SAMPLE EFFICIENT STOCHASTIC POLICY EXTRAGRADIENT ALGORITHM FOR ZERO-SUM MARKOV GAME

Anonymous authors
Paper under double-blind review

ABSTRACT

Two-player zero-sum Markov game is a fundamental problem in reinforcement learning and game theory. Although many algorithms have been proposed for solving zero-sum Markov games in the existing literature, they generally lack the desired and important features such as model-free, provably convergent, sample efficient, symmetric and private policy updates, etc. In this paper, we develop a fully decentralized stochastic policy extragradient algorithm with all these properties for solving zero-sum Markov games. In particular, our algorithm introduces multiple stochastic estimators to accurately estimate the value functions involved in the stochastic updates, and leverages entropy regularization to accelerate the convergence. Specifically, with a proper entropy-regularization parameter, we prove that the stochastic policy extragradient algorithm has a sample complexity of the order $\mathcal{O}(\frac{t_{\text{mix}} A_{\max}}{\mu_{\min}\epsilon^{5.5}(1-\gamma)^{13.5}})$ for finding a solution that achieves ϵ-Nash equilibrium duality gap. Such a sample complexity result substantially improves the state-of-the-art complexity results.

1 INTRODUCTION

Competitive reinforcement learning (RL) is an emerging and popular framework that has broad applications in various areas, including market pricing applications (Kononen and Oja, 2004), real-time strategy-making (Vinyals et al., 2019), board games (Silver et al., 2017; Moerland et al., 2018) and inverse RL (Zhang et al., 2019). In particular, an important and fundamental formulation of competitive RL is the two-player zero-sum Markov game, which involves two competing players that interact with a common environment and receive zero-sum rewards. Both players aim to learn the optimal policy that achieves the Nash equilibrium of accumulated rewards.

Algorithms for solving Markov games are very different from conventional single-agent RL algorithms. In particular, both players must learn to improve their policies based on feedback from the opponent and the environment, but usually the opponent will not reveal any sensitive information (e.g., actions or policy) or cooperate with each other. In the existing literature, numerous algorithms have been developed for solving zero-sum Markov games, including Q-learning (Fan et al., 2020; Zhu and Zhao, 2020), fitted Q iteration (Zhang et al., 2021), policy gradient (Daskalakis et al., 2020; Zhao et al., 2021), policy extragradient (Cen et al., 2021), model-based Monte Carlo estimation (Zhang et al., 2020), optimistic gradient descent ascent (Wei et al., 2021), etc. However, many of these algorithms require both players to access their opponent's actions (Wei et al., 2017; Sidford et al., 2020; Bai et al., 2020; Huang et al., 2021; Jafarnia-Jahromi et al., 2021), which causes privacy issues. On the other hand, some algorithms need to know about the environment transition kernel and reward mapping (Cen et al., 2021), which are usually unknown a priori in practice. Moreover, other algorithms require both players to perform asymmetric policy updates using different numbers of iterations, learning rates or exploration probabilities (Zhao et al., 2021; Daskalakis et al., 2020), which are generally hard to coordinate in advance between two competing players. Therefore, it is desired to develop an algorithm for solving Markov games that avoids all the aforementioned issues, and we want to ask the following question.

- **Q1:** *Can we develop a fully decentralized algorithm that is model-free and takes symmetric and private policy updates for solving zero-sum Markov games?*

Table 1: Summary of key properties and sample complexity of algorithms for solving discounted infinite-horizon zero-sum Markov games.

Work	Model -free	Private update	Symmetric update	Data type	Sample complexity (duality gap $\leq \epsilon$)
Zou et al. (2019)	✓	×	✓	Markovian	–
Zhao et al. (2021)	✓	✓	×	i.i.d.	–
Guo et al. (2021)	✓	✓	×	i.i.d.	–
Cen et al. (2021)	×	✓	✓	–	–
Wei et al. (2021)	✓	✓	✓	Markovian	$\widetilde{O}\big(\frac{A_{\max}^3 \lvert\mathcal{S}\rvert^{10.5}}{\epsilon^8(1-\gamma)^{29.5}}\big)$
Our work	✓	✓	✓	**Markovian**	$\widetilde{\mathcal{O}}\big(\frac{A_{\max}}{\epsilon^{5.5}(1-\gamma)^{13.5}}\big)$

Moreover, from a theoretical perspective, the convergence and sample complexity of these existing algorithms for solving Markov games have not been comprehensively studied. Specifically, some studies established the convergence of the algorithms with i.i.d. samples (Zhao et al., 2021; Guo et al., 2021), which violates the dependent nature of samples collected from the dynamic Markov decision process. Also, other algorithms suffer from an extremely high sample complexity to achieve an approximate Nash equilibrium solution (Wei et al., 2021). Hence, we are motivated to answer the following fundamental question.

- **Q2:** *Can we develop the algorithm mentioned in Q1 with provable convergence guarantee and an improved sample complexity for achieving a Nash equilibrium solution?*

In this paper, we provide positive and comprehensive answers to both questions by developing a fully decentralized stochastic policy extragradient algorithm. In Table 1, we compare the key properties and sample complexity of our algorithm with those of the existing algorithms. We also refer to Appendix F for more explanations on Table 1. Our contributions are summarized as follows.

1.1 Our Contributions

We consider a standard zero-sum Markov game with discounted reward over infinite horizon. To solve such a Markov game, we propose a stochastic variant of the policy extragradient algorithm (Cen et al., 2021) that satisfies the following amenable properties.

- Our algorithm uses multiple new stochastic estimators to estimate the value functions involved in the predictive updates for solving entropy-regularized matrix games, and therefore the algorithm does not rely on any prior knowledge of the environment transition kernel (model-free). Moreover, the resulting stochastic policy updates of our algorithm for both players are symmetric and do not involve any sensitive information of the opponent (private).

- Compared with the stochastic estimators used in (Wei et al., 2021), our estimators have much smaller variance that helps improve the estimation accuracy. Specifically, by developing new techniques (explained in the next bullet), we establish tight high-probability estimation error bounds for our stochastic estimators with Markovian samples. Then, with a proper entropy-regularization parameter, we show that our stochastic policy extragradient algorithm requires a sample complexity of the order $\mathcal{O}(\frac{t_{\text{mix}} A_{\max}}{\mu_{\min}\epsilon^{5.5}(1-\gamma)^{13.5}})$ to achieve an ϵ-Nash equilibrium duality gap, which substantially improves the state-of-the-art complexity result of (Wei et al., 2021).

- We develop novel techniques to bound the estimation error of the proposed stochastic estimators, whose numerator and denominator involve sample average approximations. First, we propose a special estimation error decomposition that avoids divergence of the bound caused by possibly small numerical values of the sample average involved in the denominator of the stochastic estimators. Second, we leverage this error decomposition and the recursive structure of the stochastic estimators to derive a contraction property of the estimation errors, which eventually leads to tight bounds for the estimation error. We refer to Section 4 for more elaboration on our technical novelties.

1.2 OTHER RELATED WORK

Other settings of two-player zero-sum Markov games. In this paper, we focus on a standard setting of two-player zero-sum Markov game with discount and infinite horizon in the discrete time domain. There are other settings of two-player zero-sum Markov games. For example, Bai et al. (2020); Huang et al. (2021) studied a two-player zero-sum Markov game with finite horizon and without discount, whereas Daskalakis et al. (2020) considered finite random horizon without discount. Jafarnia-Jahromi et al. (2021) also considered the setting without discount, and it allows one of the players to constantly adjust its policy based on the entire history of states and actions. Ghosh et al. (2021) studied a two-player zero-sum Markov game in the continuous time domain.

Multi-agent general-sum Markov game. Some works studied multi-agent Markov games, which extend the two-player zero-sum Markov games to multiple players without the zero-sum constraint (Wang and Sandholm, 2002; Hu and Wellman, 2003; Deng et al., 2021; Leonardos et al., 2021). More specifically, Leonardos et al. (2021) defined and studied Markov potential game which has a potential function assumption. Guo et al. (2019); Elie et al. (2020); Gu et al. (2021) studied mean-field games with a large number of players.

Entropy regularization and value iteration Our algorithm leverages entropy regularization and value iteration to accelerate convergence. Entropy regularization is a popular technique that has been widely used in reinforcement learning (Neu et al., 2017; Geist et al., 2019; Mei et al., 2020; Cen et al., 2020) and Markov game (Mertikopoulos and Sandholm, 2016; Savas et al., 2019; Cen et al., 2021) to encourage environment exploration and accelerate algorithm convergence. Value iteration is also a classical method that is widely used in both single-agent reinforcement learning (Ernst et al., 2005; Tamar et al., 2016; Farahmand and Ghavamzadeh, 2021) and Markov games (Zhu and Zhao, 2020; Cen et al., 2021). With full knowledge of the environment, it exponentially converges to the fixed point of Bellman operator (Cen et al., 2021). Compared to another similar classical method called policy iteration, value iteration does not need policy evaluation which involves additional computation (Sutton and Barto, 2018).

2 BACKGROUND OF MARKOV GAME AND ENTROPY REGULARIZATION

2.1 TWO-PLAYER ZERO-SUM MARKOV GAME

In a zero-sum Markov game, two players compete with each other in a common environment. Throughout, the state space is denoted as $\mathcal{S}$. The action spaces and policies of both players are denoted as $\mathcal{A}^{(1)}, \pi^{(1)}$ and $\mathcal{A}^{(2)}, \pi^{(2)}$, respectively. Here, $\pi^{(1)} \in \Delta(|\mathcal{A}^{(1)}|), \pi^{(2)} \in \Delta(|\mathcal{A}^{(2)}|)$ are random policies defined over the corresponding simplex sets. The reward function is denoted as $R : \mathcal{S} \times \mathcal{A}^{(1)} \times \mathcal{A}^{(2)} \to [0, 1]$, and the discount factor is denoted as $\gamma \in (0, 1)$.

At any time t, both players observe state $s_t \in \mathcal{S}$ of the environment. Then, both players respectively select their actions following their own policies, i.e., $a_t^{(1)} \sim \pi^{(1)}(\cdot|s_t)$ and $a_t^{(2)} \sim \pi^{(2)}(\cdot|s_t)$. After that, the environment state transfers to a new state s_{t+1} following the underlying transition kernel $\mathcal{P}(\cdot|s_t, a_t^{(1)}, a_t^{(2)})$, and both players receive zero-sum rewards, i.e., $R_t^{(1)} = -R_t^{(2)} = R_t$, where $R_t := R(s_t, a_t^{(1)}, a_t^{(2)})$. With this Markov decision process, we can define the following state value function associated with the players' policies $\pi^{(1)}$ and $\pi^{(2)}$ for any environment state $s \in \mathcal{S}$.

$$V_{\pi^{(1)},\pi^{(2)}}(s) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t R_t \Big| s_0 = s\Big]. \tag{1}$$

The goal of both players is to compete via the following minimax game in all states s.

$$\min_{\pi^{(2)}} \max_{\pi^{(1)}} \; V_{\pi^{(1)},\pi^{(2)}}(s). \tag{2}$$

In particular, it has been shown in (Shapley, 1953) that there exists a Nash equilibrium policy pair $\pi_*^{(1)}, \pi_*^{(2)}$ for zero-sum Markov games, i.e., $V_{\pi^{(1)},\pi_*^{(2)}}(s) \le V_{\pi_*^{(1)},\pi_*^{(2)}}(s) \le V_{\pi_*^{(1)},\pi^{(2)}}(s)$ holds for any other policies $\pi^{(1)}, \pi^{(2)}$ and for all states s.

2.2 ENTROPY-REGULARIZED MARKOV GAME

Entropy regularization is a popular technique that has been widely used in reinforcement learning (Neu et al., 2017; Geist et al., 2019; Mei et al., 2020; Cen et al., 2020) and Markov game (Mertikopoulos and Sandholm, 2016; Savas et al., 2019; Cen et al., 2021) to encourage environment exploration and accelerate algorithm convergence.

Specifically, for the zero-sum Markov game, we can define an entropy-regularized state value function by adding entropy regularization to the state value function in (1) as follows (Cen et al., 2021).

$$V^{(\tau)}_{\pi^{(1)},\pi^{(2)}}(s) := \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t\big[R_t - \tau\ln\pi^{(1)}(a_t^{(1)}|s_t) + \tau\ln\pi^{(2)}(a_t^{(2)}|s_t)\big]\Big|s_0 = s\Big], \tag{3}$$

where $\tau > 0$ is called the regularization parameter. With the above definition, we further define the following entropy-regularized state-action value function (also called Q-function) (Cen et al., 2021).

$$Q^{(\tau)}_{\pi^{(1)},\pi^{(2)}}(s, a^{(1)}, a^{(2)}) := R(s, a^{(1)}, a^{(2)}) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a^{(1)},a^{(2)})}\big[V^{(\tau)}_{\pi^{(1)},\pi^{(2)}}(s')\big]. \tag{4}$$

In particular, $V^{(\tau)}_{\pi^{(1)},\pi^{(2)}}$ can be obtained from $Q^{(\tau)}_{\pi^{(1)},\pi^{(2)}}$ as follows.

$$\begin{aligned} V^{(\tau)}_{\pi^{(1)},\pi^{(2)}}(s) =& [\pi^{(1)}(s)]^\top Q^{(\tau)}_{\pi^{(1)},\pi^{(2)}}(s)\pi^{(2)}(s) + \tau\mathcal{H}\big(\pi^{(1)}(s)\big) - \tau\mathcal{H}\big(\pi^{(2)}(s)\big) \\ :=& f_\tau\big(Q^{(\tau)}_{\pi^{(1)},\pi^{(2)}}(s), \pi^{(1)}(s), \pi^{(2)}(s)\big), \end{aligned} \tag{5}$$

where $\mathcal{H}(\pi)$ denotes the entropy of policy π, and we define this mapping as f_τ for convenience.

For the entropy regularized Markov game, it has an equilibrium policy pair that solves the minimax optimization problem $\min_{\pi^{(2)}}\max_{\pi^{(1)}} V^{(\tau)}_{\pi^{(1)},\pi^{(2)}}(s)$. Such a policy pair is called the quantal response equilibrium (QRE). Our goal is to find the equilibrium policy pair of the original Markov game in (2) by solving the entropy-regularized Markov game with a proper regularization parameter τ. In particular, compared with the equilibrium policy of the Markov game, the QRE tends to have a larger entropy due to the entropy regularization, which encourages the players to explore and obtain a better understanding of the environment. Another advantage of considering the entropy-regularized Markov game is that the entropy regularization makes the minimax problem have a better optimization geometry that accelerates the convergence of the optimization process.

