(1-\gamma)}\left(\|\widehat{x}^s_{t+1} - x^s_t\| + \|\widehat{y}^s_{t+1} - y^s_t\|\right) \\
&\hspace{25em} (\|Q^s_\star\| \le \tfrac{1}{1-\gamma}) \\
&\ge \frac{\eta}{4}\Delta^s_{t+1} - \frac{\eta\Gamma_t}{2} - \frac{\eta}{2(1-\gamma)}\|\widehat{z}^s_{t+1} - z^s_t\| \qquad \text{(by the definition of } \Delta^s_{t+1}) \\
&\ge \frac{\eta}{4}\Delta^s_{t+1} - \frac{\eta\Gamma_t}{2} - \frac{1}{16}\|\widehat{z}^s_{t+1} - z^s_t\|. \qquad (25)
\end{aligned}$$

Then notice that we have

$$\begin{aligned}
&20\theta^s_{t+1} + \eta\Gamma_t + 2\eta^2\varepsilon^2 \\
&\ge \frac{289}{256}\|\widehat{z}^s_{t+1} - z^s_t\|^2 + \|z^s_t - \widehat{z}^s_t\|^2 + \frac{\eta^2\Gamma_t^2}{4} + 2\eta^2\varepsilon^2 \\
&\hspace{12em} \text{(by the definition of } \theta^s_{t+1} \text{ and that } \eta\Gamma_t \le \tfrac{\eta}{1-\gamma} \le 1) \\
&\ge \frac{1}{4}\left(\frac{17}{16}\|\widehat{z}^s_{t+1} - z^s_t\| + \|z^s_t - \widehat{z}^s_t\| + \frac{\eta\Gamma_t}{2} + \sqrt{2}\eta\varepsilon\right)^2 \qquad \text{(Cauchy-Schwarz inequality)} \\
&\ge \frac{\eta^2}{64}\left(\Delta^s_{t+1}\right)^2. \qquad \text{(by Eq. (25) and notice that } \Delta^s_{t+1} \ge 0)
\end{aligned}$$

■

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

$$\mathrm{dist}^2_\star(\widehat{z}^s_{t+1}) \le \mathrm{dist}^2_\star(\widehat{z}^s_t) - 5\theta^s_{t+1} + \theta^s_t - \frac{\eta^2}{128}(\Delta^s_{t+1})^2 + 5\eta\Gamma_t + 8\eta^2\|Q^s_t - Q^s_{t-1}\|^2 + 7\eta\varepsilon.$$

Proof Combining Lemma 25 with Lemma 24, we get

$$
\begin{aligned}
\mathrm{dist}_\star^2(\widehat{z}_{t+1}^s) &\le \mathrm{dist}_\star^2(\widehat{z}_t^s) - 5\theta_{t+1}^s - 10\theta_{t+1}^s + \theta_t^s + 4\eta\Gamma_t + 8\eta^2\|Q_t^s - Q_{t-1}^s\|^2 + 6\eta\varepsilon \\
&\le \mathrm{dist}_\star^2(\widehat{z}_t^s) - 5\theta_{t+1}^s - \left(\frac{\eta^2}{128}\left(\Delta_{t+1}^s\right)^2 - \frac{1}{2}\eta\Gamma_t - \eta^2\varepsilon^2\right) + \theta_t^s + 4\eta\Gamma_t + 8\eta^2\|Q_t^s - Q_{t-1}^s\|^2 + 6\eta\varepsilon \\
&\le \mathrm{dist}_\star^2(\widehat{z}_t^s) - 5\theta_{t+1}^s + \theta_t^s - \frac{\eta^2}{128}\left(\Delta_{t+1}^s\right)^2 + 5\eta\Gamma_t + 8\eta^2\|Q_t^s - Q_{t-1}^s\|^2 + 7\eta\varepsilon. \\
&\hspace{20em} (\eta\varepsilon \le 1)
\end{aligned}
$$

■

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 $t \ge 1$,*

$$
\mathrm{dist}_\star(\widehat{z}_{t+1}^s) + 4.5\theta_{t+1}^s \le \frac{1}{1+\eta^2 C'^2}(\mathrm{dist}_\star(\widehat{z}_t^s) + 4.5\theta_t^s) + 5\eta\Gamma_t + 8\eta^2\|Q_t^s - Q_{t-1}^s\|^2 - 3\theta_t^s + 7\eta\varepsilon.
$$

Proof By Theorem 5 of (Wei et al., 2021) or Lemma 3 of (Gilpin et al., 2012), we have

$$
\Delta_{t+1}^s \ge C\mathrm{dist}_\star(\widehat{z}_{t+1}^s)
$$

for some problem-dependent constant $0 < C \le \frac{1}{1-\gamma}$ (C depends on $\{Q_\star^s\}_s$). Thus Theorem 26 implies

$$
\mathrm{dist}_\star^2(\widehat{z}_{t+1}^s) + 5\theta_{t+1}^s \le \mathrm{dist}_\star^2(\widehat{z}_t^s) + \theta_t^s - \frac{\eta^2 C^2}{128}\mathrm{dist}_\star^2(\widehat{z}_{t+1}^s) + 5\eta\Gamma_t + 8\eta^2\|Q_t^s - Q_{t-1}^s\|^2 + 7\eta\varepsilon.
$$

By defining $C'^2 = \frac{C^2}{128}$, we further get

$$
\begin{aligned}
\mathrm{dist}_\star^2(\widehat{z}_{t+1}^s) + \frac{5}{1+\eta^2 C'^2}\theta_{t+1}^s &\le \frac{1}{1+\eta^2 C'^2}\left(\mathrm{dist}_\star^2(\widehat{z}_t^s) + \theta_t^s + 5\eta\Gamma_t + 8\eta^2\|Q_t^s - Q_{t-1}^s\|^2 + 7\eta\varepsilon\right) \\
&\le \frac{1}{1+\eta^2 C'^2}(\mathrm{dist}_\star^2(\widehat{z}_t^s) + \theta_t^s) + 5\eta\Gamma_t + 8\eta^2\|Q_t^s - Q_{t-1}^s\|^2 + 7\eta\varepsilon.
\end{aligned}
$$

Notice that $\frac{5}{1+\eta^2 C'^2} \ge \frac{5}{1+\frac{1}{16^2}\times\frac{1}{128}} \ge 4.5$. Thus we further have

$$
\mathrm{dist}_\star^2(\widehat{z}_{t+1}^s) + 4.5\theta_{t+1}^s \le \frac{1}{1+\eta^2 C'^2}(\mathrm{dist}_\star^2(\widehat{z}_t^s) + 4.5\theta_t^s) + 5\eta\Gamma_t + 8\eta^2\|Q_t^s - Q_{t-1}^s\|^2 - 3\theta_t^s + 7\eta\varepsilon
$$

where in the last inequality we use $\frac{1}{1+\eta^2 C'^2} \le \frac{4.5}{1+\eta^2 C'^2} - 3$ because $\eta^2 C'^2 \le \frac{1}{16^2} \times \frac{1}{128}$. ■

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

Lemma 28 *We have for $t \ge 2$ and all $s \in \mathcal{S}$,*

$$
\|Q_t^s - Q_{t-1}^s\|^2 \le \frac{8\gamma^2}{(1-\gamma)^3}J_{t-1} + \frac{2\gamma^2}{1+\gamma}K_{t-1} + \frac{16\gamma^2\varepsilon^2}{1-\gamma}.
$$

Proof It is equivalent to prove that for all $t \geq 1$,

$$\|Q_{t+1}^s - Q_t^s\|^2 \leq \frac{8\gamma^2}{(1-\gamma)^3} J_t + \frac{2\gamma^2}{1+\gamma} K_t + \frac{16\gamma^2\varepsilon^2}{1-\gamma}.$$

By definition,

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

we have

$$\|Q_{t+1}^s - Q_t^s\|^2 = \max_{a,b} (Q_{t+1}^s(a,b) - Q_t^s(a,b))^2 \leq \gamma^2 \max_{s'} \left(V_t^{s'} - V_{t-1}^{s'} \right)^2 \tag{26}$$

Now it suffices to upper bound $\left(V_t^s - V_{t-1}^s\right)^2$ for any s. By Corollary 22, we have $V_{t-1}^s = \sum_{\tau=1}^{t-1} \alpha_{t-1}^\tau \rho_\tau^s$. Therefore,

$$\begin{aligned} V_t^s - V_{t-1}^s &= \alpha_t \left(\rho_t^s - V_{t-1}^s \right) \\ &= \alpha_t \left(\rho_t^s - \sum_{\tau=0}^{t-1} \alpha_{t-1}^\tau \rho_\tau^s \right) \\ &= \alpha_t \left(\sum_{\tau=0}^{t-1} \alpha_{t-1}^\tau \left(\rho_t^s - \rho_\tau^s \right) \right) && \text{(because } \textstyle\sum_{\tau=0}^{t-1} \alpha_{t-1}^\tau = 1) \end{aligned}$$