3 STOCHASTIC POLICY EXTRAGRADIENT ALGORITHM FOR ENTROPY-REGULARIZED MARKOV GAME

In this section, we develop a stochastic policy extragradient (SPE) algorithm for solving entropy-regularized Markov games. First, we recap the policy extragradient (PE) algorithm, which is introduced in (Cen et al., 2021) to solve entropy-regularized Markov games with full knowledge of the environment transition kernel and reward mapping. Then, we propose the model-free SPE algorithm that solves entropy-regularized Markov games using only stochastic samples.

3.1 REVIEW OF POLICY EXTRAGRADIENT ALGORITHM

Value iteration is a classical reinforcement learning algorithm that requires full knowledge of the environment and achieves an exponential convergence rate. In particular, for the entropy-regularized Markov game, the k-th value iteration update is defined as follows.

$$Q_k(s, a^{(1)}, a^{(2)}) = R(s, a^{(1)}, a^{(2)}) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a^{(1)},a^{(2)})}\big[V_k(s')\big], \quad \forall s, a^{(1)}, a^{(2)}, \tag{6}$$

$$V_{k+1}(s) = \min_{\pi^{(2)}(s)}\max_{\pi^{(1)}(s)} f_\tau\big(Q_k(s); \pi^{(1)}(s), \pi^{(2)}(s)\big), \quad \forall s, \tag{7}$$

where we define $Q_k(s) := Q_k(s, \cdot, \cdot) \in \mathbb{R}^{|\mathcal{A}^{(1)}|\times|\mathcal{A}^{(2)}|}$. This algorithm alternatively updates all the entries of the value functions Q_k and V_k. Thanks to the entropy regularization in the function f_τ (see (5) for the definition), the minimax matrix game in (7) is τ-strongly concave in $\pi^{(1)}$ and τ-strongly convex in $\pi^{(2)}$, and therefore it has a unique solution.

To solve the entropy-regularized minimax matrix game in (7), Cen et al. (2021) proposed a predictive update (PU) algorithm. Specifically, with uniform policy initialization, i.e., $\pi_{k,0}^{(m)}(s) = \frac{\mathbf{1}}{|\mathcal{A}^{(m)}|}, \forall m \in \{1,2\}, \forall s \in \mathcal{S}$, the PU algorithm performs the following policy updates: for $t = 0, 1, 2, \ldots$

$$\text{(PU):} \quad \begin{cases} \overline{\pi}_{k,t+1}^{(1)}(a^{(1)}|s) \propto \pi_{k,t}^{(1)}(a^{(1)}|s)^{1-\eta\tau} \exp\big(\eta Q_{k,t}^{(1)}(s,a^{(1)})\big) \\ \overline{\pi}_{k,t+1}^{(2)}(a^{(2)}|s) \propto \pi_{k,t}^{(2)}(a^{(2)}|s)^{1-\eta\tau} \exp\big(-\eta Q_{k,t}^{(2)}(s,a^{(2)})\big) \\ \pi_{k,t+1}^{(1)}(a^{(1)}|s) \propto \pi_{k,t}^{(1)}(a^{(1)}|s)^{1-\eta\tau} \exp\big(\eta \overline{Q}_{k,t+1}^{(1)}(s,a^{(1)})\big) \\ \pi_{k,t+1}^{(2)}(a^{(2)}|s) \propto \pi_{k,t}^{(2)}(a^{(2)}|s)^{1-\eta\tau} \exp\big(-\eta \overline{Q}_{k,t+1}^{(2)}(s,a^{(2)})\big) \end{cases}, \tag{8}$$

where we use the following notations (superscript $(\backslash m)$ denotes the opponent of the m-th player.).

$$Q_{k,t}^{(m)}(s,a^{(m)}) := \mathbb{E}_{a^{(\backslash m)} \sim \pi_{k,t}^{(\backslash m)}(s)}\big[Q_k(s,a^{(1)},a^{(2)})\big], \quad m \in \{1,2\} \tag{9}$$

$$\overline{Q}_{k,t+1}^{(m)}(s,a^{(m)}) := \mathbb{E}_{a^{(\backslash m)} \sim \overline{\pi}_{k,t+1}^{(\backslash m)}(s)}\big[Q_k(s,a^{(1)},a^{(2)})\big], \quad m \in \{1,2\}. \tag{10}$$

Once we obtain the output policy pair $(\pi_k^{(1)}, \pi_k^{(2)})$ of the PU algorithm, we can obtain an approximation of $V_{k+1}(s)$ as $V'_{k+1}(s) = f_\tau\big(Q_k(s); \pi_k^{(1)}(s), \pi_k^{(2)}(s)\big)$, which will be further used in the next Q-value function update (6) to replace $V_{k+1}(s')$. The updates (6), (7) & (8) are referred to as policy extragradient (PE) algorithm.

In the PE algorithm, the PU update in (8) allows both players to take symmetric updates without revealing their private actions, and has been shown to converge to the unique solution of the entropy-regularized matrix game (7) exponentially fast (Cen et al., 2021). However, the PE algorithm has several limitations. First, in the PU update, each player $m \in \{1,2\}$ needs to query the quantities $Q_{k,t}^{(m)}(s,a^{(m)}), \overline{Q}_{k,t}^{(m)}(s,a^{(m)})$ from its opponent. To compute these quantities, the opponent needs to multiply the entire Q-table by its own policy vector. This requires both players to coordinate with each other and share a Q-table. Second, the update of the Q-table in (6) requires full knowledge of the environment transition kernel $\mathcal{P}$ and the reward mapping R, which are unknown a priori in practice. To overcome these limitations, we develop a fully stochastic PE algorithm in the next subsection.

3.2 Stochastic policy extragradient algorithm

The major challenge of the PE algorithm is computing the quantities $Q_{k,t}^{(m)}, \overline{Q}_{k,t}^{(m)}$ and V'_{k+1}, which requires coordinating with the opponent and involves the environment information. Here, we develop a model-free and fully stochastic variant of PE that estimates these key quantities using Markovian stochastic samples. We refer to this algorithm as **stochastic policy extragradient (SPE)**.

Specifically, we first estimate the quantity $V'_{k+1}(s) = f_\tau\big(Q_k(s); \pi_k^{(1)}(s), \pi_k^{(2)}(s)\big)$. By definition of f_τ in (5) and the update of Q_k in (6) (now we use $V'_k(s')$ instead of $V_k(s')$) and using some standard tricks on random variables (see Lemma 2 in Appendix B for a full proof), we can rewrite $V'_{k+1}(s)$ as

$$V'_{k+1}(s) = \frac{\mathbb{E}\Big[\big(R(\widetilde{s},\widetilde{a}^{(1)},\widetilde{a}^{(2)}) + \gamma V'_k(s')\big)\mathbb{1}\{\widetilde{s}=s\}\Big]}{\mu_k(s)} + \tau\mathcal{H}\big(\pi_k^{(1)}(s)\big) - \tau\mathcal{H}\big(\pi_k^{(2)}(s)\big), \tag{11}$$

where $\mu_k(s)$ denotes the stationary state distribution associated with the policy pair $(\pi_k^{(1)}, \pi_k^{(2)})$, and the expectation is taken over $\widetilde{s} \sim \mu_k, \widetilde{a}^{(1)} \sim \pi_k^{(1)}(s), \widetilde{a}^{(2)} \sim \pi_k^{(2)}(s), s' \sim \mathcal{P}(\cdot|\widetilde{s},\widetilde{a}^{(1)},\widetilde{a}^{(2)})$. To estimate this quantity, we query a set $\mathcal{N}_{k+1}$ (with cardinality N_{k+1}) of samples from the Markov decision process following the pair of policies $(\pi_k^{(1)}, \pi_k^{(2)})$. Then, we estimate $V'_{k+1}(s)$ as

$$\widehat{V}_{k+1}(s) = \frac{\frac{1}{N_{k+1}}\sum_{i\in\mathcal{N}_{k+1}}\big(R_i + \gamma\widehat{V}_k(s_{i+1})\big)\mathbb{1}\{s_i=s\}}{\frac{1}{N_{k+1}}\sum_{i\in\mathcal{N}_{k+1}}\mathbb{1}\{s_i=s\}} + \tau\mathcal{H}\big(\pi_k^{(1)}(s)\big) - \tau\mathcal{H}\big(\pi_k^{(2)}(s)\big). \tag{12}$$

Intuitively, we use the sample average of Markovian samples to estimate the expectation terms in (11). Thanks to the concentration phenomenon of dependent samples (Paulin, 2015), these sample averages converge to the desired expected values provided that the sample size is sufficiently large.

Next, we estimate $Q_{k,t}^{(m)}, m \in 1,2$. Leveraging (9) and (6), we obtain the following equivalent characterization for both players $m \in 1,2$ (see Lemma 2 in Appendix B for the proof of equivalence).

$$Q_{k,t}^{(m)}(s,a^{(m)}) = \frac{\mathbb{E}\Big[\big(R(\widetilde{s},\widetilde{a}^{(1)},\widetilde{a}^{(2)}) + \gamma V_k'(s')\big)\mathbb{1}\{\widetilde{s}=s,\widetilde{a}^{(m)}=a^{(m)}\}\Big]}{\mu_{k,t}(s)\pi_{k,t}^{(m)}(a^{(m)}|s)}, \tag{13}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function, $\mu_{k,t}$ denotes the stationary state distribution associated with the policy pair $(\pi_{k,t}^{(1)},\pi_{k,t}^{(2)})$, and the expectation is taken over $\widetilde{s}\sim\mu_{k,t}, \widetilde{a}^{(1)}\sim\pi_{k,t}^{(1)}(s), \widetilde{a}^{(2)}\sim\pi_{k,t}^{(2)}(s), s'\sim\mathcal{P}(\cdot|\widetilde{s},\widetilde{a}^{(1)},\widetilde{a}^{(2)})$. To estimate this quantity, we query a set $\mathcal{N}_{k,t}$ (with cardinality $N_{k,t}$) of samples from the Markov decision process following a pair of *smoothed policies* $\pi_{k,t}^{\prime(m)}(s) = (1-\epsilon')\pi_{k,t}^{(m)}(s) + \frac{\epsilon'}{|\mathcal{A}^{(m)}|}\mathbf{1}$, where $\epsilon'\in[0,1]$ is a small smoothing constant that will be theoretically determined later. Then, we estimate $Q_{k,t}^{(m)}(s,a^{(m)})$ as follows.

$$\widehat{Q}_{k,t}^{(m)}(s,a^{(m)}) = \frac{\frac{1}{N_{k,t}}\sum_{i\in\mathcal{N}_{k,t}}\big(R_i+\gamma\widehat{V}_k(s_{i+1})\big)\mathbb{1}\{s_i=s,a_i^{(m)}=a^{(m)}\}}{\big(\frac{1}{N_{k,t}}\sum_{i\in\mathcal{N}_{k,t}}\mathbb{1}\{s_i=s\}\big)\pi_{k,t}^{\prime(m)}(s)}, \tag{14}$$

where we have replaced the expectations with sample averages, and replaced $V_k'(s')$ with $\widehat{V}_k(s')$. Here, the Markovian samples are queried following the ϵ'-smoothed policies $(\pi_{k,t}^{\prime(1)},\pi_{k,t}^{\prime(2)})$. On one hand, ϵ' should not be too small so that it keeps the denominator of the above estimation away from zero. On the other hand, ϵ' should not be too large so that it is sufficiently close to the original policy.

Similarly, to estimate $\overline{Q}_{k,t}^{(m)}$, we query another set $\overline{\mathcal{N}}_{k,t}$ (with cardinality $\overline{N}_{k,t}$) of samples from the Markov decision process following a pair of smoothed policies $\overline{\pi}_{k,t}^{\prime(m)}(s) = (1-\overline{\epsilon}')\overline{\pi}_{k,t}^{(m)}(s) + \frac{\overline{\epsilon}'}{|\mathcal{A}^{(m)}|}\mathbf{1}$, where $\overline{\epsilon}'\in[0,1]$ will be theoretically determined later. Then, we estimate $\overline{Q}_{k,t}^{(m)}$ as follows.

$$\widehat{\overline{Q}}_{k,t}^{(m)}(s,a^{(m)}) = \frac{\frac{1}{\overline{N}_{k,t}}\sum_{i\in\overline{\mathcal{N}}_{k,t}}\big(R_i+\gamma\widehat{V}_k(s_{i+1})\big)\mathbb{1}\{s_i=s,a_i^{(m)}=a^{(m)}\}}{\big(\frac{1}{\overline{N}_{k,t}}\sum_{i\in\overline{\mathcal{N}}_{k,t}}\mathbb{1}\{s_i=s\}\big)\overline{\pi}_{k,t}^{\prime(m)}(s)}. \tag{15}$$

Remark 1. *The above estimators improve over those introduced in (Wei et al., 2021). First, to estimate the state-action probability involved in the denominator of $Q_{k,t}^{(m)}(s,a^{(m)})$ and $\overline{Q}_{k,t}^{(m)}(s,a^{(m)})$, Wei et al. (2021) uses the state-action frequency estimator $\frac{1}{N_{k,t}}\sum_{i\in\mathcal{N}_{k,t}}\mathbb{1}\{s_i=s,a_i^{(m)}=a^{(m)}\}$, which induces a large variance due to the large state-action product space. As a comparison, we use the estimator $\big(\frac{1}{N_{k,t}}\sum_{i\in\mathcal{N}_{k,t}}\mathbb{1}\{s_i=s\}\big)\pi_{k,t}^{\prime(m)}(s)$ that involves only the state frequency and the smoothed policy $\pi_{k,t}^{\prime(m)}(s)$. Therefore, our estimator has a much smaller variance. Second, to estimate V_{k+1}', Wei et al. (2021) follows the smoothed policies to query the samples, while we follow the original policies $(\pi_k^{(1)},\pi_k^{(2)})$. This makes our estimator more accurate.*

We summarize our stochastic policy extragradient (SPE) algorithm in Algorithm 1. Specifically, in SPE, we estimate the quantities $Q_{k,t}^{(m)}$, $\overline{Q}_{k,t}^{(m)}$, V_k' using their corresponding stochastic estimators. As a result, the SPE algorithm is model-free, and the updates for both players are symmetric and private.

4 Finite-Time Convergence Analysis of SPE

Throughout our convergence analysis, we adopt the following two standard assumptions.