In the following calculation, we omit the superscript s for simplicity. By defining $\text{diff}_h \triangleq |\rho_h - \rho_{h-1}|$, we have

$$\begin{aligned} (V_t - V_{t-1})^2 &\leq (\alpha_t)^2 \left(\sum_{\tau=0}^{t-1} \alpha_{t-1}^\tau (\rho_t - \rho_\tau) \right)^2 \\ &\leq (\alpha_t)^2 \left(\sum_{\tau=0}^{t-1} \alpha_{t-1}^\tau \sum_{h=\tau+1}^{t} (\rho_h - \rho_{h-1}) \right)^2 \\ &\leq (\alpha_t)^2 \left(\sum_{\tau=0}^{t-1} \alpha_{t-1}^\tau \sum_{h=\tau+1}^{t} \text{diff}_h \right)^2 \\ &= (\alpha_t)^2 \left(\sum_{h=1}^{t} \sum_{\tau=0}^{h-1} \alpha_{t-1}^\tau \text{diff}_h \right)^2 \\ &\leq (\alpha_t)^2 \left(\sum_{h=1}^{t} \delta_{t-1}^{h-1} \text{diff}_h \right)^2. && \text{(by Lemma 35)} \end{aligned}$$

Then we continue:

$$\begin{aligned} &(V_t - V_{t-1})^2 \\ &\leq (\alpha_t)^2 \left(\sum_{h=1}^{t} \delta_{t-1}^{h-1} \text{diff}_h \right)^2 \end{aligned}$$

$$
\begin{aligned}
&\leq (\alpha_t)^2 \left(\sum_{h=1}^{t} \delta_{t-1}^{h-1}\right)\left(\sum_{h=1}^{t} \delta_{t-1}^{h-1}\text{diff}_h^2\right) && \text{(Cauchy-Schwarz inequality)}\\
&\leq \left(\sum_{h=1}^{t} \alpha_h\delta_{t-1}^{h-1}\right)\left(\sum_{h=1}^{t} \alpha_h\delta_{t-1}^{h-1}\text{diff}_h^2\right) && (\alpha_t \leq \alpha_h \text{ for } h \leq t)\\
&\leq \sum_{\tau=1}^{t} \alpha_t^\tau \text{diff}_\tau^2 && \text{(note that } \alpha_h\delta_{t-1}^{h-1} = \alpha_h\textstyle\prod_{\tau=h}^{t-1}(1-\alpha_\tau) \leq \alpha_h \prod_{\tau=h+1}^{t}(1-\alpha_\tau) = \alpha_t^h)\\
&= \sum_{\tau=1}^{t} \alpha_t^\tau \Big(\rho_t - x_\tau Q_\tau y_\tau + x_\tau(Q_\tau - Q_{\tau-1})y_\tau + (x_\tau - x_{\tau-1})Q_{\tau-1}y_\tau\\
&\qquad\qquad + x_{\tau-1}Q_{\tau-1}(y_\tau - y_{\tau-1}) + x_{\tau-1}Q_{\tau-1}y_{\tau-1} - \rho_{t-1}\Big)^2\\
&\leq \sum_{\tau=1}^{t} \alpha_t^\tau \left(\frac{8\varepsilon^2}{1-\gamma} + \frac{2}{1+\gamma}\|Q_\tau - Q_{\tau-1}\|^2 + \frac{8}{(1-\gamma)^3}\|x_\tau - x_{\tau-1}\|^2 + \frac{8}{(1-\gamma)^3}\|y_\tau - y_{\tau-1}\|^2 + \frac{8\varepsilon^2}{1-\gamma}\right),
\end{aligned}
$$

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,

$$
\begin{aligned}
&\sum_{\tau=1}^{t} \alpha_t^\tau \|Q_\tau^s - Q_{\tau-1}^s\|^2 = K_t^s \leq K_t,\\
&\sum_{\tau=1}^{t} \alpha_t^\tau \|z_\tau^s - z_{\tau-1}^s\|^2 \leq J_t^s \leq J_t.
\end{aligned}
$$

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

$$
\|Q_{t+1}^s - Q_t^s\|^2 \leq \frac{8\gamma^2}{(1-\gamma)^3}J_t + \frac{2\gamma^2}{1+\gamma}K_t + \frac{16\gamma^2\varepsilon^2}{1-\gamma}.
$$

■

Appendix E. Proof for Steps 4 and 5: Bounding $\|Q_t^s - Q_\star^s\|$

Lemma 29 *For all $t \geq 2$,*

$$
\Gamma_t \leq \gamma\left(\sum_{\tau=1}^{t-1} \alpha_{t-1}^\tau \Gamma_\tau + \overline{\text{Reg}}_{t-1} + \varepsilon\right).
$$

Proof We proceed with

$$
\begin{aligned}
& Q^s_{t+1}(a,b) \\
&= \sigma(s,a,b) + \gamma \mathbb{E}_{s'\sim p(\cdot|s,a,b)}\left[V^{s'}_t\right] \\
&= \sigma(s,a,b) + \gamma \mathbb{E}_{s'\sim p(\cdot|s,a,b)}\left[\sum_{\tau=1}^{t} \alpha_t^\tau \rho^{s'}_\tau\right] && \text{(Corollary 22)} \\
&\le \sigma(s,a,b) + \gamma \mathbb{E}_{s'\sim p(\cdot|s,a,b)}\left[\sum_{\tau=1}^{t} \alpha_t^\tau x^{s'}_\tau Q^{s'}_\tau y^{s'}_\tau + \varepsilon\right] \\
& && \text{(by the definition of } \rho^{s'}_\tau \text{ and that } \textstyle\sum_{\tau=1}^t \alpha_t^\tau = 1 \text{ for } t \ge 1) \\
&\le \sigma(s,a,b) + \gamma \mathbb{E}_{s'\sim p(\cdot|s,a,b)}\left[\sum_{\tau=1}^{t} \alpha_t^\tau \widehat{x}^{s'}_{t\star} Q^{s'}_\tau y^{s'}_\tau + \overline{\text{Reg}}_t + \varepsilon\right] && \text{(by the definition of } \overline{\text{Reg}}_t) \\
&\le \sigma(s,a,b) + \gamma \mathbb{E}_{s'\sim p(\cdot|s,a,b)}\left[\sum_{\tau=1}^{t} \alpha_t^\tau \widehat{x}^{s'}_{t\star} Q^{s'}_\star y^{s'}_\tau + \sum_{\tau=1}^{t} \alpha_t^\tau \Gamma_\tau + \overline{\text{Reg}}_t + \varepsilon\right] \\
& && \text{(by the definition of } \Gamma_\tau) \\
&\le \sigma(s,a,b) + \gamma \mathbb{E}_{s'\sim p(\cdot|s,a,b)}\left[\sum_{\tau=1}^{t} \alpha_t^\tau \widehat{x}^{s'}_{t\star} Q^{s'}_\star y^{s'}_\star\right] + \gamma\left(\sum_{\tau=1}^{t} \alpha_t^\tau \Gamma_\tau + \overline{\text{Reg}}_t + \varepsilon\right) \\
& && \text{(by definition of } y^{s'}_\star) \\
&= \sigma(s,a,b) + \gamma \mathbb{E}_{s'\sim p(\cdot|s,a,b)}\left[V^{s'}_\star\right] + \gamma\left(\sum_{\tau=1}^{t} \alpha_t^\tau \Gamma_\tau + \overline{\text{Reg}}_t + \varepsilon\right) && (\textstyle\sum_{\tau=1}^t \alpha_t^\tau = 1 \text{ for } t \ge 1) \\
&= Q^s_\star(a,b) + \gamma\left(\sum_{\tau=1}^{t} \alpha_t^\tau \Gamma_\tau + \overline{\text{Reg}}_t + \varepsilon\right).
\end{aligned}
$$

Similarly,

$$
Q^s_{t+1}(a,b) \ge Q^s_\star(a,b) - \gamma\left(\sum_{\tau=1}^{t} \alpha_t^\tau \Gamma_\tau + \overline{\text{Reg}}_t + \varepsilon\right).
$$

They jointly imply

$$
\Gamma_{t+1} \le \gamma\left(\sum_{\tau=1}^{t} \alpha_t^\tau \Gamma_\tau + \overline{\text{Reg}}_t + \varepsilon\right).
$$

■

Lemma 30 *For any state s and time $t \ge 1$,*

$$
\overline{\text{Reg}}_t \le \frac{1}{2\eta} Z_t + \frac{4\eta}{(1-\gamma)^2} J_t + 4\eta K_t + 3\varepsilon.
$$

Proof Summing the first bound in Lemma 23 over $\tau = 1, \ldots, t$ with weights α_t^τ, and dropping negative terms $-\|\widehat{x}_{t+1}^s - x_t^s\|^2 - \|x_t^s - \widehat{x}_t^2\|^2$, we get

$$\begin{aligned}
&\sum_{\tau=1}^{t} \alpha_t^\tau \left(x_\tau^s - \widehat{x}_{\tau\star}^s\right) Q_\tau^s y_\tau^s \\
&\le \sum_{\tau=1}^{t} \frac{\alpha_t^\tau}{2\eta} \left(\mathrm{dist}_\star^2(\widehat{x}_\tau^s) - \mathrm{dist}_\star^2(\widehat{x}_{\tau+1}^s)\right) + \frac{4\eta}{(1-\gamma)^2} \sum_{\tau=1}^{t} \alpha_t^\tau \|y_\tau^s - y_{\tau-1}^s\|^2 + 4\eta \sum_{\tau=1}^{t} \alpha_t^\tau \|Q_\tau^s - Q_{\tau-1}^s\|^2 + 3\varepsilon \\
&\le \frac{\alpha_t^1}{2\eta} \mathrm{dist}_\star^2(\widehat{x}_1^s) + \sum_{\tau=2}^{t} \frac{\alpha_t^\tau - \alpha_t^{\tau-1}}{2\eta} \mathrm{dist}_\star(\widehat{x}_\tau^s) + \frac{4\eta}{(1-\gamma)^2} \sum_{\tau=1}^{t} \alpha_t^\tau \|y_\tau^s - y_{\tau-1}^s\|^2 + 4\eta \sum_{\tau=1}^{t} \alpha_t^\tau \|Q_\tau^s - Q_{\tau-1}^s\|^2 + 3\varepsilon \\
&\le \frac{\alpha_t^1}{2\eta} \mathrm{dist}_\star^2(\widehat{x}_1^s) + \sum_{\tau=2}^{t} \frac{\alpha_t^\tau - \alpha_t^{\tau-1}}{2\eta} \mathrm{dist}_\star(\widehat{x}_\tau^s) + \frac{4\eta}{(1-\gamma)^2} J_t^s + 4\eta K_t^s + 3\varepsilon. \qquad (27)
\end{aligned}$$

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

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

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

$$\begin{aligned}
\sum_{\tau=1}^{t} \alpha_t^\tau \left(x_\tau^s - \widehat{x}_{\tau*}^s\right) Q_\tau^s y_\tau^s &\le \frac{\alpha_t^1}{2\eta} \mathrm{dist}_\star(\widehat{x}_1^s) + \frac{1}{2\eta} \sum_{\tau=2}^{t} \alpha_t^\tau \alpha_{\tau-1} \mathrm{dist}_\star(\widehat{x}_\tau^s) + \frac{4\eta}{(1-\gamma)^2} J_t^s + 4\eta K_t^s + 3\varepsilon \\
&= \frac{1}{2\eta} \sum_{\tau=1}^{t} \alpha_t^\tau \alpha_{\tau-1} \mathrm{dist}_\star(\widehat{x}_\tau^s) + \frac{4\eta}{(1-\gamma)^2} J_t^s + 4\eta K_t^s + 3\varepsilon \\
&\qquad \text{(recall that } \alpha_0 = 1\text{)}
\end{aligned}$$

Using $J_n^s \le J_n, K_n^s \le K_n$, and the definition of $Z_t, \overline{\mathrm{Reg}}_t$ 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 $t \ge 1$,*

$$\begin{aligned}
&5\eta\Gamma_t + 8\eta^2 \left\|Q_t^s - Q_{t-1}^s\right\|^2 \\
&\le \max_{s'} \left(\frac{C_1 \eta^2}{(1-\gamma)^4} \sum_{\tau=1}^{t-1} \beta_t^\tau (\theta_\tau^{s'} + \theta_{\tau+1}^{s'}) + C_2 \sum_{\tau=1}^{t-1} \beta_t^\tau \alpha_{\tau-1} \mathrm{dist}_\star^2(\widehat{z}_\tau^{s'}) \right) + 80\beta_t^1 + \frac{80\eta\varepsilon}{(1-\gamma)^2}.
\end{aligned}$$

where $C_1 = 1152 \times 80$ and $C_2 = 10$.

Proof By Lemma 28, for all $t \geq 2$,

$$\eta^2 \|Q_t^s - Q_{t-1}^s\|^2 \leq \frac{8\eta^2\gamma^2}{(1-\gamma)^3} J_{t-1} + \frac{2\eta^2\gamma^2}{1+\gamma} K_{t-1} + \frac{16\eta^2\varepsilon^2}{1-\gamma}. \tag{28}$$

By Lemma 29 and Lemma 30, for all $t \geq 2$,

$$\eta\Gamma_t \leq \gamma \sum_{\tau=1}^{t-1} \alpha_{t-1}^\tau \eta\Gamma_\tau + \frac{4\eta^2}{(1-\gamma)^2} J_{t-1} + 4\eta^2 K_{t-1} + \frac{1}{2} Z_{t-1} + 4\eta\varepsilon. \tag{29}$$

Now, multiply Eq. (29) with $\frac{1-\gamma}{16}$, and then add it to Eq. (28). Then we get that for $t \geq 2$,

$$\begin{aligned}
&\eta^2 \|Q_t^s - Q_{t-1}^s\|^2 + \frac{1-\gamma}{16}\eta\Gamma_t \\
&\leq \gamma \sum_{\tau=1}^{t-1} \alpha_{t-1}^\tau \left(\frac{1-\gamma}{16}\eta\Gamma_\tau\right) + \left(\frac{8\gamma^2}{(1-\gamma)^3} + \frac{1}{4(1-\gamma)}\right)\eta^2 J_{t-1} + \\
&\qquad\qquad \left(\frac{2\gamma^2}{1+\gamma} + \frac{1-\gamma}{4}\right)\eta^2 K_{t-1} + \frac{(1-\gamma)Z_{t-1}}{32} + \frac{\eta\varepsilon}{2} \\
&\leq \gamma \sum_{\tau=1}^{t-1} \alpha_{t-1}^\tau \left(\frac{1-\gamma}{16}\eta\Gamma_\tau\right) + \frac{9}{(1-\gamma)^3}\eta^2 J_{t-1} + \gamma\eta^2 K_{t-1} + \frac{(1-\gamma)Z_{t-1}}{32} + \frac{\eta\varepsilon}{2} \\
&\qquad\qquad \text{(see explanation below)} \\
&\leq \gamma \sum_{\tau=1}^{t-1} \alpha_{t-1}^\tau \left(\frac{1-\gamma}{16}\eta\Gamma_\tau + \eta^2 \max_{s'} \|Q_\tau^{s'} - Q_{\tau-1}^{s'}\|^2\right) + \frac{9}{(1-\gamma)^3}\eta^2 J_{t-1} + \frac{(1-\gamma)Z_{t-1}}{32} + \frac{\eta\varepsilon}{2},
\end{aligned}$$

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 $K_{t-1}^s = \sum_{\tau=1}^{t-1} \alpha_{t-1}^\tau \|Q_\tau^s - Q_{\tau-1}^s\|^2$. Define the new variable

$$u_t = \eta^2 \max_s \|Q_t^s - Q_{t-1}^s\|^2 + \frac{1-\gamma}{16}\eta\Gamma_t.$$

Then the above implies that for all $t \geq 2$,

$$u_t \leq \gamma \sum_{\tau=1}^{t-1} \alpha_{t-1}^\tau u_\tau + \frac{9}{(1-\gamma)^3}\eta^2 J_{t-1} + \frac{(1-\gamma)Z_{t-1}}{32} + \frac{\eta\varepsilon}{2}. \tag{30}$$

Observe that Eq. (30) is in the form of Lemma 33 with the following choices:

$$\begin{aligned}
&g_t = u_t, \quad \forall t \geq 1, \\
&h_t = \begin{cases} u_t + \frac{\eta\varepsilon}{2} & \text{for } t = 1 \\ \frac{9\eta^2}{(1-\gamma)^3} J_{t-1} + \frac{(1-\gamma)}{32} Z_{t-1} + \frac{\eta\varepsilon}{2} & \text{for } t \geq 2 \end{cases}
\end{aligned}$$

and get that for $t \geq 2$,

$$\begin{aligned}
u_t &\leq \frac{9\eta^2}{(1-\gamma)^3} \sum_{\tau=2}^{t} \beta_t^\tau J_{\tau-1} + \frac{1-\gamma}{32} \sum_{\tau=2}^{t} \beta_t^\tau Z_{\tau-1} + \beta_t^1 u_1 + \frac{\eta\varepsilon}{2} \sum_{\tau=1}^{t} \beta_t^\tau \\
&\leq \frac{9\eta^2}{(1-\gamma)^3} \sum_{\tau=2}^{t} \beta_t^\tau J_{\tau-1} + \frac{1-\gamma}{32} \sum_{\tau=2}^{t} \beta_t^\tau Z_{\tau-1} + (1-\gamma)\beta_t^1 + \frac{\eta\varepsilon}{1-\gamma} \qquad \text{(by Lemma 38)}
\end{aligned}$$

because $u_t \le \frac{\eta^2}{(1-\gamma)^2} + \frac{1-\gamma}{16} \le \frac{1-\gamma}{2}$. Further using Lemma 34 on the first two terms on the right-hand side, and noticing that $\frac{1}{\gamma^2} \le 4$, we further get that for $t \ge 2$,

$$\begin{aligned} u_t &\le \max_s \frac{36\eta^2}{(1-\gamma)^3} \sum_{\tau=1}^{t-1} \beta_t^\tau \left\| z_\tau^s - z_{\tau-1}^s \right\|^2 + \frac{1-\gamma}{8} \sum_{\tau=1}^{t-1} \beta_t^\tau \alpha_{\tau-1} \mathrm{dist}_\star^2(\widehat{z}_\tau^s) + (1-\gamma)\beta_t^1 + \frac{\eta\varepsilon}{1-\gamma} \\ &\le \max_s \frac{72\eta^2}{(1-\gamma)^3} \sum_{\tau=1}^{t-1} \beta_t^\tau (\| z_\tau^s - \widehat{z}_\tau^s \|^2 + \| \widehat{z}_\tau^s - z_{\tau-1}^s \|^2) + \\ &\qquad\qquad \frac{1-\gamma}{8} \sum_{\tau=1}^{t-1} \beta_t^\tau \alpha_{\tau-1} \mathrm{dist}_\star^2(\widehat{z}_\tau^s) + (1-\gamma)\beta_t^1 + \frac{\eta\varepsilon}{1-\gamma} \\ &\le \max_s \frac{1152\eta^2}{(1-\gamma)^3} \sum_{\tau=1}^{t-1} \beta_t^\tau (\theta_{\tau+1}^s + \theta_\tau^s) + \frac{1-\gamma}{8} \sum_{\tau=1}^{t-1} \beta_t^\tau \alpha_{\tau-1} \mathrm{dist}_\star^2(\widehat{z}_\tau^s) + (1-\gamma)\beta_t^1 + \frac{\eta\varepsilon}{1-\gamma}. \end{aligned} \tag{31}$$

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

Proof of Theorem 1. Define $C_\alpha(T) \triangleq 1 + \sum_{t=1}^T \alpha_t$ and let C_β be an upper bound of $\sum_{t=\tau}^\infty \beta_t^\tau$ for any τ. With the choice of α_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 $C_\beta \le \frac{2}{1-\gamma} + 3$. Define $S = |\mathcal{S}|$.