Assumption 1. *Denote $T_{\pi^{(1)},\pi^{(2)}}(s,s') := \inf\{t\geq 1 : s_t = s'|s_0=s\}$ as the first-visit time under the policy pair $\pi^{(1)},\pi^{(2)}$. We assume that*

$$\sup_{s,s'\in\mathcal{S}}\sup_{\pi^{(1)},\pi^{(2)}}\mathbb{E}_{\pi^{(1)},\pi^{(2)}}\big[T_{\pi^{(1)},\pi^{(2)}}(s,s')\big] < +\infty. \tag{16}$$

Algorithm 1 Stochastic policy extragradient (SPE) for entropy-regularized Markov game

Initialize: $V_0'(s)$ for all $s \in \mathcal{S}$.
for *value iterations* $k = 0, 1, \ldots, K-1$ **do**
  Initialize $\pi_{k,0}^{(1)}, \pi_{k,0}^{(2)}$ with uniform distribution.
  for *PU iterations* $t = 0, 1, \ldots, T_k - 1$ **do**
    Players 1,2 sample $N_{k,t}$ Markovian samples following smoothed policies $\pi_{k,t}'^{(1)}, \pi_{k,t}'^{(2)}$.
    Every player $m \in \{1, 2\}$ computes $\widehat{Q}_{k,t}^{(m)}(s, a^{(m)})$ for all $s, a^{(m)}$ using (14).
    Players 1,2 sample $\overline{N}_{k,t}$ Markovian samples following smoothed policies $\overline{\pi}_{k,t}'^{(1)}, \overline{\pi}_{k,t}'^{(2)}$.
    Every player $m \in \{1, 2\}$ computes $\widehat{\overline{Q}}_{k,t}^{(m)}(s, a^{(m)})$ for all $s, a^{(m)}$ using (15).
    Implement the t-th PU iteration for all $s, a^{(1)}, a^{(2)}$ using (8) with estimations (14)&(15).
  end
  Let $\pi_k^{(m)} = \overline{\pi}_{k,T_k}^{(m)}$, $m = 1, 2$. Players sample N_k Markovian samples following $\pi_k^{(1)}, \pi_k^{(2)}$.
  Compute $\widehat{V}_{k+1}(s)$ for all s using (12).
end
Output: $\pi_{K-1}^{(1)}, \pi_{K-1}^{(2)}$.

Assumption 1 is widely used in the reinforcement learning literature (Ortner and Auer, 2007; Ortner, 2020; Wei et al., 2021; Jafarnia-Jahromi et al., 2021). It ensures that every state will be visited at least once within a finite duration of time, thus ensuring that all the states will be visited infinitely often. This guarantees sufficient exploration. In our analysis, we use the following equivalent statement of Assumption 1 for convenience, which means that the stationary state distribution $\mu_{\pi^{(1)},\pi^{(2)}}$ has a uniform lower bound $\mu_{\min} > 0$. Their equivalence is based on Theorem 5.5.11 of (Durrett, 2019).

$$\mu_{\min} := \inf_{s\in\mathcal{S}} \inf_{\pi^{(1)},\pi^{(2)}} \mu_{\pi^{(1)},\pi^{(2)}}(s) = \Big[\sup_{s\in\mathcal{S}} \sup_{\pi^{(1)},\pi^{(2)}} \mathbb{E}_{\pi^{(1)},\pi^{(2)}} T_{\pi^{(1)},\pi^{(2)}}(s,s)\Big]^{-1} > 0. \tag{17}$$

Assumption 2. *There exists a mixing time $t_{mix} \in \mathbb{N}$ such that for any policy pair $\pi^{(1)}, \pi^{(2)}$ and its corresponding stationary state distribution $\mu_{\pi^{(1)},\pi^{(2)}}$, we have*

$$d_{\mathrm{TV}}\big(\mathcal{P}_{\pi^{(1)},\pi^{(2)}}(s_{t_{mix}}), \mu_{\pi^{(1)},\pi^{(2)}}\big) \le \frac{1}{4}.$$

where $\mathcal{P}_{\pi^{(1)},\pi^{(2)}}(s_{t_{mix}})$ denotes the state distribution under the policy pair $\pi^{(1)}, \pi^{(2)}$ at time t_{mix}, and d_{TV} denotes the total variation distance between two probability distributions.

Assumption 2 is also widely adopted in the existing literature (Greensmith et al., 2004; Ortner, 2020; Ciosek, 2021). It ensures that the state distribution is not too far away from its stationary distribution within a finite time t_{mix}. We note that this assumption is much weaker than the popular uniformly ergodic Markov chain assumption, which assumes that $\mathrm{d}_{TV}(\mathcal{P}_{\pi^{(1)},\pi^{(2)}}(s_t), \mu_{\pi^{(1)},\pi^{(2)}})$ decays exponentially fast as $t \to \infty$ (Bhandari et al., 2018; Qiu et al., 2019; Xu and Liang, 2021).

In this subsection, we analyze the finite-time convergence of Algorithm 1 for solving the entropy-regularized Markov game (5). We focus on the convergence rate of the following Nash equilibrium duality gap, which is a standard optimality metric widely adopted in the existing literature (Xu et al., 2020; Jin and Sidford, 2021; Wei et al., 2021).

$$D^{(\tau)}(\pi^{(1)}, \pi^{(2)}) := \max_s \Big(\max_\pi V_{\pi,\pi^{(2)}}^{(\tau)}(s) - \min_{\pi'} V_{\pi^{(1)},\pi'}^{(\tau)}(s)\Big).$$

In particular, when $\tau = 0$, $D^{(0)}(\pi^{(1)}, \pi^{(2)})$ corresponds to the duality gap of the original Markov game. Throughout, we define $A_{\max} := \max\{|\mathcal{A}^{(1)}|, |\mathcal{A}^{(2)}|\}$, $Q_{\max} := \frac{1+\gamma\tau\ln A_{\max}}{1-\gamma}$ and $V_{\max} := \frac{1+\tau\ln A_{\max}}{1-\gamma}$. We also require that the batch sizes of Algorithm 1 satisfy the following conditions.

$$N_{k,t}, \overline{N}_{k,t} \ge \frac{650 t_{\text{mix}} A_{\max}}{\mu_{\min}} \ln\Big(\frac{20 T_{\text{sum}} |\mathcal{S}| A_{\max}}{\delta\sqrt{\mu_{\min}}}\Big), \quad \forall k, t, \tag{18}$$

$$N_{k+1} \ge \frac{650 t_{\text{mix}}}{\mu_{\min}(1-\gamma)^2} \ln\Big(\frac{4}{\delta\sqrt{\mu_{\min}}}\Big), \quad \forall k. \tag{19}$$

Then, we obtain the following convergence result of Algorithm 1, where $T_{\text{sum}} := \sum_{k=0}^{K-1} T_k$.

Theorem 1 (Finite-time convergence rate)**.** *Apply Algorithm 1 to solve the entropy-regularized Markov game with $\tau \in (0, 1]$. Choose learning rate $\eta = [2(\tau + Q_{\max})]^{-1}$, initialization $\|\widehat{V}_0\|_\infty \leq V_{\max}$ and batch sizes $N_{k,t}, \overline{N}_{k,t}, N_{k+1}$ that satisfy (18) & (19). Then, the Nash equilibrium duality gap converges at the following rate with probability at least $1 - \delta$.*

$$D^{(\tau)}\big(\pi^{(1)}_{K-1}, \pi^{(2)}_{K-1}\big) \leq \mathcal{O}\bigg(\frac{V_{\max} \ln A_{\max}}{1-\gamma} \sum_{k=0}^{K-1} \gamma^{K-k} (1-\eta\tau)^{T_k - 1}$$
$$+ \frac{V_{\max}}{1-\gamma} \Big[\frac{t_{mix} A_{\max}}{\mu_{\min}} \ln\Big(\frac{T_{sum} |\mathcal{S}| A_{\max}}{\delta \mu_{\min}} \Big) \Big]^{2/3} \sum_{k=0}^{K-1} \gamma^{K-k-1} \sum_{t=0}^{T_k - 1} (1-\eta\tau)^{T_k - 2 - t} \Big(\frac{1}{N_{k,t}^{2/3}} + \frac{V_{\max}}{\tau \overline{N}_{k,t+1}^{2/3}} \Big)$$
$$+ \frac{V_{\max}^3 \gamma^K}{\tau^2 (1-\gamma)^2} + \frac{t_{mix} V_{\max}^3}{\tau^2 \mu_{\min} (1-\gamma)^3} \ln\Big(\frac{K|\mathcal{S}|}{\delta \mu_{\min}} \Big) \sum_{k=0}^{K-1} \frac{\gamma^{K-k-1}}{N_{k+1}} \bigg). \tag{20}$$

Remark 2. *In the proof of Theorem 1, we also prove that the convergence rate of the Q-function estimation error $\|Q_K - Q_*^{(\tau)}\|_\infty$ is $(1-\gamma)$ times the convergence rate in (20). Here, Q_K corresponds to the Q-function associated with the policy pair $(\pi_K^{(1)}, \pi_K^{(2)})$ produced by Algorithm 1 in the K-th iteration, and $Q_*^{(\tau)}$ corresponds to the optimal Q-function associated with the Nash equilibrium policy pair $\pi_{*\tau}^{(1)}, \pi_{*\tau}^{(2)}$ of the entropy-regularized Markov game.*

Theorem 1 characterizes the convergence of duality gap of the SPE algorithm under general hyperparameter scheduling of the batch sizes $N_{k,t}, \overline{N}_{k,t}, N_{k+1}$ and number of inner iterations T_k. Specifically, it can be seen that as the number of outer iterations K and inner iterations T_k increase, the duality gap converges to an exponentially weighted average of $N_{k,t}^{-2/3}, \overline{N}_{k,t+1}^{-2/3}$ and N_{k+1}^{-1}, and the gap can be made arbitrarily small by choosing sufficiently large batch sizes. Moreover, we show in Theorem 2 later that the above tight convergence rate leads to a substantially improved sample complexity over the state-of-the-art result. In particular, due to the exponentially weighted average structure in (20), the sample complexity is optimized by an adaptive scheduling of the batch sizes.

We further comment on the technical proof of Theorem 1. Note that the analysis of the PE algorithm in (Cen et al., 2021) requires full knowledge of the environment and does not characterize the convergence of duality gap. As a comparison, to establish the duality gap convergence rate (20) of the model-free SPE, we need to make substantial new developments to tightly bound the estimation errors of the proposed stochastic estimators. We elaborate our technical contributions below.

- As we explained in Remark 1, our proposed stochastic estimators in (12), (14) and (15) are more accurate than those used in (Wei et al., 2021). Moreover, note that the sample averages involved in these estimators are over Markovian samples, and we need to apply the concentration inequalities developed in (Paulin, 2015) for dependent samples to establish high-probability estimation error bounds. As a comparison, Wei et al. (2021) assumes independence of these samples (although their setting is Markovian sampling).
- We develop a much refined analysis of the state value function estimation error $\|\widehat{V}_{k+1} - V'_{k+1}\|_\infty$, which is the key to develop tight bounds for all the other estimation errors. Specifically, we first propose the following error decomposition for any state s

  $$|\widehat{V}_{k+1}(s) - V'_{k+1}(s)| = \Big| \frac{\widehat{v}_{k+1}(s)}{\widehat{\mu}_k(s)} - \frac{v_{k+1}(s)}{\mu_k(s)} \Big| \leq \frac{|\widehat{v}_{k+1}(s) - v_{k+1}(s)|}{\mu_k(s)} + \big|\widehat{v}_{k+1}(s)\big| \Big| \frac{\mu_k(s) - \widehat{\mu}_k(s)}{\mu_k(s)\widehat{\mu}_k(s)} \Big|,$$

  where $\widehat{v}_{k+1}(s), \widehat{\mu}_{k+1}(s)$ are sample average estimators of $v_{k+1}(s), \mu_{k+1}(s)$, respectively, and we refer to Appendix B for the definitions of these terms. The motivation is that the $\big|\widehat{v}_{k+1}(s)\big|$ in the second term helps cancel out the estimator $\widehat{\mu}_k(s)$ in the denominator, and then all the denominators do not involve any sample average estimators, which may take a small numerical value that causes divergence and a loose concentration bound. By leveraging this special decomposition and the recursive structure of the stochastic estimator (12), we are able to establish the following key contraction property of the estimation error (see (84) in Appendix B).

  $$\|\widehat{V}_{k+1} - V'_{k+1}\|_\infty \leq \gamma \|\widehat{V}_k - V'_k\|_\infty + \mathcal{O}\big(N_{k+1}^{-1/2}\big).$$

By telescoping the above contraction bound, we obtain tight estimation error bounds for all the proposed stochastic estimators. As a comparison, Wei et al. (2021) directly applied the Azuma-Hoeffding inequality with independent samples to bound the entire estimator and obtain a loose error bound, and Liu et al. (2021) simply assumed a small upper bound for the estimation error.

- We develop a stochastic predictive update (SPU) algorithm with general inexact value function estimations and a finite-time convergence analysis of its duality gap (see Lemma 1 for the SPU algorithm and its convergence proof). This generalizes the convergence result of the PU algorithm established in (Cen et al., 2021), which uses exact value functions based on full knowledge of the environment. Finally, by incorporating our developed tight estimation error bounds into the finite-time duality gap bound of SPU, we obtain the desired convergence rate in Theorem 1.

Based on Theorem 1, we obtain the following sample complexity of SPE for achieving an ϵ-Nash equilibrium duality gap of the original Markov game, i.e., $D^{(0)}(\pi^{(1)}_{K-1}, \pi^{(2)}_{K-1}) \leq \epsilon$. Here, we adopt an adaptive batch size scheduling scheme to optimize the complexity order. The overall sample complexity is given by $\sum_{k=0}^{K-1}\left[N_{k+1} + 2\sum_{t=0}^{T_k-1}(N_{k,t} + \overline{N}_{k,t+1})\right]$.

Theorem 2 (Sample complexity). *Implement Algorithm 1 with $\eta = \mathcal{O}(1-\gamma)$, $\tau = \mathcal{O}(\frac{\epsilon(1-\gamma)}{\ln A_{\max}})$, $K = \mathcal{O}\left[\frac{1}{1-\gamma}\ln\left(\frac{\ln A_{\max}}{\epsilon(1-\gamma)}\right)\right]$ and $T_k = 1 + \frac{k\ln\gamma^{-1}}{\ln(1-\eta\tau)^{-1}}$. Choose the following adaptive batch sizes.*

$$N_{k+1} = \widetilde{\mathcal{O}}\Big(\frac{t_{mix}(\ln^2 A_{\max})\gamma^{-\frac{k}{2}}}{\epsilon^2\mu_{\min}(1-\gamma)^8}\Big),\ N_{k,t} = \overline{N}_{k,t}\epsilon^{\frac{3}{2}}(1-\gamma)^3 = \widetilde{\mathcal{O}}\Big(\frac{t_{mix}A_{\max}(1-\eta\tau)^{\frac{-3(t+1)}{5}}}{\mu_{\min}(1-\gamma)^3}\Big).$$

Then, for any $\epsilon \leq \frac{\ln A_{\max}}{1-\gamma}$, the overall sample complexity to achieve $D^{(0)}(\pi^{(1)}_{K-1}, \pi^{(2)}_{K-1}) \leq \epsilon$ is $\widetilde{\mathcal{O}}(\frac{t_{mix}A_{\max}}{\mu_{\min}\epsilon^{5.5}(1-\gamma)^{13.5}})$. Please refer to (117) in Appendix E for a complete expression.