Combining Lemma 31 and Theorem 26, we get that for $t \ge 1$,

$$\begin{aligned} \frac{\eta^2}{128}(\Delta_{t+1}^s)^2 &\le \mathrm{dist}_\star^2(\widehat{z}_t^s) - \mathrm{dist}_\star^2(\widehat{z}_{t+1}^s) - 5\theta_{t+1}^s + \theta_t^s \\ &\quad + \max_{s'} \left(\frac{C_1\eta^2}{(1-\gamma)^4} \sum_{\tau=1}^{t-1} \beta_t^\tau (\theta_\tau^{s'} + \theta_{\tau+1}^{s'}) + C_2 \sum_{\tau=1}^{t-1} \beta_t^\tau \alpha_{\tau-1} \mathrm{dist}_\star^2(\widehat{z}_\tau^{s'}) \right) + 80\beta_t^1 + \frac{87\eta\varepsilon}{(1-\gamma)^2}. \end{aligned}$$

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

$$\begin{aligned} \frac{\eta^2}{128} \sum_{t=1}^T \sum_s (\Delta_t^s)^2 &\le \mathcal{O}(S) - \sum_{t=1}^T 4\Theta_t + \frac{C_1 S \eta^2}{(1-\gamma)^4} \sum_{t=1}^T \sum_{\tau=1}^{t-1} \beta_t^\tau (\Theta_\tau + \Theta_{\tau+1}) \\ &\quad + \mathcal{O}\left(S \sum_{t=1}^T \sum_{\tau=1}^{t-1} \beta_t^\tau \alpha_{\tau-1} + S \sum_{t=1}^T \beta_t^1 + \frac{S\eta\varepsilon T}{(1-\gamma)^2} \right) \end{aligned} \tag{32}$$

since $\mathrm{dist}_\star^2(\widehat{z}_\tau^s) = \mathcal{O}(1)$ and $\Theta_1 = \mathcal{O}(S)$. Notice that the following hold

$$\sum_{t=1}^T \sum_{\tau=1}^{t-1} \beta_t^\tau (\Theta_\tau + \Theta_{\tau+1}) \le \sum_{\tau=1}^{T-1} \sum_{t=\tau}^T \beta_t^\tau (\Theta_\tau + \Theta_{\tau+1}) \le 2C_\beta \sum_{\tau=1}^T \Theta_\tau,$$

$$\sum_{t=1}^T \sum_{\tau=1}^{t-1} \beta_t^\tau \alpha_{\tau-1} \le \sum_{\tau=1}^T \sum_{t=\tau}^T \beta_t^\tau \alpha_{\tau-1} \le C_\beta \sum_{\tau=1}^T \alpha_{\tau-1} = C_\alpha(T) C_\beta,$$

and $\sum_{t=1}^{T}\beta_t^1 \le C_\beta$. Combining these three inequalities with Eq. (32), we get

$$\begin{aligned}\sum_{t=1}^{T}\sum_{s}(\Delta_t^s)^2 &= \frac{128}{\eta^2}\sum_{t=1}^{T}\left(-4+2C_\beta\frac{C_1S\eta^2}{(1-\gamma)^4}\right)\Theta_t + \mathcal{O}\left(\frac{SC_\alpha(T)C_\beta}{\eta^2}+\frac{S\varepsilon T}{\eta(1-\gamma)^2}\right)\\ &= \mathcal{O}\left(\frac{SC_\alpha(T)C_\beta}{\eta^2}+\frac{S\varepsilon T}{\eta(1-\gamma)^2}\right).\end{aligned}$$

(by our choice of η, we have $2C_\beta\frac{C_1S\eta^2}{(1-\gamma)^4}\le\frac{10C_1S\eta^2}{(1-\gamma)^5}\le 4$)

By Cauchy-Schwarz inequality, we further have

$$\sum_{t=1}^{T}\sum_{s}\Delta_t^s \le \sqrt{ST}\left(\sum_{t=1}^{T}\sum_{s}(\Delta_t^s)^2\right)^{\frac{1}{2}} = \mathcal{O}\left(\frac{S\sqrt{C_\alpha(T)C_\beta T}}{\eta}+\frac{ST\sqrt{\epsilon}}{\sqrt{\eta}(1-\gamma)}\right).$$

Finally, by Lemma 32, we get

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\max_{s,x',y'}\left(V^s_{\widehat{x}_t,y'}-V^s_{x',\widehat{y}_t}\right) &\le \frac{2}{1-\gamma}\frac{1}{T}\sum_{t=1}^{T}\max_{s}\Delta_t^s = \mathcal{O}\left(\frac{S\sqrt{C_\alpha(T)C_\beta}}{\eta(1-\gamma)\sqrt{T}}+\frac{S\sqrt{\epsilon}}{\sqrt{\eta}(1-\gamma)^2}\right)\\ &= \mathcal{O}\left(\frac{S\sqrt{\log T}}{\eta(1-\gamma)^2\sqrt{T}}+\frac{S\sqrt{\epsilon}}{\sqrt{\eta}(1-\gamma)^2}\right).\end{aligned}$$

■

Proof of Theorem 2. Combining Lemma 27 and Lemma 31, we get that for all $t\ge 1$,

$$\begin{aligned}&\mathrm{dist}_\star^2(\widehat{z}_{t+1}^s)+4.5\theta_{t+1}^s\\ &\le \frac{1}{1+\eta^2C'^2}\left(\mathrm{dist}_\star^2(\widehat{z}_t^s)+4.5\theta_t^s\right)\\ &+\max_{s'}\left(\frac{C_1\eta^2}{(1-\gamma)^4}\sum_{\tau=1}^{t-1}\beta_t^\tau(\theta_\tau^{s'}+\theta_{\tau+1}^{s'})+C_2\sum_{\tau=1}^{t-1}\beta_t^\tau\alpha_{\tau-1}\mathrm{dist}_\star^2(\widehat{z}_\tau^{s'})\right)+80\beta_t^1-3\theta_t^s+\frac{87\eta\varepsilon}{(1-\gamma)^2}.\end{aligned}$$

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

$$\begin{aligned}L_{t+1}+4.5\Theta_{t+1}\le\frac{1}{1+\eta^2C'^2}(L_t+4.5\Theta_t)+\frac{C_1S\eta^2}{(1-\gamma)^4}\sum_{\tau=1}^{t-1}\beta_t^\tau(\Theta_\tau+\Theta_{\tau+1})+\\ C_2S\sum_{\tau=1}^{t-1}\beta_t^\tau\alpha_{\tau-1}L_\tau+80S\beta_t^1-3\Theta_t+\frac{87S\eta\varepsilon}{(1-\gamma)^2}.\end{aligned}\tag{33}$$

The key idea of the following analysis is to use the negative (bonus) term $-3\Theta_t$ to cancel the positive (penalty) term $\frac{C_1S\eta^2}{(1-\gamma)^4}\sum_{\tau=1}^{t}\beta_t^\tau\Theta_\tau$. 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 λ_{t+1}^τ:

$$\sum_{\tau=2}^{t+1}\lambda_{t+1}^\tau(L_\tau+4.5\Theta_\tau)$$

$$
\begin{aligned}
&= \sum_{\tau=1}^{t} \lambda_{t+1}^{\tau+1}(L_{\tau+1} + 4.5\Theta_{\tau+1}) && \text{(re-indexing)}\\
&\leq \sum_{\tau=1}^{t} \lambda_{t}^{\tau}(L_{\tau+1} + 4.5\Theta_{\tau+1}) && \text{(Lemma 37)}\\
&\leq \frac{1}{1+\eta^2 C'^2}\sum_{\tau=1}^{t}\lambda_t^\tau(L_\tau + 4.5\Theta_\tau) + \frac{C_1 S\eta^2}{(1-\gamma)^4}\sum_{\tau=1}^{t}\lambda_t^\tau\sum_{i=1}^{\tau-1}\beta_\tau^i(\Theta_i+\Theta_{i+1})\\
&\qquad + C_2 S\sum_{\tau=1}^{t}\lambda_t^\tau\sum_{i=1}^{\tau-1}\beta_\tau^i\alpha_{i-1}L_i + 80S\sum_{\tau=1}^{t}\lambda_t^\tau\beta_\tau^1 - 3\sum_{\tau=1}^{t}\lambda_t^\tau\Theta_\tau + \frac{87S\eta\varepsilon}{(1-\gamma)^2}\sum_{\tau=1}^{t}\lambda_t^\tau\\
&\leq \frac{1}{1+\eta^2 C'^2}\sum_{\tau=1}^{t}\lambda_t^\tau(L_\tau + 4.5\Theta_\tau) + \frac{C_1 S\eta^2}{(1-\gamma)^4}\sum_{i=1}^{t-1}\left(\sum_{\tau=i}^{t}\lambda_t^\tau\beta_\tau^i\right)(\Theta_i+\Theta_{i+1})\\
&\qquad + C_2 S\sum_{i=1}^{t-1}\left(\sum_{\tau=i}^{t}\lambda_t^\tau\beta_\tau^i\right)\alpha_{i-1}L_i + 80S\sum_{\tau=1}^{t}\lambda_t^\tau\beta_\tau^1 - 3\sum_{\tau=1}^{t}\lambda_t^\tau\Theta_\tau + \frac{87S\eta\varepsilon}{(1-\gamma)^2}\sum_{\tau=1}^{t}\lambda_t^\tau\\
&\leq \frac{1}{1+\eta^2 C'^2}\sum_{\tau=1}^{t}\lambda_t^\tau(L_\tau + 4.5\Theta_\tau) + \frac{3C_1 S\eta^2}{(1-\gamma)^5}\sum_{\tau=1}^{t-1}\lambda_t^\tau(\Theta_\tau+\Theta_{\tau+1})\\
&\qquad + \frac{3C_2 S}{1-\gamma}\sum_{\tau=1}^{t-1}\lambda_t^\tau\alpha_{\tau-1}L_\tau + \frac{240S}{1-\gamma}\lambda_t^1 - 3\sum_{\tau=1}^{t}\lambda_t^\tau\Theta_\tau + \frac{87S\eta\varepsilon}{(1-\gamma)^2}\sum_{\tau=1}^{t}\lambda_t^\tau && \text{(by Lemma 36)}\\
&\leq \frac{1}{1+\eta^2 C'^2}\sum_{\tau=1}^{t}\lambda_t^\tau(L_\tau + 4.5\Theta_\tau) + \frac{6C_1 S\eta^2}{(1-\gamma)^5}\sum_{\tau=1}^{t}\lambda_t^\tau\Theta_\tau\\
&\qquad + \frac{3C_2 S}{1-\gamma}\sum_{\tau=1}^{t-1}\lambda_t^\tau\alpha_{\tau-1}L_\tau + \frac{240S}{1-\gamma}\lambda_t^1 - 3\sum_{\tau=1}^{t}\lambda_t^\tau\Theta_\tau + \frac{87S\eta\varepsilon}{(1-\gamma)^2}\sum_{\tau=1}^{t}\lambda_t^\tau\\
&\leq \frac{1}{1+\eta^2 C'^2}\sum_{\tau=1}^{t}\lambda_t^\tau(L_\tau + 4.5\Theta_\tau) + \frac{3C_2 S}{1-\gamma}\sum_{\tau=1}^{t-1}\lambda_t^\tau\alpha_{\tau-1}L_\tau + \frac{240S}{1-\gamma}\lambda_t^1 + \frac{87S\eta\varepsilon}{(1-\gamma)^2}\sum_{\tau=1}^{t}\lambda_t^\tau\\
&\hspace{20em} \text{(by our choice of } \eta,\ \tfrac{6C_1S\eta^2}{(1-\gamma)^5} \leq 3)
\end{aligned}
$$

where in the second-to-last inequality we use Lemma 41: with the special choice of α_t specified in the theorem, we have $\lambda_t^\tau = \alpha_t \leq \lambda_t^{\tau+1}$ for $\tau \leq t-1$.

Let $t_0 = \min\left\{\tau : \frac{3C_2S}{1-\gamma}\alpha_\tau \leq \frac{\eta^2C'^2}{2}\right\}$. Then we have

$$
\begin{aligned}
&\sum_{\tau=2}^{t+1}\lambda_{t+1}^\tau(L_\tau + 4.5\Theta_\tau)\\
&\leq \left(\frac{1}{1+\eta^2C'^2} + \frac{\eta^2C'^2}{2}\right)\sum_{\tau=1}^{t}\lambda_t^\tau(L_\tau + 4.5\Theta_\tau) + \frac{3C_2S}{1-\gamma}\sum_{\tau=1}^{\min\{t_0,t\}}\lambda_t^\tau\alpha_{\tau-1}L_\tau + \frac{240S}{1-\gamma}\lambda_t^1 + \frac{87S\eta\varepsilon}{(1-\gamma)^2}\sum_{\tau=1}^{t}\lambda_t^\tau\\
&\leq \frac{1}{1+0.1\eta^2C'^2}\sum_{\tau=3}^{t}\lambda_t^\tau(L_\tau + 4.5\Theta_\tau) + \frac{12C_2S^2}{1-\gamma}\sum_{\tau=1}^{\min\{t_0,t\}}\lambda_t^\tau\alpha_{\tau-1} + \frac{240S}{1-\gamma}\lambda_t^1 + \frac{87S\eta\varepsilon}{(1-\gamma)^2}\sum_{\tau=1}^{t}\lambda_t^\tau.\\
&\hspace{12em} (\eta C' \leq 2^{-15} \text{ according to Lemma 27, } L_\tau \leq S\cdot\max_{z,z'}\|z-z'\|^2 \leq 4S)
\end{aligned}
$$

Finally, we add $\lambda_{t+1}^1(L_1+4.5\Theta_1)$ to both sides, and note that

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

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

$$\begin{aligned}&\sum_{\tau=1}^{t+1}\lambda_{t+1}^\tau(L_\tau+4.5\Theta_\tau)\\&\le\frac{1}{1+0.1\eta^2C'^2}\sum_{\tau=1}^{t}\lambda_t^\tau(L_\tau+4.5\Theta_\tau)+\frac{12C_2S^2}{1-\gamma}\sum_{\tau=1}^{\min\{t_0,t\}}\lambda_t^\tau\alpha_{\tau-1}+\frac{240S}{1-\gamma}\lambda_t^1+22S\lambda_t^1+\frac{87S\eta\varepsilon}{(1-\gamma)^2}\sum_{\tau=1}^{t}\lambda_t^\tau\\&\le\frac{1}{1+0.1\eta^2C'^2}\sum_{\tau=1}^{t}\lambda_t^\tau(L_\tau+4.5\Theta_\tau)+\frac{274C_2S^2}{1-\gamma}\sum_{\tau=1}^{\min\{t_0,t\}}\lambda_t^\tau\alpha_{\tau-1}+\frac{87S\eta\varepsilon}{(1-\gamma)^2}\sum_{\tau=1}^{t}\lambda_t^\tau.\end{aligned}$$

Define

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

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

$$Y_{t+1}\le\frac{1}{1+0.1\eta^2C'^2}Y_t+\frac{274C_2S^2t_0}{1-\gamma}\lambda_t^{\min\{t_0,t\}}+\frac{87S\eta\varepsilon}{(1-\gamma)^2}\sum_{\tau=1}^{t}\lambda_t^\tau.$$

(upper bounding $\alpha_{\tau-1}$ 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

$$\begin{aligned}Y_t\le Y_1(1+0.1\eta^2C'^2)^{-t}&+\frac{2}{0.1\eta^2C'^2}\left(\frac{274C_2S^2t_0}{1-\gamma}+\frac{87S\eta\varepsilon}{(1-\gamma)^2}\sup_{t'\in[1,\frac{t}{2}]}\sum_{\tau=1}^{t'}\lambda_t^\tau\right)(1+0.1\eta^2C'^2)^{-\frac{t}{2}}\\&+\frac{2}{0.1\eta^2C'^2}\left(\frac{274C_2S^2t_0}{1-\gamma}\sup_{t'\in[\frac{t}{2},t]}\lambda_{t'}^{\min\{t_0,t'\}}+\frac{87S\eta\varepsilon}{(1-\gamma)^2}\sup_{t'\in[\frac{t}{2},t]}\sum_{\tau=1}^{t'}\lambda_t^\tau\right).\quad(\lambda_t^\tau\le 1)\end{aligned}$$

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

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

$$\sup_{t'\in[1,t]}\sum_{\tau=1}^{t'}\lambda_t^\tau\le\sup_{t'\in[1,t]}t'\alpha_t\le\frac{t(H+1)}{H+t}\le H+1=\mathcal{O}\left(\frac{1}{1-\gamma}\right)\qquad\text{(Lemma 41)}$$

$$\sup_{t'\in[\frac{t}{2},t]}\lambda_{t'}^{\min\{t_0,t'\}}=\begin{cases}1&\text{if }\frac{t}{2}\le t_0\\\alpha_{\frac{t}{2}}&\text{else}\end{cases}\qquad\text{(Lemma 41)}$$

Combining them and noticing that $(1+0.1\eta^2C'^2)^{-\frac{t}{2}} = \mathcal{O}(\frac{1}{t})$ when $t \geq \frac{20}{\eta^2 C'^2}$, we get that for $t \geq 2t_0 = \Theta\left(\frac{S}{(1-\gamma)^2\eta^2C'^2}\right)$,

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

Since $Y_t \leq 22S + 22S(t-1)\alpha_t \leq \mathcal{O}(\frac{S\log t}{1-\gamma})$, 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:*

$$\max_{s,x',y'}\left(V^s_{x,y'} - V^s_{x',y}\right) \leq \frac{2}{1-\gamma}\max_{s,x',y'}\left(x^sQ^s_\star y'^s - x'^sQ^s_\star y^s\right).$$

Proof Notice that for any policy x and state s,

$$\begin{aligned}
\max_{y'} V^s_{x,y'} - V^s_\star &= \sum_{a,b} x^s(a)y'^s(b)Q^s_{x,y'}(a,b) - \sum_{a,b} x^s_\star(a)y^s_\star(b)Q^s_\star(a,b)\\
&= \sum_{a,b} x^s(a)y'^s(b)\left(Q^s_{x,y'}(a,b) - Q^s_\star(a,b)\right) + \sum_{a,b}\left(x^s(a)y'^s(b) - x^s_\star(a)y^s_\star(b)\right)Q^s_\star(a,b)\\
&= \gamma\sum_{a,b} x^s(a)y'^s(b)p(s'|s,a,b)\left(V^{s'}_{x,y'} - V^{s'}_\star\right) + x^sQ^s_\star y'^s - x^s_\star Q^s_\star y^s_\star\\
&\leq \gamma\max_{s',y'}\left(V^{s'}_{x,y'} - V^{s'}_\star\right) + x^sQ^s_\star y'^s - x^s_\star Q^s_\star y^s_\star.
\end{aligned}$$

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

$$\max_{s,y'}\left(V^s_{x,y'} - V^s_\star\right) \leq \frac{1}{1-\gamma}\max_{s,y'}\left(x^sQ^s_\star y'^s - x^s_\star Q^s_\star y^s_\star\right) \leq \frac{1}{1-\gamma}\max_{s,x',y'}\left(x^sQ^s_\star y'^s - x'^sQ^s_\star y^s\right).$$

Similarly,

$$\max_{s,x'}\left(V^s_\star - V^s_{x',y}\right) \leq \frac{1}{1-\gamma}\max_{s,x',y'}\left(x^sQ^s_\star y'^s - x'^sQ^s_\star y^s\right).$$

Combining the two inequalities, we get

$$\max_{s,x',y'}\left(V^s_{x,y'} - V^s_{x',y}\right) \leq \frac{2}{1-\gamma}\max_{s,x',y'}\left(x^sQ^s_\star y'^s - x'^sQ^s_\star y^s\right).$$

■

Appendix G. Auxiliary Lemmas

G.1. Interactions between the Auxiliary Coefficients

Lemma 33 *Let $\{g_t\}_{t=1,2,\dots}$, $\{h_t\}_{t=1,2,\dots}$ be non-negative sequences that satisfy $g_t \le \gamma \sum_{\tau=1}^{t-1} \alpha_{t-1}^{\tau} g_\tau + h_t$ for all $t \ge 1$. Then $g_t \le \sum_{\tau=1}^{t} \beta_t^\tau h_\tau$.*

Proof We prove it by induction. When $t = 1$, the condition guarantees $g_1 \le h_1 = \beta_1^1 h_1$. Suppose that it holds for $1, \dots, t-1$. Then

$$\begin{aligned} g_t &\le \gamma \sum_{\tau=1}^{t-1} \alpha_{t-1}^{\tau} g_\tau + h_t \\ &\le \gamma \sum_{\tau=1}^{t-1} \alpha_{t-1}^{\tau} \left(\sum_{i=1}^{\tau} \beta_\tau^i h_i \right) + h_t \\ &= \sum_{i=1}^{t-1} \left(\sum_{\tau=i}^{t-1} \gamma \alpha_{t-1}^{\tau} \beta_\tau^i \right) h_i + h_t \end{aligned}$$

It remains to prove that $\sum_{\tau=i}^{t-1} \gamma \alpha_{t-1}^{\tau} \beta_\tau^i \le \beta_t^i$ for all $i \le t-1$. We use another induction to show this. Fix i and t, and define the partial sum $\zeta_r = \sum_{\tau=i}^{r} \gamma \alpha_{t-1}^{\tau} \beta_\tau^i$ for $r \in [i, t-1]$. Below we show that

$$\zeta_r \le \alpha_i \prod_{\tau=i}^{r} (1 - \alpha_\tau + \alpha_\tau \gamma) \prod_{\tau=r+1}^{t-1} (1 - \alpha_\tau). \tag{34}$$

Notice that the right-hand side above is β_t^i when $r = t-1$, which is exactly what we want to prove.

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

$$\begin{aligned} \zeta_{r+1} &= \zeta_r + \beta_{r+1}^i \gamma \alpha_{t-1}^{r+1} \\ &\le \alpha_i \prod_{\tau=i}^{r} (1 - \alpha_\tau + \alpha_\tau \gamma) \prod_{\tau=r+1}^{t-1} (1 - \alpha_\tau) + \left(\alpha_i \prod_{\tau=i}^{r} (1 - \alpha_\tau + \alpha_\tau \gamma) \right) \gamma \alpha_{r+1} \prod_{\tau=r+2}^{t-1} (1 - \alpha_\tau) \\ &= \alpha_i \prod_{\tau=i}^{r+1} (1 - \alpha_\tau + \alpha_\tau \gamma) \prod_{\tau=r+2}^{t-1} (1 - \alpha_\tau). \end{aligned}$$

This finishes the induction. ■

Lemma 34 *Let $\{h_t\}_{t=1,2,\dots}$ and $\{k_t\}_{t=1,2,\dots}$ be non-negative sequences that satisfy $h_t = \sum_{\tau=1}^{t} \alpha_t^\tau k_\tau$. Then $\sum_{\tau=2}^{t} \beta_t^\tau h_{\tau-1} \le \frac{1}{\gamma^2} \sum_{\tau=1}^{t-1} \beta_t^\tau k_\tau$.*

Proof By the assumption on h_τ, we have

$$\sum_{\tau=2}^{t} \beta_t^\tau h_{\tau-1} \le \sum_{\tau=2}^{t} \beta_t^\tau \left(\sum_{i=1}^{\tau-1} \alpha_{\tau-1}^i k_i \right) = \sum_{i=1}^{t-1} \left(\sum_{\tau=i+1}^{t} \beta_t^\tau \alpha_{\tau-1}^i \right) k_i$$

It remains to prove that for $i < t$, $\sum_{\tau=i+1}^{t} \beta_t^\tau \alpha_{\tau-1}^i \leq \frac{1}{\gamma^2}\beta_t^i$, or equivalently, $\gamma \sum_{\tau=i+1}^{t} \beta_t^\tau \alpha_{\tau-1}^i \leq \frac{1}{\gamma}\beta_t^i$. Below we use another induction to prove this. Fix i and t, and define the partial sum $\zeta_r = \gamma \sum_{\tau=r}^{t} \beta_t^\tau \alpha_{\tau-1}^i$ for $r \in [i+1, t]$. We will show that

$$\zeta_r \leq \alpha_i \prod_{\tau=i+1}^{r-1} (1-\alpha_\tau) \prod_{\tau=r}^{t-1} (1-\alpha_\tau+\alpha_\tau\gamma). \tag{35}$$

For the base case $r = t$, we have

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

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

$$\begin{aligned}
\zeta_{r-1} &= \zeta_r + \alpha_{r-2}^i \gamma \beta_t^{r-1} \\
&\leq \alpha_i \prod_{\tau=i+1}^{r-1} (1-\alpha_\tau) \prod_{\tau=r}^{t-1} (1-\alpha_\tau+\alpha_\tau\gamma) + \alpha_i \left(\prod_{\tau=i+1}^{r-2} (1-\alpha_\tau) \right) \gamma\alpha_{r-1} \prod_{\tau=r-1}^{t-1} (1-\alpha_\tau+\alpha_\tau\gamma) \\
&\leq \left(\alpha_i \prod_{\tau=i+1}^{r-2} (1-\alpha_\tau) \prod_{\tau=r}^{t-1} (1-\alpha_\tau+\alpha_\tau\gamma) \right) \times (1-\alpha_{r-1}+\gamma\alpha_{r-1}(1-\alpha_{r-1}+\alpha_{r-1}\gamma)) \\
&\leq \left(\alpha_i \prod_{\tau=i+1}^{r-2} (1-\alpha_\tau) \prod_{\tau=r}^{t-1} (1-\alpha_\tau+\alpha_\tau\gamma) \right) \times (1-\alpha_{r-1}+\alpha_{r-1}\gamma) \\
&= \alpha_i \prod_{\tau=i+1}^{r-2} (1-\alpha_\tau) \prod_{\tau=r-1}^{t-1} (1-\alpha_\tau+\alpha_\tau\gamma).
\end{aligned}$$

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

$$\zeta_{i+1} \leq \alpha_i \prod_{\tau=i+1}^{t-1} (1-\alpha_\tau+\alpha_\tau\gamma) = \frac{\beta_t^i}{1-\alpha_i+\alpha_i\gamma} \leq \frac{\beta_t^i}{\gamma}$$

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

Lemma 35 *For $0 \leq h \leq t$, $\sum_{\tau=0}^{h} \alpha_t^\tau = \delta_t^h$.*

Proof We prove it by induction on h. When $h = 0$, $\sum_{\tau=0}^{h} \alpha_t^\tau = \alpha_t^0 = \prod_{\tau=1}^{t}(1-\alpha_\tau) = \delta_t^h$ since $\alpha_0 = 1$. Suppose that the formula holds for h. Then $\sum_{\tau=0}^{h+1} \alpha_t^\tau = \sum_{\tau=0}^{h} \alpha_t^\tau + \alpha_t^{h+1} = \prod_{\tau=h+1}^{t}(1-\alpha_\tau) + \alpha_{h+1}\prod_{\tau=h+2}^{t}(1-\alpha_\tau) = \prod_{\tau=h+2}^{t}(1-\alpha_\tau) = \delta_t^{h+1}$, which finishes the induction. ■