The above complexity result is obtained by choosing a small $\tau = \mathcal{O}(\frac{\epsilon(1-\gamma)}{\ln A_{\max}})$ for the convergence rate result in Theorem 1. Specifically, we show in Lemma 6 that the duality gap is Lipschitz continuous with regard to the entropy regularization parameter, i.e., $\left|D^{(\tau)}(\pi^{(1)}, \pi^{(2)}) - D^{(0)}(\pi^{(1)}, \pi^{(2)})\right| \leq \frac{2\tau\ln A_{\max}}{1-\gamma}$. Therefore, by choosing a proper small τ, convergence of the duality gap $D^{(\tau)}$ of the entropy-regularized Markov game implies the convergence of the duality gap $D^{(0)}$ of the original Markov game.

Our sample complexity is substantially lower than that of (Wei et al., 2021), which is of the order $\widetilde{O}(\frac{A^3_{\max}|\mathcal{S}|^{10.5}}{\epsilon^8\mu_{\min}(1-\gamma)^{29.5}})$. This is due to multiple advantages in our algorithm design. First, our SPE algorithm uses entropy regularization to handle the simplex probability constraint, and this leads to a faster convergence than the Euclidean projected optimistic gradient update used in (Wei et al., 2021). Second, the optimistic gradient algorithm of (Wei et al., 2021) only updates the policy once per iteration, whereas our SPE performs multiple policy updates ($T_k \gg 1$), which turns out to help improve the sample complexity. Moreover, our SPE allows to use a large constant learning rate $\eta = \mathcal{O}(1-\gamma)$, whereas the algorithm in (Wei et al., 2021) uses a substantially smaller learning rate $\eta \leq \mathcal{O}(\sqrt{(1-\gamma)^5|\mathcal{S}|^{-1}})$ that significantly slows down the convergence.

5 Conclusion

In this paper, we developed a model-free, provably convergent, sample efficient, symmetric and private stochastic policy extra gradient algorithm for solving two-player zero-sum Markov games. Our algorithm leverages entropy regularization to facilitate the algorithm convergence and develops new stochastic estimators to accurately estimate the value functions. We proved that our SPE algorithm achieved a fast convergence rate in terms of the Nash equilibrium duality gap and moreover, achieves a substantially improved sample complexity over the state-of-the-art result. We believe our algorithm deepens the understanding of Markov games from a computation complexity perspective. In the future study, we are excited about exploring other topics along this direction. For example, it is interesting to extend SPE algorithm to the multi-agent setting for solving general-sum Markov games and competitive games that involve multiple cooperative teams.

REFERENCES

Bai, Y., Jin, C., and Yu, T. (2020). Near-optimal reinforcement learning with self-play. *Proc. Advances in Neural Information Processing Systems (Neurips)*, 33.

Bhandari, J., Russo, D., and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. In *Proc. Conference on Learning Theory (COLT)*, volume 75, pages 1691–1692.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. (2020). Fast global convergence of natural policy gradient methods with entropy regularization. *ArXiv:2007.06558*.

Cen, S., Wei, Y., and Chi, Y. (2021). Fast policy extragradient methods for competitive games with entropy regularization. *ArXiv:2105.15186*.

Ciosek, K. (2021). Imitation learning by reinforcement learning. *ArXiv:2108.04763*.

Daskalakis, C., Foster, D. J., and Golowich, N. (2020). Independent policy gradient methods for competitive reinforcement learning. In *Proc. Advances in Neural Information Processing Systems (Neurips)*, volume 33.

Deng, X., Li, Y., Mguni, D. H., Wang, J., and Yang, Y. (2021). On the complexity of computing markov perfect equilibrium in general-sum stochastic games. *ArXiv:2109.01795*.

Durrett, R. (2019). *Probability: theory and examples*, volume 49. Cambridge university press.

Elie, R., Perolat, J., Laurière, M., Geist, M., and Pietquin, O. (2020). On the convergence of model free learning in mean field games. In *Proc. the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 7143–7150.

Ernst, D., Geurts, P., and Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2020). A theoretical analysis of deep q-learning. In *Proc. Learning for Dynamics and Control (L4DC)*, pages 486–489.

Farahmand, A.-M. and Ghavamzadeh, M. (2021). Pid accelerated value iteration algorithm. In *Proc. International Conference on Machine Learning (ICML)*, pages 3143–3153.

Geist, M., Scherrer, B., and Pietquin, O. (2019). A theory of regularized markov decision processes. In *Proc. International Conference on Machine Learning (ICML)*, pages 2160–2169.

Ghosh, M. K., Golui, S., Pal, C., and Pradhan, S. (2021). Zero-sum games for continuous-time markov decision processes with risk-sensitive average cost criterion. *ArXiv:2109.08837*.

Greensmith, E., Bartlett, P. L., and Baxter, J. (2004). Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(9).

Gu, H., Guo, X., Wei, X., and Xu, R. (2021). Mean-field multi-agent reinforcement learning: A decentralized network approach. *ArXiv:2108.02731*.

Guo, H., Fu, Z., Yang, Z., and Wang, Z. (2021). Decentralized single-timescale actor-critic on zero-sum two-player stochastic games. In *Proc. International Conference on Machine Learning (ICML)*, pages 3899–3909.

Guo, X., Hu, A., Xu, R., and Zhang, J. (2019). Learning mean-field games. *Proc. Advances in Neural Information Processing Systems (Neurips)*, 32:4966–4976.

Hu, J. and Wellman, M. P. (2003). Nash q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069.

Huang, B., Lee, J. D., Wang, Z., and Yang, Z. (2021). Towards general function approximation in zero-sum markov games. *ArXiv:2107.14702*.

Jafarnia-Jahromi, M., Jain, R., and Nayyar, A. (2021). Learning zero-sum stochastic games with posterior sampling. *ArXiv:2109.03396*.

Jin, Y. and Sidford, A. (2021). Towards tight bounds on the sample complexity of average-reward mdps. In *Proc. International Conference on Machine Learning (ICML)*, pages 5055–5064.

Kononen, V. and Oja, E. (2004). Asymmetric multiagent reinforcement learning in pricing applications. In *Proc. IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, volume 2, pages 1097–1102.

Leonardos, S., Overman, W., Panageas, I., and Piliouras, G. (2021). Global convergence of multi-agent policy gradient in markov potential games. *ArXiv:2106.01969*.

Liu, B., Yang, Z., and Wang, Z. (2021). Policy optimization in zero-sum markov games: Fictitious self-play provably attains nash equilibria.

Mei, J., Xiao, C., Szepesvari, C., and Schuurmans, D. (2020). On the global convergence rates of softmax policy gradient methods. In *Proc. International Conference on Machine Learning (ICML)*, pages 6820–6829.

Mertikopoulos, P. and Sandholm, W. H. (2016). Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324.

Moerland, T. M., Broekens, J., Plaat, A., and Jonker, C. M. (2018). A0c: Alpha zero in continuous action space. *ArXiv:1805.09613*.

Neu, G., Jonsson, A., and Gómez, V. (2017). A unified view of entropy-regularized markov decision processes. *ArXiv:1705.07798*.

Ortner, P. and Auer, R. (2007). Logarithmic online regret bounds for undiscounted reinforcement learning. In *Proc. Advances in Neural Information Processing Systems (Neurips)*, volume 19, page 49.

Ortner, R. (2020). Regret bounds for reinforcement learning via markov chain concentration. *Journal of Artificial Intelligence Research*, 67:115–128.

Paulin, D. (2015). Concentration inequalities for markov chains by marton couplings and spectral methods. *Electronic Journal of Probability*, 20:1–32.

Qiu, S., Yang, Z., Ye, J., and Wang, Z. (2019). On the finite-time convergence of actor-critic algorithm. In *NeurIPS Optimization Foundations for Reinforcement Learning Workshop*.

Savas, Y., Ahmadi, M., Tanaka, T., and Topcu, U. (2019). Entropy-regularized stochastic games. In *Proc. IEEE 58th Conference on Decision and Control (CDC)*, pages 5955–5962.

Shapley, L. S. (1953). Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100.

Sidford, A., Wang, M., Yang, L., and Ye, Y. (2020). Solving discounted stochastic two-player games with near-optimal time and sample complexity. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2992–3002.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *nature*, 550(7676):354–359.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tamar, A., Wu, Y., Thomas, G., Levine, S., and Abbeel, P. (2016). Value iteration networks. In *Proc. Advances in neural information processing systems (Neurips)*, pages 2154–2162.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. (2019). Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354.

Wang, X. and Sandholm, T. (2002). Reinforcement learning to play an optimal nash equilibrium in team markov games. volume 15, pages 1603–1610.

Wei, C.-Y., Hong, Y.-T., and Lu, C.-J. (2017). Online reinforcement learning in stochastic games. In *Proc. Advances in neural information processing systems (Neurips)*, pages 4994–5004.

Wei, C.-Y., Lee, C.-W., Zhang, M., and Luo, H. (2021). Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In *Proc. Conference on Learning Theory (COLT)*.

Xu, T. and Liang, Y. (2021). Sample complexity bounds for two timescale value-based reinforcement learning algorithms. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 811–819.

Xu, Y., Deng, Z., Wang, M., Xu, W., So, A. M.-C., and Cui, S. (2020). Voting-based multiagent reinforcement learning for intelligent iot. *IEEE Internet of Things Journal*, 8(4):2681–2693.

Zhang, K., Kakade, S., Basar, T., and Yang, L. (2020). Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity. In *Proc. Advances in Neural Information Processing Systems (Neurips)*, volume 33.

Zhang, K., Yang, Z., Liu, H., Zhang, T., and Basar, T. (2021). Finite-sample analysis for decentralized batch multi-agent reinforcement learning with networked agents. *IEEE Transactions on Automatic Control*.

Zhang, X., Zhang, K., Miehling, E., and Basar, T. (2019). Non-cooperative inverse reinforcement learning. In *Proc. Advances in Neural Information Processing Systems (Neurips)*, volume 32, pages 9487–9497.

Zhao, Y., Tian, Y., Lee, J. D., and Du, S. S. (2021). Provably efficient policy gradient methods for two-player zero-sum markov games. *ArXiv:2102.08903*.

Zhu, Y. and Zhao, D. (2020). Online minimax q network learning for two-player zero-sum markov games. *IEEE Transactions on Neural Networks and Learning Systems*.

Zou, S., Xu, T., and Liang, Y. (2019). Finite-sample analysis for sarsa and q-learning with linear function approximation. *ArXiv:1902.02234*.

Appendix

Table of Contents

A	Analysis of stochastic PU (SPU) for entropy-regularized matrix game	13
B	Estimation error bounds	18
C	Properties of the Duality Gap	25
D	Proof of Theorem 1	27
E	Proof of Theorem 2	31
F	Explanation of Table 1	34

A Analysis of stochastic PU (SPU) for entropy-regularized matrix game

We first consider the following zero-sum entropy regularized matrix game, which can be considered as a simple special case of Markov game with only one state.

$$\max_{\mu\in\Delta(d_1)}\min_{\nu\in\Delta(d_2)} h_\tau(\mu,\nu) := \mu^\top Q\nu + \tau\mathcal{H}(\mu) - \tau\mathcal{H}(\nu). \tag{21}$$

where $Q \in \mathbb{R}^{d_1\times d_2}$ is fixed. The solution (μ_τ^*, ν_τ^*) of the above problem, also known as the quantal response equilibrium (QRE), satisfies the following condition (Cen et al., 2021)

$$\mu_\tau^* \propto \exp(Q\nu_\tau^*/\tau), \nu_\tau^* \propto \exp(-Q^\top\mu_\tau^*/\tau). \tag{22}$$

Our stochastic PU (SPU) algorithm for the above matrix game is written as follows,

$$\overline{\mu}_{t+1}(a) \propto \mu_t(a)^{1-\eta\tau}\exp\big(\eta\big[Q\nu_t + \delta_t^{(1)}\big]_a\big) \tag{23}$$

$$\overline{\nu}_{t+1}(b) \propto \nu_t(b)^{1-\eta\tau}\exp\big(-\eta\big[Q^\top\mu_t + \delta_t^{(2)}\big]_b\big) \tag{24}$$

$$\mu_{t+1}(a) \propto \mu_t(a)^{1-\eta\tau}\exp\big(\eta\big[Q\overline{\nu}_{t+1} + \overline{\delta}_t^{(1)}\big]_a\big) \tag{25}$$

$$\nu_{t+1}(b) \propto \nu_t(b)^{1-\eta\tau}\exp\big(-\eta\big[Q^\top\overline{\mu}_{t+1} + \overline{\delta}_t^{(2)}\big]_b\big), \tag{26}$$

where $\delta_t^{(1)}$ is an additive noise to the underlying true quantity $Q\nu_t$, and the other noises $\delta_t^{(2)}$, $\overline{\delta}_t^{(1)}$, $\overline{\delta}_t^{(2)}$ are similar. If all these noises are zero, the above stochastic PU algorithm reduces to the PU algorithm for matrix game defined in Section 2 of (Cen et al., 2021).

The convergence rate of stochastic PU for matrix game is shown below. The proof logic follows from that of (Cen et al., 2021).

Lemma 1. *Choose learning rate $\eta \le \big[2(\tau + \|Q\|_\infty)\big]^{-1}$ and use the uniform policy initialization $\mu_0 = \mathbf{1}/d_1, \nu_0 = \mathbf{1}/d_2$. Then, the SPU algorithm defined by eqs. (23)-(26) for entropy-regularized matrix game has the following convergence rates.*

$$\begin{aligned}
&KL(\mu_\tau^*\|\mu_t) + KL(\nu_\tau^*\|\nu_t)\\
&\le (1-\eta\tau)^t \ln(d_1 d_2)\\
&\quad + \sum_{j=0}^{t-1}(1-\eta\tau)^{t-1-j}\Big[2\eta^2\big(\|\delta_j^{(1)}\|_\infty^2 + \|\delta_j^{(2)}\|_\infty^2\big) + \frac{2\eta}{\tau}\big(\|\overline{\delta}_j^{(1)}\|_\infty^2 + \|\overline{\delta}_j^{(2)}\|_\infty^2\big)\Big].
\end{aligned} \tag{27}$$

$$
\begin{aligned}
&\left|h_\tau(\mu_\tau^\star,\nu_\tau^\star)-h_\tau(\overline{\mu}_{t+1},\overline{\nu}_{t+1})\right| \\
&\le \frac{4}{\eta}(1-\eta\tau)^t\ln(d_1d_2) \\
&\quad+\sum_{j=0}^{t}(1-\eta\tau)^{t-1-j}\Big[8\eta\big(\|\delta_j^{(1)}\|_\infty^2+\|\delta_j^{(2)}\|_\infty^2\big)+\frac{33}{4\tau}\big(\|\overline{\delta}_j^{(1)}\|_\infty^2+\|\overline{\delta}_j^{(2)}\|_\infty^2\big)\Big]. \qquad (28)\\
&\max_{\mu\in\Delta(d_1)}h_\tau(\mu,\overline{\nu}_{t+1})-\min_{\nu\in\Delta(d_2)}h_\tau(\overline{\mu}_{t+1},\nu) \\
&\le \frac{2}{\eta}(1-\eta\tau)^t\ln(d_1d_2) \\
&\quad+\sum_{j=0}^{t}(1-\eta\tau)^{t-1-j}\Big[16\eta\big(\|\delta_j^{(1)}\|_\infty^2+\|\delta_j^{(2)}\|_\infty^2\big)+\frac{12}{\tau}\big(\|\overline{\delta}_j^{(1)}\|_\infty^2+\|\overline{\delta}_j^{(2)}\|_\infty^2\big)\Big]. \qquad (29)
\end{aligned}
$$

Proof. Eq. (25) implies that for some constant $c\in\mathbb{R}$,

$$
\begin{aligned}
&\ln\mu_{t+1}=(1-\eta\tau)\ln\mu_t+\eta(Q\overline{\nu}_{t+1}+\overline{\delta}_t^{(1)})+c\mathbf{1} \\
&\Rightarrow\big\langle\ln\mu_{t+1}-(1-\eta\tau)\ln\mu_t-\eta(Q\overline{\nu}_{t+1}+\overline{\delta}_t^{(1)}),\overline{\mu}_{t+1}-\mu_\tau^*\big\rangle=0. \qquad (30)
\end{aligned}
$$

where we used $\langle\overline{\mu}_{t+1}-\mu_\tau^*,\mathbf{1}\rangle=0$. Similarly, eqs. (26)&(22) respectively imply that

$$
\begin{aligned}
\big\langle\ln\nu_{t+1}-(1-\eta\tau)\ln\nu_t+\eta(Q^\top\overline{\mu}_{t+1}+\overline{\delta}_t^{(2)}),\overline{\nu}_{t+1}-\nu_\tau^*\big\rangle&=0 \qquad (31)\\
\big\langle\ln\mu_\tau^*-Q\nu_\tau^*/\tau,\overline{\mu}_{t+1}-\mu_\tau^*\big\rangle&=0 \qquad (32)\\
\big\langle\ln\nu_\tau^*+Q^\top\mu_\tau^*/\tau,\overline{\nu}_{t+1}-\nu_\tau^*\big\rangle&=0. \qquad (33)
\end{aligned}
$$

Similarly, eqs. (23)&(25) imply that

$$
\begin{aligned}
&\big\langle\ln(\overline{\mu}_{t+1}/\mu_{t+1}),\overline{\mu}_{t+1}-\mu_{t+1}\big\rangle \\
&=\big\langle\eta\big[Q(\nu_t-\overline{\nu}_{t+1})+\delta_t^{(1)}-\overline{\delta}_t^{(1)}\big],\overline{\mu}_{t+1}-\mu_{t+1}\big\rangle \\
&\le\eta\|Q\|_\infty\|\overline{\mu}_{t+1}-\mu_{t+1}\|_1\|\nu_t-\overline{\nu}_{t+1}\|_1+\eta\|\delta_t^{(1)}-\overline{\delta}_t^{(1)}\|_\infty\|\overline{\mu}_{t+1}-\mu_{t+1}\|_1 \\
&\le\frac{\eta}{2}\|Q\|_\infty\big(\|\overline{\mu}_{t+1}-\mu_{t+1}\|_1^2+\|\nu_t-\overline{\nu}_{t+1}\|_1^2\big)+\frac{\eta^2}{2(1-\eta\|Q\|_\infty)}\|\delta_t^{(1)}-\overline{\delta}_t^{(1)}\|_\infty^2 \\
&\quad+\frac{1}{2}(1-\eta\|Q\|_\infty)\|\overline{\mu}_{t+1}-\mu_{t+1}\|_1^2 \\
&\overset{(i)}{\le}KL(\mu_{t+1}\|\overline{\mu}_{t+1})+\eta\|Q\|_\infty KL(\overline{\nu}_{t+1}\|\nu_t)+2\eta^2\big(\|\delta_t^{(1)}\|_\infty^2+\|\overline{\delta}_t^{(1)}\|_\infty^2\big), \qquad (34)
\end{aligned}
$$

where (i) uses the Pinsker's inequality and $\eta\le\big[2(\tau+\|Q\|_\infty)\big]^{-1}$.

Similarly to the above derivation, we can infer from eqs. (24)&(26) that

$$
\begin{aligned}
\big\langle\ln(\overline{\nu}_{t+1}/\nu_{t+1}),\overline{\nu}_{t+1}-\nu_{t+1}\big\rangle\le&KL(\nu_{t+1}\|\overline{\nu}_{t+1})+\eta\|Q\|_\infty KL(\overline{\mu}_{t+1}\|\mu_t) \\
&+2\eta^2\big(\|\delta_t^{(2)}\|_\infty^2+\|\overline{\delta}_t^{(2)}\|_\infty^2\big). \qquad (35)
\end{aligned}
$$

Computing eq. (30)+eq. (31)-$\eta\tau$[eq. (32)+eq. (33)] yields that

$$
\begin{aligned}
&\eta\langle\overline{\delta}_t^{(1)},\overline{\mu}_{t+1}-\mu_\tau^*\rangle-\eta\langle\overline{\delta}_t^{(2)},\overline{\nu}_{t+1}-\nu_\tau^*\rangle \\
=&\big\langle\ln\mu_{t+1}-(1-\eta\tau)\ln\mu_t-\eta\tau\ln\mu_\tau^*-\eta Q(\overline{\nu}_{t+1}-\nu_\tau^*),\overline{\mu}_{t+1}-\mu_\tau^*\big\rangle \\
&+\big\langle\ln\nu_{t+1}-(1-\eta\tau)\ln\nu_t-\eta\tau\ln\nu_\tau^*+\eta Q^\top(\overline{\mu}_{t+1}-\mu_\tau^*),\overline{\nu}_{t+1}-\nu_\tau^*\big\rangle \\
=&\big\langle\ln(\mu_{t+1}/\mu_\tau^*)-(1-\eta\tau)\ln(\mu_t/\mu_\tau^*),\overline{\mu}_{t+1}-\mu_\tau^*\big\rangle \\
&+\big\langle\ln(\nu_{t+1}/\nu_\tau^*)-(1-\eta\tau)\ln(\nu_t/\nu_\tau^*),\overline{\nu}_{t+1}-\nu_\tau^*\big\rangle \\
=&KL(\mu_\tau^*\|\mu_{t+1})-(1-\eta\tau)KL(\mu_\tau^*\|\mu_t)
\end{aligned}
$$

$$
\begin{aligned}
&+\big\langle \ln(\overline{\mu}_{t+1}/\mu_\tau^*) - \ln(\overline{\mu}_{t+1}/\mu_{t+1}) - (1-\eta\tau)\big[\ln(\overline{\mu}_{t+1}/\mu_\tau^*) - \ln(\overline{\mu}_{t+1}/\mu_t)\big], \overline{\mu}_{t+1}\big\rangle \\
&+ KL(\nu_\tau^*\|\nu_{t+1}) - (1-\eta\tau)KL(\nu_\tau^*\|\nu_t) \\
&+\big\langle \ln(\overline{\nu}_{t+1}/\nu_\tau^*) - \ln(\overline{\nu}_{t+1}/\nu_{t+1}) - (1-\eta\tau)\big[\ln(\overline{\nu}_{t+1}/\nu_\tau^*) - \ln(\overline{\nu}_{t+1}/\nu_t)\big], \overline{\nu}_{t+1}\big\rangle \\
=&KL(\mu_\tau^*\|\mu_{t+1}) - (1-\eta\tau)KL(\mu_\tau^*\|\mu_t) + \eta\tau KL(\overline{\mu}_{t+1}\|\mu_\tau^*) \\
&+ (1-\eta\tau)KL(\overline{\mu}_{t+1}\|\mu_t) + KL(\mu_{t+1}\|\overline{\mu}_{t+1}) \\
&+ KL(\nu_\tau^*\|\nu_{t+1}) - (1-\eta\tau)KL(\nu_\tau^*\|\nu_t) + \eta\tau KL(\overline{\nu}_{t+1}\|\nu_\tau^*) \\
&+ (1-\eta\tau)KL(\overline{\nu}_{t+1}\|\nu_t) + KL(\nu_{t+1}\|\overline{\nu}_{t+1}) \\
&- \big\langle \ln(\overline{\mu}_{t+1}/\mu_{t+1}), \overline{\mu}_{t+1} - \mu_{t+1}\big\rangle - \big\langle \ln(\overline{\nu}_{t+1}/\nu_{t+1}), \overline{\nu}_{t+1} - \nu_{t+1}\big\rangle \\
\overset{(i)}{\geq}&KL(\mu_\tau^*\|\mu_{t+1}) - (1-\eta\tau)KL(\mu_\tau^*\|\mu_t) + \eta\tau KL(\overline{\mu}_{t+1}\|\mu_\tau^*) + (1-\eta\tau-\eta\|Q\|_\infty)KL(\overline{\mu}_{t+1}\|\mu_t) \\
&+ KL(\nu_\tau^*\|\nu_{t+1}) - (1-\eta\tau)KL(\nu_\tau^*\|\nu_t) + \eta\tau KL(\overline{\nu}_{t+1}\|\nu_\tau^*) + (1-\eta\tau-\eta\|Q\|_\infty)KL(\overline{\nu}_{t+1}\|\nu_t) \\
&- 2\eta^2\big(\|\delta_t^{(1)}\|_\infty^2 + \|\overline{\delta}_t^{(1)}\|_\infty^2 + \|\delta_t^{(2)}\|_\infty^2 + \|\overline{\delta}_t^{(2)}\|_\infty^2\big),
\end{aligned} \tag{36}
$$

where (i) uses eqs. (34)&(35). The left side of the above inequality has the following upper bound.

$$
\begin{aligned}
&\eta\langle \overline{\delta}_t^{(1)}, \overline{\mu}_{t+1} - \mu_\tau^*\rangle - \eta\langle \overline{\delta}_t^{(2)}, \overline{\nu}_{t+1} - \nu_\tau^*\rangle \\
\leq&\eta\|\overline{\delta}_t^{(1)}\|_\infty\|\overline{\mu}_{t+1} - \mu_\tau^*\|_1 + \eta\|\overline{\delta}_t^{(2)}\|_\infty\|\overline{\nu}_{t+1} - \nu_\tau^*\|_1 \\
\leq&\frac{\eta}{2}\Big(\frac{2\|\overline{\delta}_t^{(1)}\|_\infty^2}{\tau} + \frac{\tau}{2}\|\overline{\mu}_{t+1} - \mu_\tau^*\|_1^2 + \frac{2\|\overline{\delta}_t^{(2)}\|_\infty^2}{\tau} + \frac{\tau}{2}\|\overline{\nu}_{t+1} - \nu_\tau^*\|_1^2\Big) \\
\leq&\frac{\eta}{\tau}\big(\|\overline{\delta}_t^{(1)}\|_\infty^2 + \|\overline{\delta}_t^{(2)}\|_\infty^2\big) + \frac{\eta\tau}{2}\big(KL(\overline{\mu}_{t+1}\|\mu_\tau^*) + KL(\overline{\nu}_{t+1}\|\nu_\tau^*)\big).
\end{aligned} \tag{37}
$$

Substituting eq. (37) into eq. (36) and rearranging it yields that

$$
\begin{aligned}
&KL(\mu_\tau^*\|\mu_{t+1}) + KL(\nu_\tau^*\|\nu_{t+1}) + \frac{\eta\tau}{2}\big[KL(\overline{\mu}_{t+1}\|\mu_\tau^*) + KL(\overline{\nu}_{t+1}\|\nu_\tau^*)\big] \\
&+ \frac{1}{2}\big[KL(\overline{\mu}_{t+1}\|\mu_t) + KL(\overline{\nu}_{t+1}\|\nu_t)\big] \\
&\leq (1-\eta\tau)\big[KL(\mu_\tau^*\|\mu_t) + KL(\nu_\tau^*\|\nu_t)\big] + 2\eta^2\big(\|\delta_t^{(1)}\|_\infty^2 + \|\delta_t^{(2)}\|_\infty^2\big) \\
&\quad + \frac{2\eta}{\tau}\big(\|\overline{\delta}_t^{(1)}\|_\infty^2 + \|\overline{\delta}_t^{(2)}\|_\infty^2\big),
\end{aligned} \tag{38}
$$

where we use $\eta \leq \big[2(\tau + \|Q\|_\infty)\big]^{-1} \leq (2\tau)^{-1}$. This further implies that

$$
\begin{aligned}
&KL(\mu_\tau^*\|\mu_{t+1}) + KL(\nu_\tau^*\|\nu_{t+1}) \\
&\leq (1-\eta\tau)\big[KL(\mu_\tau^*\|\mu_t) + KL(\nu_\tau^*\|\nu_t)\big] \\
&\quad + 2\eta^2\big(\|\delta_t^{(1)}\|_\infty^2 + \|\delta_t^{(2)}\|_\infty^2\big) + \frac{2\eta}{\tau}\big(\|\overline{\delta}_t^{(1)}\|_\infty^2 + \|\overline{\delta}_t^{(2)}\|_\infty^2\big).
\end{aligned} \tag{39}
$$

Consequently, eq. (27) can be proved via iterating the above inequality and using the facts that $KL(\mu_\tau^*\|\mu_0) \leq \ln d_1$, $KL(\nu_\tau^*\|\nu_0) \leq \ln d_2$ (Since μ_0 and ν_0 are uniform probability vectors).