Lemma 36 *For any positive integers i, t with $i \leq t$, $\sum_{\tau=i}^{t} \lambda_t^\tau \beta_\tau^i \leq \frac{3}{1-\gamma}\lambda_t^i$.*

Proof Notice that

$$\sum_{\tau=i}^{t} \lambda_t^\tau \beta_\tau^i = \lambda_t^i + \sum_{\tau=i+1}^{t} \lambda_t^\tau \beta_\tau^i. \tag{36}$$

Below we use induction to prove

$$\sum_{\tau=r}^{t} \lambda_t^\tau \beta_\tau^i \le \frac{2}{1-\gamma}\alpha_i \prod_{\tau=i}^{r-1}\Big(1-\alpha_\tau(1-\gamma)\Big)\prod_{\tau=r}^{t-1}\lambda_\tau$$

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

$$\begin{aligned}
\sum_{\tau=r-1}^{t} \lambda_t^\tau \beta_\tau^i &\le \frac{2}{1-\gamma}\alpha_i \prod_{\tau=i}^{r-1}\Big(1-\alpha_\tau(1-\gamma)\Big)\prod_{\tau=r}^{t-1}\lambda_\tau + \beta_{r-1}^i \lambda_t^{r-1} \\
&\le \frac{2}{1-\gamma}\alpha_i \prod_{\tau=i}^{r-1}\Big(1-\alpha_\tau(1-\gamma)\Big)\prod_{\tau=r}^{t-1}\lambda_\tau + \left(\alpha_i \prod_{\tau=i}^{r-2}\Big(1-\alpha_\tau(1-\gamma)\Big)\right)\alpha_{r-1}\prod_{\tau=r-1}^{t-1}\lambda_\tau \\
&\le \left[\alpha_i \prod_{\tau=i}^{r-2}\Big(1-\alpha_\tau(1-\gamma)\Big)\prod_{\tau=r}^{t-1}\lambda_\tau\right]\left(\frac{2}{1-\gamma}\left(1-\alpha_{r-1}(1-\gamma)\right)+\alpha_{r-1}\right) \\
&\qquad\qquad (\lambda_{r-1} \le 1) \\
&= \left[\alpha_i \prod_{\tau=i}^{r-2}\Big(1-\alpha_\tau(1-\gamma)\Big)\prod_{\tau=r}^{t-1}\lambda_\tau\right]\left(\frac{2}{1-\gamma}\left(1-\frac{1}{2}\alpha_{r-1}(1-\gamma)\right)\right) \\
&\le \frac{2}{1-\gamma}\alpha_i \prod_{\tau=i}^{r-2}\Big(1-\alpha_\tau(1-\gamma)\Big)\prod_{\tau=r-1}^{t-1}\lambda_\tau, \\
&\qquad\qquad (\lambda_{r-1} \ge 1-\tfrac{1}{2}\alpha_{r-1}(1-\gamma) \text{ by the definition of } \lambda_{r-1})
\end{aligned}$$

which finishes the induction. Notice that this implies

$$\sum_{\tau=i+1}^{t} \lambda_t^\tau \beta_\tau^i \le \frac{2}{1-\gamma}\alpha_i\left(1-\alpha_i(1-\gamma)\right)\prod_{\tau=i+1}^{t-1}\lambda_\tau \le \frac{2}{1-\gamma}\alpha_i\prod_{\tau=i}^{t-1}\lambda_\tau = \frac{2}{1-\gamma}\lambda_t^i$$

where the second inequality is by the definition of λ_i. Thus,

$$\sum_{\tau=i}^{t} \lambda_t^\tau \beta_\tau^i \le \frac{3}{1-\gamma}\lambda_t^i. \tag{37}$$

■

Lemma 37 $\lambda_{t+1}^{\tau+1} \le \lambda_t^\tau$.

Proof When $\tau < t$, we have

$$\frac{\lambda_{t+1}^{\tau+1}}{\lambda_t^{\tau}} = \frac{\alpha_{\tau+1}\Pi_{i=\tau+1}^{t}\lambda_i}{\alpha_\tau \Pi_{i=\tau}^{t-1}\lambda_i} \leq \frac{\alpha_{\tau+1}}{\alpha_\tau\lambda_\tau} \leq 1$$

where in the first inequality we use $\lambda_t \leq 1$ and in the second inequality we use the definition of λ_τ. When $\tau = t$, we have $\frac{\lambda_{t+1}^{\tau+1}}{\lambda_t^{\tau}} = \frac{1}{1} = 1$. ■

Lemma 38 $\sum_{\tau=1}^{t}\beta_t^\tau \leq \frac{2}{1-\gamma}$.

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

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

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

Suppose that this holds for r, then

$$\begin{aligned}
\sum_{\tau=1}^{r+1}\beta_t^\tau &= \beta_t^{r+1} + \sum_{\tau=1}^{r}\beta_t^\tau \\
&\leq \alpha_{r+1}\prod_{i=r+1}^{t-1}(1-\alpha_i+\alpha_i\gamma) + \frac{1}{1-\gamma}\prod_{i=r}^{t-1}(1-\alpha_i+\alpha_i\gamma) \\
&\leq \left(\frac{1}{1-\gamma}\prod_{i=r+1}^{t-1}(1-\alpha_i+\alpha_i\gamma)\right)(\alpha_{r+1}(1-\gamma)+1-\alpha_r(1-\gamma)) \\
&\leq \frac{1}{1-\gamma}\prod_{i=r+1}^{t-1}(1-\alpha_i+\alpha_i\gamma) \qquad \text{(because } \alpha_{r+1}\leq\alpha_r\text{)}
\end{aligned}$$

which finishes the induction.

Therefore, $\sum_{\tau=1}^{t}\beta_t^\tau = 1 + \sum_{\tau=1}^{t-1}\beta_t^\tau \leq 1 + \frac{1}{1-\gamma} \leq \frac{2}{1-\gamma}$. ■

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*

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

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

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

which finishes the induction. Therefore,

$$\begin{aligned} g_t &\le g_0(1-c)^t + \sum_{\tau=1}^{t/2}(1-c)^{t-\tau}\max_{\tau\in[1,t/2]} h_\tau + \sum_{\tau=t/2+1}^{t}(1-c)^{t-\tau}\max_{\tau\in[t/2,t]} h_\tau \\ &\le g_0(1-c)^t + \frac{\max_{\tau\in[1,t/2]} h_\tau}{c}(1-c)^{\frac{t}{2}} + \frac{\max_{\tau\in[t/2,t]} h_\tau}{c}. \end{aligned}$$

■

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 $H \ge \frac{2}{1-\gamma}$, we have $\sum_{t=\tau}^{\infty} \beta_t^\tau \le H + 3$.*

Proof When $t \ge \tau + 2$,

$$\begin{aligned} \beta_t^\tau &= \alpha_\tau \prod_{i=\tau}^{t-1}(1-\alpha_i(1-\gamma)) \\ &\le \alpha_\tau \prod_{i=\tau}^{t-1}\left(1-\alpha_i \times \frac{2}{H+1}\right) && (H+1 \ge \tfrac{2}{1-\gamma}) \\ &= \alpha_\tau \prod_{i=\tau}^{t-1}\left(1-\frac{2}{H+i}\right) \\ &= \frac{H+1}{H+\tau} \times \frac{H+\tau-2}{H+\tau} \times \frac{H+\tau-1}{H+\tau+1} \times \cdots \times \frac{H+t-3}{H+t-1} \\ &= \frac{H+1}{H+\tau} \times \frac{(H+\tau-2)(H+\tau-1)}{(H+t-2)(H+t-1)} \\ &= \frac{H+1}{H+\tau}(H+\tau-2)(H+\tau-1)\left(\frac{1}{H+t-2} - \frac{1}{H+t-1}\right) \\ &\le (H+1)(H+\tau-2)\left(\frac{1}{H+t-2} - \frac{1}{H+t-1}\right). \end{aligned}$$

Therefore,

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

and thus $\sum_{t=\tau}^{\infty} \beta_t^\tau \le H + 3$. ■

Lemma 41 *For the choice $\alpha_t = \frac{H+1}{H+t}$ with $H \ge \frac{2}{1-\gamma}$, we have $\lambda_t^\tau = \alpha_t$ for $\tau < t$.*

Proof With this choice of α_t,

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

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

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

■

Appendix H. Analysis on Sample Complexity

H.1. Proof of Theorem 3

Proof As long as we can make

$$\left|\ell_t^s(a) - \mathbf{e}_a^\top Q_t^s \widetilde{y}_t^s\right| \le \frac{\varepsilon}{2|\mathcal{A}|} \tag{38}$$

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

$$\begin{aligned}&\text{Prob}\left[\left|\ell_t^s(a) - \mathbf{e}_a^\top Q_t^s \widetilde{y}_t^s\right| \le \mathcal{O}\left(\sqrt{\frac{\log(L/\delta)}{N_{s,a}}}\right)\right] \\ &\le \text{Prob}\left[\exists m \in [L], \left|\widetilde{\ell}_{t,m}^s - \mathbf{e}_a^\top Q_t^s \widetilde{y}_t^s\right| \le \mathcal{O}\left(\sqrt{\frac{\log(L/\delta)}{m}}\right)\right] \\ &\le \sum_{m=1}^L \text{Prob}\left[\left|\widetilde{\ell}_{t,m}^s - \mathbf{e}_a^\top Q_t^s \widetilde{y}_t^s\right| \le \mathcal{O}\left(\sqrt{\frac{\log(L/\delta)}{m}}\right)\right] \le \delta.\end{aligned}$$

Therefore, with probability at least $1-\delta$, Eq. (38) holds if

$$N_{s,a} = \Omega\left(\frac{|\mathcal{A}|^2}{\varepsilon^2}\log(L/\delta)\right). \tag{39}$$

Now it remains to determine L to make Eq. (39) hold with high probability. Note that by Assumption 1, we know that $T^{s'\to s}_{\widetilde{x}_t,\widetilde{y}_t} \le \frac{1}{\mu}$ for any s'. Let $\mathcal{T}_{s,a}$ 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 $\widetilde{x}^s_t$ and $\widetilde{y}^s_t$. The mean of this distribution is $t_{s,a} \le 1 + \frac{2|\mathcal{A}|}{\varepsilon'\mu} \le \frac{3|\mathcal{A}|}{\varepsilon'\mu}$. Then by Markov inequality,

$$\text{Prob}\left[\text{the number of rounds before reaching } (s,a) \le \frac{6|\mathcal{A}|}{\varepsilon'\mu}\right] \ge \frac{1}{2}.$$

Therefore, with probability at least $1-\frac{\delta}{L}$, within $\Theta(\frac{|A|}{\varepsilon'\mu}\log(L/\delta))$ rounds, we reach (s,a) state-action at least pair once. Thus, Eq. (39) holds when $L = \Omega\left(\frac{|\mathcal{A}|^3}{\varepsilon'\mu\varepsilon^2}\log^2(L/\delta)\right)$ with probability $1-\delta$. Solving L gives $L = \widetilde{\Omega}\left(\frac{|\mathcal{A}|^3}{(1-\gamma)\mu\varepsilon^3}\log^2(1/\delta)\right)$. The cases for $r^s_t(b)$ and ρ^s_t are similar. Finally, using a union bound on all $\mathcal{A}$, $\mathcal{B}$, $\mathcal{S}$, and all iterations, we know that with probability $1-\delta$, the ε-approximations are always guaranteed if we use the estimation above and take $L = \widetilde{\Omega}\left(\frac{|\mathcal{A}|^3+|\mathcal{B}|^3}{(1-\gamma)\mu\varepsilon^3}\log^2(T/\delta)\right)$. ■

H.2. Proof of Corollary 4

Proof From Theorem 1, we know that in order to show $\frac{1}{T}\sum_{t=1}^T \max_{s,x',y'}\left(V^s_{\widehat{x}_t,y'} - V^s_{x',\widehat{y}_t}\right) \le \xi$, it is sufficient to show $\frac{|\mathcal{S}|}{\eta(1-\gamma)^2}\sqrt{\frac{\log T}{T}} \le \xi$ and $\frac{|\mathcal{S}|\sqrt{\varepsilon}}{\sqrt{\eta}(1-\gamma)^2} \le \xi$. Solving these two inequalities, we get $T = \Omega\left(\frac{|\mathcal{S}|^2}{\eta^2(1-\gamma)^4\xi^2}\log\frac{|\mathcal{S}|}{\eta(1-\gamma)\xi}\right)$ and $\varepsilon = \mathcal{O}\left(\frac{\eta(1-\gamma)^4\xi^2}{|\mathcal{S}|^2}\right)$. Plugging ε 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 $\frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\text{dist}^2_\star(\widehat{z}^s_T) \le \xi$, it is sufficient to show $\frac{|\mathcal{S}|^2}{\eta^4C^4(1-\gamma)^4T} \le \xi$ and $\frac{\varepsilon}{\eta C^2(1-\gamma)^3} \le \xi$. Solving these two inequalities, we get $T = \Omega\left(\frac{|\mathcal{S}|^2}{\eta^4C^4(1-\gamma)^4\xi}\right)$ and $\varepsilon = \mathcal{O}\left(\eta C^2(1-\gamma)^3\xi\right)$. Plugging ε into L in Theorem 3 gives the required L. ■

Appendix I. Analysis on Rationalily

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: $\gamma \in [\frac{1}{2}, 1), \eta \le \frac{1}{10^4}\sqrt{\frac{(1-\gamma)^5}{S}}, \varepsilon \in \left[0, \frac{1}{1-\gamma}\right]$
Parameters: a non-increasing sequence $\{\alpha_t\}_{t=1}^{T}$ that goes to zero.
Initialization: arbitrarily initialize $\widehat{x}_1^s = x_1^s \in \Delta_{\mathcal{A}}$ for all $s \in \mathcal{S}$.
$V_0^s \leftarrow 0$ for all $s \in \mathcal{S}$.
for $t = 1, \ldots, T$ do
    For all $s$, define $Q_t^s \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{B}|}$ 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
        $\widehat{x}_{t+1}^s = \Pi_{\Delta_{\mathcal{A}}}\left\{\widehat{x}_t^s - \eta \ell_t^s\right\},$    (40)
        $x_{t+1}^s = \Pi_{\Delta_{\mathcal{A}}}\left\{\widehat{x}_{t+1}^s - \eta \ell_t^s\right\},$    (41)
        $V_t^s = (1-\alpha_t)V_{t-1}^s + \alpha_t \rho_t^s,$    (42)
    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.
end
```

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 $\underline{\sigma}(s,a,1) = \mathbb{E}_{b\sim y^s}[\sigma(s,a,b)]$, and the transition kernel is redefined as $\underline{p}(s'|s,a,1) = \mathbb{E}_{b\sim y^s}[p(s'|s,a,b)]$. Correspondingly, we define

$$
\begin{aligned}
\underline{Q}_t^s(a,1) &= \underline{\sigma}(s,a,1) + \gamma \mathbb{E}_{s'\sim \underline{p}(\cdot|s,a,1)}\left[\underline{V}_{t-1}^{s'}\right], \\
\underline{\widehat{x}}_{t+1}^s &= \Pi_{\Delta_{\mathcal{A}}}\left\{\underline{\widehat{x}}_t^s - \eta \ell_t^s\right\}, \\
\underline{x}_{t+1}^s &= \Pi_{\Delta_{\mathcal{A}}}\left\{\underline{\widehat{x}}_{t+1}^s - \eta \ell_t^s\right\}, \\
\underline{V}_t^s &= (1-\alpha_t)\underline{V}_{t-1}^s + \alpha_t \rho_t^s,
\end{aligned}
$$

where $\underline{V}_0^s = 0$ for all s, and ℓ_t^s and ρ_t^s are the same as in Algorithm 2. Clearly, the sequences $\{\widehat{x}_t, x_t\}_{t\in[T]}$ and $\{\underline{\widehat{x}}_t, \underline{x}_t\}_{t\in[T]}$ are exactly the same (assuming their initializations are the same, that is, $\widehat{x}_1 = \underline{\widehat{x}}_1$ and $x_1 = \underline{x}_1$). In the following lemma, we show that ℓ_t^s and ρ_t^s are indeed ε-approximation of $\underline{Q}_t^s(\cdot,1)$ and $\underline{x}_t^{s\top}\underline{Q}_t^s(a,\cdot)$, which then implies that the sequence $\{\underline{\widehat{x}}_t\}_{t\in[T]}$ 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 $\mathcal{X}_\star^s$ 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, ℓ_t^s and ρ_t^s are ε-approximation of $\underline{Q}_t^s(\cdot,1)$ and $\underline{x}_t^{s\top}\underline{Q}_t^s(a,\cdot)$ respectively.*

Proof We prove the result together with $\underline{V}_t^s = V_t^s$, $\underline{Q}_t^s(\cdot, 1) = Q_t^s y^s$ for all $t \in [T]$ by induction. When $t = 1$, $\underline{V}_t^s = V_t^s$ clearly holds. In addition, $Q_1^s(a, \cdot)y^s = \mathbb{E}_{b\sim y^s}[\sigma(s, a, b)] = \underline{Q}_1^s(a, 1)$. Therefore ℓ_1^s and ρ_1^s are indeed ε-approximation of $\underline{Q}_1^s(a, \cdot)$ and $\underline{x}_1^{s\top} \underline{Q}_t^s(a, \cdot)$.

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

$$\begin{aligned}\underline{Q}_{t+1}^s(a, 1) &= \underline{\sigma}(s, a, 1) + \gamma \mathbb{E}_{s'\sim \underline{p}(\cdot|s,a,1)}\left[\underline{V}_t^{s'}\right] \\ &= \mathbb{E}_{b\sim y^s}\left[\sigma(s, a, b) + \gamma \mathbb{E}_{s'\sim p(s'|s,a,b)}\left[V_t^{s'}\right]\right] \\ &= Q_{t+1}^s(a, \cdot)y^s,\end{aligned}$$

which also shows that ℓ_{t+1}^s and ρ_{t+1}^s are indeed ε-approximation of $\underline{Q}_{t+1}^s(\cdot, 1)$ and $\underline{x}_{t+1}^{s\top}\underline{Q}_t^s(a, \cdot)$ (recall $\underline{x}_{t+1}^s = x_{t+1}^s$). By definition of $\underline{V}_{t+1}^s$, we also have $\underline{V}_{t+1}^s = (1 - \alpha_{t+1})\underline{V}_t^s + \alpha_t\rho_t^s = (1 - \alpha_t)V_t^s + \alpha_t\rho_t^s = V_{t+1}^s$, which finishes the induction. ■