Next, we will prove eq. (28).

Computing eq. (30) $-\eta\tau\cdot$eq. (32) yields that

$$
\begin{aligned}
&\langle Q(\overline{\nu}_{t+1} - \nu_\tau^*), \overline{\mu}_{t+1} - \mu_\tau^*\rangle \\
&= \eta^{-1}\big\langle \ln\mu_{t+1} - (1-\eta\tau)\ln\mu_t - \eta\tau\ln\mu_\tau^* - \eta\overline{\delta}_t^{(1)}, \overline{\mu}_{t+1} - \mu_\tau^*\big\rangle.
\end{aligned} \tag{40}
$$

Hence, we conclude that

$$
\begin{aligned}
&h_\tau(\overline{\mu}_{t+1}, \overline{\nu}_{t+1}) - h_\tau(\mu_\tau^\star, \nu_\tau^\star) \\
&\leq h_\tau(\overline{\mu}_{t+1}, \overline{\nu}_{t+1}) - h_\tau(\overline{\mu}_{t+1}, \nu_\tau^\star) \\
&= \overline{\mu}_{t+1}^\top Q(\overline{\nu}_{t+1} - \nu_\tau^\star) + \tau\overline{\nu}_{t+1}^\top\ln\overline{\nu}_{t+1} - \tau\nu_\tau^{\star\top}\ln\nu_\tau^\star \\
&= (\overline{\mu}_{t+1} - \mu_\tau^\star)^\top Q(\overline{\nu}_{t+1} - \nu_\tau^\star) + \mu_\tau^{\star\top}Q(\overline{\nu}_{t+1} - \nu_\tau^\star) + \tau\overline{\nu}_{t+1}^\top\ln\overline{\nu}_{t+1} - \tau\nu_\tau^{\star\top}\ln\nu_\tau^\star
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(i)}{=} \eta^{-1}\big\langle \ln \mu_{t+1} - (1-\eta\tau)\ln \mu_t - \eta\tau \ln \mu_\tau^* - \eta\overline{\delta}_t^{(1)}, \overline{\mu}_{t+1} - \mu_\tau^* \big\rangle - \tau\big\langle \ln \nu_\tau^*, \overline{\nu}_{t+1} - \nu_\tau^* \big\rangle \\
&\quad + \tau \overline{\nu}_{t+1}^\top \ln \overline{\nu}_{t+1} - \tau \nu_\tau^{\star\top} \ln \nu_\tau^\star \\
&\overset{(ii)}{=} \eta^{-1}\big[KL(\mu_\tau^*\|\mu_{t+1}) - (1-\eta\tau)KL(\mu_\tau^*\|\mu_t) + \eta\tau KL(\overline{\mu}_{t+1}\|\mu_\tau^*) + (1-\eta\tau)KL(\overline{\mu}_{t+1}\|\mu_t) \\
&\quad + KL(\mu_{t+1}\|\overline{\mu}_{t+1}) - \big\langle \ln(\overline{\mu}_{t+1}/\mu_{t+1}), \overline{\mu}_{t+1} - \mu_{t+1}\big\rangle\big] + \tau KL(\overline{\nu}_{t+1}\|\nu_\tau^\star) - \langle \overline{\delta}_t^{(1)}, \overline{\mu}_{t+1} - \mu_\tau^* \rangle \\
&\le \eta^{-1}\big[KL(\mu_\tau^*\|\mu_{t+1}) - (1-\eta\tau)KL(\mu_\tau^*\|\mu_t) + \eta\tau\big(KL(\overline{\mu}_{t+1}\|\mu_\tau^*) + KL(\overline{\nu}_{t+1}\|\nu_\tau^\star)\big) \\
&\quad + (1-\eta\tau)KL(\overline{\mu}_{t+1}\|\mu_t) - KL(\overline{\mu}_{t+1}\|\mu_{t+1})\big] + \frac{1}{2\tau}\|\overline{\delta}_t^{(1)}\|_\infty^2 + \frac{\tau}{2}\|\overline{\mu}_{t+1} - \mu_\tau^*\|_1^2 \\
&\overset{(iii)}{\le} \eta^{-1}\big[KL(\mu_\tau^*\|\mu_{t+1}) + 2\eta\tau\big(KL(\overline{\mu}_{t+1}\|\mu_\tau^*) + KL(\overline{\nu}_{t+1}\|\nu_\tau^\star)\big) + KL(\overline{\mu}_{t+1}\|\mu_t)\big] + \frac{1}{4\tau}\|\overline{\delta}_t^{(1)}\|_\infty^2 \\
&\overset{(iv)}{\le} 4\eta^{-1}\Big[KL(\mu_\tau^*\|\mu_t) + KL(\nu_\tau^*\|\nu_t) + 2\eta^2\big(\|\delta_t^{(1)}\|_\infty^2 + \|\delta_t^{(2)}\|_\infty^2\big) + \frac{2\eta}{\tau}\big(\|\overline{\delta}_t^{(1)}\|_\infty^2 + \|\overline{\delta}_t^{(2)}\|_\infty^2\big)\Big] \\
&\quad + \frac{1}{4\tau}\|\overline{\delta}_t^{(1)}\|_\infty^2 \\
&\overset{(v)}{\le} 4\eta^{-1}\big[KL(\mu_\tau^*\|\mu_t) + KL(\nu_\tau^*\|\nu_t)\big] \\
&\quad + 8\eta\big(\|\delta_t^{(1)}\|_\infty^2 + \|\delta_t^{(2)}\|_\infty^2\big) + \frac{33}{4\tau}\big(\|\overline{\delta}_t^{(1)}\|_\infty^2 + \|\overline{\delta}_t^{(2)}\|_\infty^2\big),
\end{aligned} \tag{41}
$$

where (i) uses eq. (40)&(33), (ii) follows the derivation of eq. (36), (iii) uses the Pinsker's inequality, (iv) uses eq. (38), and (v) uses $\eta \le \big[2(\tau + \|Q\|_\infty)\big]^{-1}$. In a similar way, we can also prove that

$$
\begin{aligned}
h_\tau(\mu_\tau^\star, \nu_\tau^\star) - h_\tau(\overline{\mu}_{t+1}, \overline{\nu}_{t+1}) &\le h_\tau(\mu_\tau^\star, \overline{\nu}_{t+1}) - h_\tau(\overline{\mu}_{t+1}, \overline{\nu}_{t+1}) \\
&\le 4\eta^{-1}\big[KL(\mu_\tau^*\|\mu_t) + KL(\nu_\tau^*\|\nu_t)\big] \\
&\quad + 8\eta\big(\|\delta_t^{(1)}\|_\infty^2 + \|\delta_t^{(2)}\|_\infty^2\big) + \frac{33}{4\tau}\big(\|\overline{\delta}_t^{(1)}\|_\infty^2 + \|\overline{\delta}_t^{(2)}\|_\infty^2\big).
\end{aligned} \tag{42}
$$

Combining eqs. (41)&(42) yields that

$$
\begin{aligned}
&\big|h_\tau(\mu_\tau^\star, \nu_\tau^\star) - h_\tau(\overline{\mu}_{t+1}, \overline{\nu}_{t+1})\big| \\
&\le 4\eta^{-1}\big[KL(\mu_\tau^*\|\mu_t) + KL(\nu_\tau^*\|\nu_t)\big] + 8\eta\big(\|\delta_t^{(1)}\|_\infty^2 + \|\delta_t^{(2)}\|_\infty^2\big) + \frac{33}{4\tau}\big(\|\overline{\delta}_t^{(1)}\|_\infty^2 + \|\overline{\delta}_t^{(2)}\|_\infty^2\big) \\
&\overset{(i)}{\le} 4\eta^{-1}(1-\eta\tau)^t \ln(d_1 d_2) \\
&\quad + \sum_{j=0}^{t-1}(1-\eta\tau)^{t-1-j}\Big[8\eta\big(\|\delta_j^{(1)}\|_\infty^2 + \|\delta_j^{(2)}\|_\infty^2\big) + \frac{8}{\tau}\big(\|\overline{\delta}_j^{(1)}\|_\infty^2 + \|\overline{\delta}_j^{(2)}\|_\infty^2\big)\Big] \\
&\quad + 8\eta\big(\|\delta_t^{(1)}\|_\infty^2 + \|\delta_t^{(2)}\|_\infty^2\big) + \frac{33}{4\tau}\big(\|\overline{\delta}_t^{(1)}\|_\infty^2 + \|\overline{\delta}_t^{(2)}\|_\infty^2\big) \\
&\le 4\eta^{-1}(1-\eta\tau)^t \ln(d_1 d_2) \\
&\quad + \sum_{j=0}^{t}(1-\eta\tau)^{t-1-j}\Big[8\eta\big(\|\delta_j^{(1)}\|_\infty^2 + \|\delta_j^{(2)}\|_\infty^2\big) + \frac{33}{4\tau}\big(\|\overline{\delta}_j^{(1)}\|_\infty^2 + \|\overline{\delta}_j^{(2)}\|_\infty^2\big)\Big],
\end{aligned}
$$

where (i) uses eq. (27). This proves eq. (28).

Next, to prove the duality gap (29), we first derive an upper bound of $KL(\mu_\tau^\star\|\overline{\mu}_{t+1}) + KL(\nu_\tau^\star\|\overline{\nu}_{t+1})$.

$$
\begin{aligned}
&KL(\nu_\tau^\star\|\overline{\nu}_{t+1}) \\
&= KL(\nu_\tau^\star\|\nu_{t+1}) - KL(\overline{\nu}_{t+1}\|\nu_{t+1}) + \big\langle \overline{\nu}_{t+1} - \nu_\tau^\star, \ln \overline{\nu}_{t+1} - \ln \nu_{t+1}\big\rangle \\
&\le KL(\nu_\tau^\star\|\nu_{t+1}) + \big\langle \overline{\nu}_{t+1} - \nu_\tau^\star, \eta Q^\top(\overline{\mu}_{t+1} - \mu_t) + \eta(\overline{\delta}_t^{(2)} - \delta_t^{(2)})\big\rangle \\
&\le KL(\nu_\tau^\star\|\nu_{t+1}) + \|\overline{\nu}_{t+1} - \nu_\tau^\star\|_1\big(\eta\|Q\|_\infty\|\overline{\mu}_{t+1} - \mu_t\|_1 + \eta\|\overline{\delta}_t^{(2)} - \delta_t^{(2)}\|_\infty\big) \\
&\le KL(\nu_\tau^\star\|\nu_{t+1}) \\
&\quad + \frac{1}{2}\Big[\Big(\eta\|Q\|_\infty + \frac{1}{4}\Big)\|\overline{\nu}_{t+1} - \nu_\tau^\star\|_1^2 + \eta\|Q\|_\infty\|\overline{\mu}_{t+1} - \mu_t\|_1^2 + 4\eta^2\|\overline{\delta}_t^{(2)} - \delta_t^{(2)}\|_\infty^2\Big]
\end{aligned}
$$

$$\overset{(i)}{\leq} KL(\nu_\tau^\star \| \nu_{t+1}) + \frac{3}{4} KL(\nu_\tau^\star \| \overline{\nu}_{t+1}) + \frac{1}{2} KL(\overline{\mu}_{t+1} \| \mu_t) + 4\eta^2 \big(\|\delta_t^{(2)}\|_\infty^2 + \|\overline{\delta}_t^{(2)}\|_\infty^2 \big), \tag{43}$$

where (i) uses the Pinker's inequality and $\eta \leq \big[2(\tau + \|Q\|_\infty)\big]^{-1}$. Rearranging the above inequality yields that

$$KL(\nu_\tau^\star \| \overline{\nu}_{t+1}) \leq 4KL(\nu_\tau^\star \| \nu_{t+1}) + 2KL(\overline{\mu}_{t+1} \| \mu_t) + 16\eta^2 \big(\|\delta_t^{(2)}\|_\infty^2 + \|\overline{\delta}_t^{(2)}\|_\infty^2 \big). \tag{44}$$

Similarly, we can prove that

$$KL(\mu_\tau^\star \| \overline{\mu}_{t+1}) \leq 4KL(\mu_\tau^\star \| \mu_{t+1}) + 2KL(\overline{\nu}_{t+1} \| \nu_t) + 16\eta^2 \big(\|\delta_t^{(1)}\|_\infty^2 + \|\overline{\delta}_t^{(1)}\|_\infty^2 \big). \tag{45}$$

Summing up eqs. (44)&(45) yields that

$$\begin{aligned}
& KL(\mu_\tau^\star \| \overline{\mu}_{t+1}) + KL(\nu_\tau^\star \| \overline{\nu}_{t+1}) \\
& \leq 4\big[KL(\mu_\tau^* \| \mu_{t+1}) + KL(\nu_\tau^* \| \nu_{t+1})\big] + 2\big[KL(\overline{\mu}_{t+1} \| \mu_t) + KL(\overline{\nu}_{t+1} \| \nu_t)\big] \\
& \quad + 16\eta^2 \big(\|\delta_t^{(1)}\|_\infty^2 + \|\overline{\delta}_t^{(1)}\|_\infty^2 + \|\delta_t^{(2)}\|_\infty^2 + \|\overline{\delta}_t^{(2)}\|_\infty^2 \big) \\
& \overset{(i)}{\leq} 4\big[KL(\mu_\tau^* \| \mu_t) + KL(\nu_\tau^* \| \nu_t)\big] + 24\eta^2 \big(\|\delta_t^{(1)}\|_\infty^2 + \|\delta_t^{(2)}\|_\infty^2 \big) \\
& \quad + \frac{16\eta}{\tau} \big(\|\overline{\delta}_t^{(1)}\|_\infty^2 + \|\overline{\delta}_t^{(2)}\|_\infty^2 \big),
\end{aligned} \tag{46}$$

where (i) uses eq. (38) and $\eta \leq \big[2(\tau + \|Q\|_\infty)\big]^{-1} \leq (2\tau)^{-1}$.

Finally, we prove the duality gap bound (29).

$$\begin{aligned}
h_\tau(\mu_\tau^\star, \nu_\tau^\star) - h_\tau(\mu, \nu_\tau^\star) &= (\mu_\tau^\star - \mu)^\top Q \nu_\tau^\star + \tau \mu^\top \ln \mu - \tau \mu_\tau^{\star\top} \ln \mu_\tau^\star \\
&\overset{(i)}{=} \tau(\langle \mu_\tau^\star - \mu, \ln \mu_\tau^\star \rangle + \mu^\top \ln \mu - \mu_\tau^{\star\top} \ln \mu_\tau^\star) = \tau KL(\mu \| \mu_\tau^\star),
\end{aligned} \tag{47}$$

where (i) uses eq. (32).

$$\begin{aligned}
& h_\tau(\mu, \overline{\nu}_{t+1}) - h_\tau(\mu, \nu_\tau^\star) \\
& = \mu^\top Q(\overline{\nu}_{t+1} - \nu_\tau^\star) + \tau \overline{\nu}_{t+1}^\top \ln \overline{\nu}_{t+1} - \tau \nu_\tau^{\star\top} \ln \nu_\tau^\star \\
& = (\mu - \mu_\tau^\star)^\top Q(\overline{\nu}_{t+1} - \nu_\tau^\star) + \mu_\tau^{\star\top} Q(\overline{\nu}_{t+1} - \nu_\tau^\star) + \tau \overline{\nu}_{t+1}^\top \ln \overline{\nu}_{t+1} - \tau \nu_\tau^{\star\top} \ln \nu_\tau^\star \\
& \overset{(i)}{\leq} \|\mu - \mu_\tau^\star\|_1 \|Q\|_\infty \|\overline{\nu}_{t+1} - \nu_\tau^\star\|_1 - \tau \big\langle \ln \nu_\tau^*, \overline{\nu}_{t+1} - \nu_\tau^* \big\rangle + \tau \overline{\nu}_{t+1}^\top \ln \overline{\nu}_{t+1} - \tau \nu_\tau^{\star\top} \ln \nu_\tau^\star \\
& \leq \frac{1}{2} \big[\tau \|\mu - \mu_\tau^\star\|_1^2 + \|Q\|_\infty^2 \|\overline{\nu}_{t+1} - \nu_\tau^\star\|_1^2 / \tau \big] + \tau KL(\overline{\nu}_{t+1} \| \nu_\tau^\star) \\
& \overset{(ii)}{\leq} \tau KL(\mu \| \mu_\tau^\star) + \tau KL(\overline{\nu}_{t+1} \| \nu_\tau^\star) + \tau^{-1} \|Q\|_\infty^2 KL(\nu_\tau^\star \| \overline{\nu}_{t+1}),
\end{aligned} \tag{48}$$

where (i) uses eq. (33) and (ii) uses the Pinsker's inequality. Then, eq. (48) & eq. (47) imply that

$$h_\tau(\mu, \overline{\nu}_{t+1}) - h_\tau(\mu_\tau^\star, \nu_\tau^\star) \leq \tau KL(\overline{\nu}_{t+1} \| \nu_\tau^\star) + \tau^{-1} \|Q\|_\infty^2 KL(\nu_\tau^\star \| \overline{\nu}_{t+1}), \tag{49}$$

where (i) uses eq. (44), (ii) uses the Pinker's inequality and $\eta \leq \big[2(\tau + \|Q\|_\infty)\big]^{-1}$.

Similarly, it can be proved that

$$h_\tau(\mu_\tau^\star, \nu_\tau^\star) - h_\tau(\overline{\mu}_{t+1}, \nu) \leq \tau KL(\overline{\mu}_{t+1} \| \mu_\tau^\star) + \tau^{-1} \|Q\|_\infty^2 KL(\mu_\tau^\star \| \overline{\mu}_{t+1}). \tag{50}$$

Therefore, the duality gap (29) can be proved as follows.

$$\begin{aligned}
& \max_{\mu \in \Delta(d_1)} h_\tau(\mu, \overline{\nu}_{t+1}) - \min_{\nu \in \Delta(d_2)} h_\tau(\overline{\mu}_{t+1}, \nu) \\
& = \max_{\mu \in \Delta(d_1), \nu \in \Delta(d_2)} \big[h_\tau(\mu, \overline{\nu}_{t+1}) - h_\tau(\overline{\mu}_{t+1}, \nu) \big] \\
& \overset{(i)}{\leq} \tau [KL(\overline{\mu}_{t+1} \| \mu_\tau^\star) + KL(\overline{\nu}_{t+1} \| \nu_\tau^\star)] + \tau^{-1} \|Q\|_\infty^2 [KL(\mu_\tau^\star \| \overline{\mu}_{t+1}) + KL(\nu_\tau^\star \| \overline{\nu}_{t+1})] \\
& \overset{(ii)}{\leq} \tau [KL(\overline{\mu}_{t+1} \| \mu_\tau^\star) + KL(\overline{\nu}_{t+1} \| \nu_\tau^\star)] + \tau^{-1} \|Q\|_\infty^2
\end{aligned}$$

$$
\begin{aligned}
&\Big[4\big[KL(\mu_\tau^*\|\mu_t)+KL(\nu_\tau^*\|\nu_t)\big]+24\eta^2\big(\|\delta_t^{(1)}\|_\infty^2+\|\delta_t^{(2)}\|_\infty^2\big)+\frac{16\eta}{\tau}\big(\|\overline{\delta}_t^{(1)}\|_\infty^2+\|\overline{\delta}_t^{(2)}\|_\infty^2\big)\Big]\\
&\overset{(iii)}{\leq}\max\Big(\frac{2}{\eta},\frac{4\|Q\|_\infty^2}{\tau}\Big)\\
&\quad\Big[KL(\mu_\tau^*\|\mu_t)+KL(\nu_\tau^*\|\nu_t)+2\eta^2\big(\|\delta_t^{(1)}\|_\infty^2+\|\delta_t^{(2)}\|_\infty^2\big)+\frac{2\eta}{\tau}\big(\|\overline{\delta}_t^{(1)}\|_\infty^2+\|\overline{\delta}_t^{(2)}\|_\infty^2\big)\Big]\\
&\quad+\tau^{-1}\|Q\|_\infty^2\Big[24\eta^2\big(\|\delta_t^{(1)}\|_\infty^2+\|\delta_t^{(2)}\|_\infty^2\big)+\frac{16\eta}{\tau}\big(\|\overline{\delta}_t^{(1)}\|_\infty^2+\|\overline{\delta}_t^{(2)}\|_\infty^2\big)\Big]\\
&\overset{(iv)}{\leq}\frac{2}{\eta}\Big((1-\eta\tau)^t\ln(d_1d_2)\\
&\quad+\sum_{j=0}^{t-1}(1-\eta\tau)^{t-1-j}\Big[2\eta^2\big(\|\delta_j^{(1)}\|_\infty^2+\|\delta_j^{(2)}\|_\infty^2\big)+\frac{2\eta}{\tau}\big(\|\overline{\delta}_j^{(1)}\|_\infty^2+\|\overline{\delta}_j^{(2)}\|_\infty^2\big)\Big]\Big)\\
&\quad+16\eta\big(\|\delta_t^{(1)}\|_\infty^2+\|\delta_t^{(2)}\|_\infty^2\big)+\frac{12}{\tau}\big(\|\overline{\delta}_t^{(1)}\|_\infty^2+\|\overline{\delta}_t^{(2)}\|_\infty^2\big)\\
&\leq\frac{2}{\eta}(1-\eta\tau)^t\ln(d_1d_2)\\
&\quad+\sum_{j=0}^{t}(1-\eta\tau)^{t-1-j}\Big[16\eta\big(\|\delta_j^{(1)}\|_\infty^2+\|\delta_j^{(2)}\|_\infty^2\big)+\frac{12}{\tau}\big(\|\overline{\delta}_j^{(1)}\|_\infty^2+\|\overline{\delta}_j^{(2)}\|_\infty^2\big)\Big],
\end{aligned}
\tag{51}
$$

where (i) adds up eqs. (49) & (50), (ii) uses eq. (46), (iii) uses eq. (38), and (iv) uses eq. (27) and $4\tau^{-1}\|Q\|_\infty^2\leq 2/\eta$ (since $\eta\leq\big[2(\tau+\|Q\|_\infty)\big]^{-1}$, $\tau\leq 1$). This proves eq. (29) □

B ESTIMATION ERROR BOUNDS

In this section, we derive error bounds for the estimators $\widehat{V}_{k+1}\approx V'_{k+1}$, $\widehat{Q}_{k,t}^{(m)}\approx Q_{k,t}^{(m)}$, $\widehat{\overline{Q}}_{k,t}^{(m)}\approx\overline{Q}_{k,t}^{(m)}$ used in Algorithm 1. For convenience, we define the following additional notations.

$$\widehat{\mu}_k(s):=\frac{1}{N_{k+1}}\sum_{i\in\mathcal{N}_{k+1}}\mathbb{1}\{s_i=s\}\approx\mu_k(s) \tag{52}$$

$$v_{k+1}(s):=\mathbb{E}_{\mu_k,\pi_k^{(1)},\pi_k^{(2)}}\Big[\big[R(\widetilde{s},\widetilde{a}^{(1)},\widetilde{a}^{(2)})+\gamma V'_k(s')\big]\mathbb{1}\{\widetilde{s}=s\}\Big] \tag{53}$$

$$v'_{k+1}(s):=\mathbb{E}_{\mu_k,\pi_k^{(1)},\pi_k^{(2)}}\Big[\big[R(\widetilde{s},\widetilde{a}^{(1)},\widetilde{a}^{(2)})+\gamma\widehat{V}_k(s')\big]\mathbb{1}\{\widetilde{s}=s\}\Big] \tag{54}$$

$$\widehat{v}_{k+1}(s):=\frac{1}{N_{k+1}}\sum_{i\in\mathcal{N}_{k+1}}\Big[\big[R_i+\gamma\widehat{V}_k(s_{i+1})\big]\mathbb{1}\{s_i=s\}\Big]\approx v'_{k+1}(s)\approx v_{k+1}(s) \tag{55}$$

$$Q_k(s,a^{(1)},a^{(2)}):=R(s,a^{(1)},a^{(2)})+\gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a^{(1)},a^{(2)})}\big[V'_k(s')\big] \tag{56}$$

$$V'_{k+1}(s)=f_\tau\big(Q_k(s);\pi_k^{(1)}(s),\pi_k^{(2)}(s)\big)=\frac{v_{k+1}(s)}{\mu_k(s)}+\tau\mathcal{H}\big(\pi_k^{(1)}(s)\big)-\tau\mathcal{H}\big(\pi_k^{(2)}(s)\big) \tag{57}$$

$$\widehat{V}_{k+1}(s):=\frac{\widehat{v}_{k+1}(s)}{\widehat{\mu}_k(s)}+\tau\mathcal{H}\big(\pi_k^{(1)}(s)\big)-\tau\mathcal{H}\big(\pi_k^{(2)}(s)\big)\approx V'_{k+1}(s) \tag{58}$$

$$\widehat{\mu}_{k,t}(s):=\frac{1}{N_{k,t}}\sum_{i\in\mathcal{N}_{k,t}}\mathbb{1}\{s_i=s\}\approx\mu'_{k,t}(s)\approx\mu_{k,t}(s), \tag{59}$$

$$q_{k,t}^{(m)}(s,a^{(m)}):=\mathbb{E}_{\mu'_{k,t},\pi_{k,t}^{\prime(1)},\pi_{k,t}^{\prime(2)}}\Big[\big[R(\widetilde{s},\widetilde{a}^{(1)},\widetilde{a}^{(2)})+\gamma V'_k(s')\big]\mathbb{1}\{\widetilde{s}=s,\widetilde{a}^{(m)}=a^{(m)}\}\Big] \tag{60}$$

$$q_{k,t}^{\prime(m)}(s,a^{(m)}):=\mathbb{E}_{\mu'_{k,t},\pi_{k,t}^{\prime(1)},\pi_{k,t}^{\prime(2)}}\Big[\big[R(\widetilde{s},\widetilde{a}^{(1)},\widetilde{a}^{(2)})+\gamma\widehat{V}_k(s')\big]\mathbb{1}\{\widetilde{s}=s,\widetilde{a}^{(m)}=a^{(m)}\}\Big] \tag{61}$$

$$
\begin{aligned}
\widehat{q}_{k,t}^{(m)}(s,a^{(m)})&:=\frac{1}{N_{k,t}}\sum_{i\in\mathcal{N}_{k,t}}\Big[\big[R_i+\gamma\widehat{V}_k(s_{i+1})\big]\mathbb{1}\{s_i=s,a_i^{(m)}=a^{(m)}\}\Big]\\
&\approx q_{k,t}^{\prime(m)}(s,a^{(m)})\approx q_{k,t}^{(m)}(s,a^{(m)})
\end{aligned}
\tag{62}
$$

$$Q'^{(m)}_{k,t}(s,a^{(m)}) := \mathbb{E}_{a^{(\setminus m)}\sim\pi'^{(\setminus m)}_{k,t}(s)}\big[Q_k(s,a^{(1)},a^{(2)})\big] = \frac{q^{(m)}_{k,t}(s,a^{(m)})}{\mu'_{k,t}(s)\pi'^{(m)}_{k,t}(a^{(m)}|s)} \tag{63}$$

$$\widehat{Q}^{(m)}_{k,t}(s,a^{(m)}) := \frac{\widehat{q}^{(m)}_{k,t}(s,a^{(m)})}{\widehat{\mu}_{k,t}(s)\pi'^{(m)}_{k,t}(a^{(m)}|s)} \approx Q'^{(m)}_{k,t}(s,a^{(m)}) \approx Q^{(m)}_{k,t}(s,a^{(m)}), \tag{64}$$

where eqs. (52)-(58) and (59)-(64) are introduced for the description and error bound proof of the estimations $\widehat{V}_{k+1} \approx V'_{k+1}$ and $\widehat{Q}^{(m)}_{k,t} \approx Q^{(m)}_{k,t}$, respectively. Specifically, eq. (52) estimates the stationary state distribution μ_k associated with the policy pair $\pi^{(1)}_k, \pi^{(2)}_k$ using state frequency $\widehat{\mu}_k(s)$. In eqs. (53)&(54), the expectation is taken over $\widetilde{s}\sim\mu_k, \widetilde{a}^{(1)}\sim\pi^{(1)}_k(s), \widetilde{a}^{(2)}\sim\pi^{(2)}_k(s), s'\sim\mathcal{P}(\cdot|\widetilde{s},\widetilde{a}^{(1)},\widetilde{a}^{(2)})$. The definition of $\mathbb{E}_{\mu'_{k,t},\pi'^{(1)}_{k,t},\pi'^{(2)}_{k,t}}$ in eqs. (60)&(61) is similar. Equations (57) & (63) give two equivalent definitions of V'_{k+1} and $Q'^{(m)}_{k,t}(s,a^{(m)})$, respectively [1] (We will prove the equivalence in Lemma 2 below). Equation (59) estimates the stationary state distribution $\mu'_{k,t}$ associated with the smoothed policy pair $\pi'^{(1)}_{k,t}, \pi'^{(2)}_{k,t}$, and thus approximates the stationary state distribution $\mu_{k,t}$ associated with the current policy pair $\pi^{(1)}_{k,t}, \pi^{(2)}_{k,t}$.

To prove the estimation error bounds, we first prove the following two lemmas.

Lemma 2. *Eq. (13) and the second "=" of eqs. (57) & (63) hold for all s, $a^{(1)}, a^{(2)}$.*

Proof. The second "=" of eq. (63) under $m=1$ can be proved as follows.

$$\begin{aligned}
&q^{(1)}_{k,t}(s,a^{(1)})\\
&\overset{(i)}{=} \mathbb{E}_{\mu'_{k,t},\pi'^{(1)}_{k,t},\pi'^{(2)}_{k,t}}\Big[\big[R(\widetilde{s},\widetilde{a}^{(1)},\widetilde{a}^{(2)}) + \gamma V'_k(s')\big]\mathbb{1}\{\widetilde{s}=s,\widetilde{a}^{(m)}=a^{(m)}\}\Big]\\
&= \mathbb{E}_{\widetilde{s}\sim\mu'_{k,t},\widetilde{a}^{(1)}\sim\pi'^{(1)}_{k,t}(s),\widetilde{a}^{(2)}\sim\pi'^{(2)}_{k,t}(s),s'\sim\mathcal{P}(\cdot|s,a^{(1)},\widetilde{a}^{(2)})}\\
&\quad\Big[\big[R(s,a^{(1)},\widetilde{a}^{(2)}) + \gamma V'_k(s')\big]\mathbb{1}\{\widetilde{s}=s,\widetilde{a}^{(1)}=a^{(1)}\}\Big]\\
&\overset{(ii)}{=} \mathbb{E}_{\widetilde{s}\sim\mu'_{k,t},\widetilde{a}^{(1)}\sim\pi'^{(1)}_{k,t}(s)}\big[\mathbb{1}\{\widetilde{s}=s,\widetilde{a}^{(1)}=a^{(1)}\}\big]\\
&\quad\mathbb{E}_{\widetilde{a}^{(2)}\sim\pi'^{(2)}_{k,t}(s),s'\sim\mathcal{P}(\cdot|s,a^{(1)},\widetilde{a}^{(2)})}\big[R(s,a^{(1)},\widetilde{a}^{(2)}) + \gamma V'_k(s')\big]\\
&\overset{(iii)}{=} \mu'_{k,t}(s)\pi'^{(1)}_{k,t}(a^{(1)}|s)\mathbb{E}_{\widetilde{a}^{(2)}\sim\pi'^{(2)}_{k,t}(s)}\big[Q_k(s,a^{(1)},\widetilde{a}^{(2)})\big]\\
&\overset{(iv)}{=} \mu'_{k,t}(s)\pi'^{(1)}_{k,t}(a^{(1)}|s)Q'_{k,t}(s,a^{(1)}),
\end{aligned}$$

where (i) uses eq. (60) which denotes $\mathbb{E}_{\mu'_{k,t},\pi'^{(1)}_{k,t},\pi'^{(2)}_{k,t}}$ as expectation over $\widetilde{s}\sim\mu'_{k,t}$, $\widetilde{a}^{(1)}\sim\pi'^{(1)}_{k,t}(\widetilde{s})$, $\widetilde{a}^{(2)}\sim\pi'^{(2)}_{k,t}(\widetilde{s})$, $s'\sim\mathcal{P}(\cdot|\widetilde{s},\widetilde{a}^{(1)},\widetilde{a}^{(2)})$, (ii) uses independence between $\widetilde{s}\sim\mu'_{k,t}, \widetilde{a}^{(1)}\sim\pi'^{(1)}_{k,t}(s)$ and $\widetilde{a}^{(2)}\sim\pi'^{(2)}_{k,t}(s), s'\sim\mathcal{P}(\cdot|s,a^{(1)},\widetilde{a}^{(2)})$, (iii) uses eq. (56), and (iv) uses the first definition of eq. (63). The second "=" of eq. (63) under $m=2$ and eq. (13) can be proved in a similar way.

The second "=" of eq. (57) can be proved as follows.

$$\begin{aligned}
&v_{k+1}(s)\\
&\overset{(i)}{=} \mathbb{E}_{\mu_k,\pi^{(1)}_k,\pi^{(2)}_k}\Big[\big[R(\widetilde{s},\widetilde{a}^{(1)},\widetilde{a}^{(2)}) + \gamma V'_k(s')\big]\mathbb{1}\{\widetilde{s}=s\}\Big]\\
&= \mathbb{E}_{\widetilde{s}\sim\mu_k,\widetilde{a}^{(1)}\sim\pi^{(1)}_k(s),\widetilde{a}^{(2)}\sim\pi^{(2)}_k(s),s'\sim\mathcal{P}(\cdot|s,\widetilde{a}^{(1)},\widetilde{a}^{(2)})}\Big[\big[R(s,\widetilde{a}^{(1)},\widetilde{a}^{(2)}) + \gamma V'_k(s')\big]\mathbb{1}\{\widetilde{s}=s\}\Big]\\
&\overset{(ii)}{=} \mathbb{E}_{\widetilde{s}\sim\mu_k}\big[\mathbb{1}\{\widetilde{s}=s\}\big]\mathbb{E}_{\widetilde{a}^{(1)}\sim\pi^{(1)}_k(s),\widetilde{a}^{(2)}\sim\pi^{(2)}_k(s),s'\sim\mathcal{P}(\cdot|s,\widetilde{a}^{(1)},\widetilde{a}^{(2)})}\big[R(s,\widetilde{a}^{(1)},\widetilde{a}^{(2)}) + \gamma V'_k(s')\big]\\
&\overset{(iii)}{=} \mu_k(s)\mathbb{E}_{\widetilde{a}^{(1)}\sim\pi^{(1)}_k(s),\widetilde{a}^{(2)}\sim\pi^{(2)}_k(s)}\big[Q_k(s,\widetilde{a}^{(1)},\widetilde{a}^{(2)})\big]
\end{aligned}$$

[1] The definition after the second "=" of eq. (57) is the same as eq. (11)

$$= \mu_{k,t}(s)\big[f_\tau\big(Q_k(s);\pi_k^{(1)}(s),\pi_k^{(2)}(s)\big) - \tau\mathcal{H}\big(\pi_k^{(1)}(s)\big) + \tau\mathcal{H}\big(\pi_k^{(2)}(s)\big)\big]$$
$$\overset{(iv)}{=} \mu_k(s)\big[V'_{k+1}(s) - \tau\mathcal{H}\big(\pi_k^{(1)}(s)\big) + \tau\mathcal{H}\big(\pi_k^{(2)}(s)\big)\big]$$

where (i) uses eq. (53), (ii) uses independence between $s' \sim \mu_k$ and $\widetilde{a}^{(1)} \sim \pi_k^{(1)}(s), \widetilde{a}^{(2)} \sim \pi_k^{(2)}(s), s' \sim \mathcal{P}(\cdot|s,\widetilde{a}^{(1)},\widetilde{a}^{(2)})$, (iii) uses eq. (56), and (iv) uses eq. (57). □

Lemma 3. *If* $\|\widehat{V}_0\|_\infty \le V_{\max} := \frac{1+\tau\ln A_{\max}}{1-\gamma}$, *then for all* $k \ge 0$, $0 \le t \le T_k - 1$, *we have*

$$\|\widehat{V}_k\|_\infty, \|V'_k\|_\infty, \|V^*\|_\infty \le V_{\max}, \tag{65}$$
$$|\widehat{v}_{k+1}(s)| \le 2V_{\max}\widehat{\mu}_k(s), \forall s \in \mathcal{S}, \tag{66}$$
$$\max\big(\|Q_k\|_\infty, \|Q_{k,t}'^{(m)}\|_\infty, \|Q^*\|_\infty\big) \le Q_{\max}, \tag{67}$$

where $Q_{\max} := 1 + \gamma V_{\max} = \frac{1+\gamma\tau\ln A_{\max}}{1-\gamma}$ *with* $A_{\max} := \max\big(|\mathcal{A}^{(1)}|, |\mathcal{A}^{(2)}|\big)$.

Proof. We will first prove $\|\widehat{V}_k\|_\infty \le V_{\max}$ in eq. (65) by induction. Suppose $\|\widehat{V}_{k'}\|_\infty \le V_{\max}$ holds for a certain $k' \in \mathbb{N}$. Then, eq. (58) implies that for all $s \in \mathcal{S}$,

$$\begin{aligned}
|\widehat{V}_{k'+1}(s)| \le& \Big|\frac{\widehat{v}_{k'+1}(s)}{\widehat{\mu}'_{k'}(s)}\Big| + \tau\big[\mathcal{H}\big(\pi_{k'}^{(1)}(s)\big) - \mathcal{H}\big(\pi_{k'}^{(2)}(s)\big)\big] \\
\le& \frac{\sum_{i\in\mathcal{N}_{k'+1}}\Big[\big[|R_i| + \gamma|\widehat{V}_{k'}(s_{i+1})|\big]\mathbb{1}\{s_i = s\}\Big]}{\sum_{i\in\mathcal{N}_{k'+1}}\mathbb{1}\{s_i = s\}} + \tau\ln A_{\max} \\
\overset{(i)}{\le}& \frac{\sum_{i\in\mathcal{N}_{k'+1}}\Big[\big(1+\gamma V_{\max}\big)\mathbb{1}\{s_i = s\}\Big]}{\sum_{i\in\mathcal{N}_{k'+1}}\mathbb{1}\{s_i = s\}} + \tau\ln A_{\max} \\
=& 1 + \frac{\gamma(1+\tau\ln A_{\max})}{1-\gamma} + \tau\ln A_{\max} = V_{\max},
\end{aligned} \tag{68}$$

where (i) uses the inequality that $0 \le \mathcal{H}(\pi^{(m)}) \le \ln|\mathcal{A}^{(m)}| \le \ln A_{\max}, \forall\pi^{(m)} \in \Delta(|\mathcal{A}^{(m)}|)$. Since $\|\widehat{V}_0\|_\infty \le V_{\max}$, $\|\widehat{V}_k\|_\infty \le V_{\max}$ for all $k \in \mathbb{N}$. A similar induction yields that $\|V'_k\|_\infty \le V_{\max}$.

Next, we prove $\|V^*\|_\infty \le V_{\max}$, $\|Q^*\|_\infty \le Q_{\max}$ in eqs. (65) & (67) respectively. Notice that V^* and Q^* have the following relation.

$$Q^*(s,a^{(1)},a^{(2)}) = R(s,a^{(1)},a^{(2)}) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a^{(1)},a^{(2)})}V^*(s'),$$
$$V^*(s) = \pi^{*(1)}(s)^\top Q^*(s)\pi^{*(2)}(s) + \tau\big[\mathcal{H}(\pi^{*(1)}(s)) - \mathcal{H}(\pi^{*(2)}(s))\big].$$

Hence,

$$\|Q^*\|_\infty \le 1 + \gamma\|V^*\|_\infty \tag{69}$$
$$\|V^*\|_\infty \le \|Q^*\|_\infty + \tau\ln A_{\max} \tag{70}$$

Substituting eq. (69) into eq. (70) yields that

$$\|V^*\|_\infty \le 1 + \gamma\|V^*\|_\infty + \tau\ln A_{\max} \Rightarrow \|V^*\|_\infty \le V_{\max}.$$

The proof of eq. (65) is finished. Then, substituting $\|V^*\|_\infty \le V_{\max}$ into eq. (69) proves that $\|Q^*\|_\infty \le Q_{\max}$.

Equation (66) can be proved as follows.

$$\Big|\frac{\widehat{v}_{k+1}(s)}{\widehat{\mu}_k(s)}\Big| \overset{(i)}{\le} |\widehat{V}_{k+1}(s)| + \tau\big|\mathcal{H}\big(\pi_k^{(2)}(s)\big) - \mathcal{H}\big(\pi_k^{(1)}(s)\big)\big| \le V_{\max} + \tau\ln A_{\max} \overset{(ii)}{\le} 2V_{\max},$$

where (i) uses eq. (57) and (ii) uses $V_{\max} := \frac{1+\tau\ln A_{\max}}{1-\gamma} \ge \tau\ln A_{\max}$.

Next, we prove eq. (67). It can be easily seen from eqs. (56) & (65) that

$$\big|Q_k(s,a^{(1)},a^{(2)})\big| \le \big|R(s,a^{(1)},a^{(2)})\big| + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a^{(1)},a^{(2)})}\big|V'_k(s')\big| \le 1 + \gamma V_{\max} = Q_{\max},$$

which implies that $\|Q_k\|_\infty \leq Q_{\max}$. Hence,

$$\left|Q'^{(m)}_{k,t}(s,a^{(m)})\right| \overset{(i)}{\leq} \mathbb{E}_{a^{(\setminus m)}\sim\pi'^{(\setminus m)}_{k,t}(s)}\left|Q_k(s,a^{(1)},a^{(2)})\right| \leq Q_{\max}$$

where (i) uses eq. (63). This proves eq. (67). □

With the above two Lemmas, we can derive the estimation error bounds as follows.

Lemma 4. *Use hyperparameter choices* $N_{k,t}, \overline{N}_{k,t} \geq \frac{650 t_{mix} A_{\max}}{\mu_{\min}} \ln\left(\frac{20 T_{sum}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\right)$ *and* $N_{k+1} \geq \frac{650 t_{mix}}{\mu_{\min}(1-\gamma)^2} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)$ *in Algorithm 1. Then with probability at least* $1-\delta$*, all the following bounds hold for all* $0 \leq k \leq K-1$*,* $0 \leq t \leq T_k - 1$*,* $m = 1,2$*,* $s \in \mathcal{S}$*,* $a^{(m)} \in \mathcal{A}^{(m)}$*, where* $T_{sum} := \sum_{k=0}^{K-1} T_k$*.*

$$\left\|\widehat{V}_k - V'_k\right\|_\infty \leq 2V_{\max}\gamma^k + \frac{171 V_{\max}}{1-\gamma}\sqrt{\frac{t_{mix}}{N_{k+1}\mu_{\min}} \ln\left(\frac{4K|\mathcal{S}|}{\delta\sqrt{\mu_{\min}}}\right)} \tag{71}$$

$$\begin{aligned}
&\left\|\widehat{Q}^{(m)}_{k,t} - Q^{(m)}_{k,t}\right\|_\infty \\
&\leq Q_{\max}\left[\frac{640 t_{mix} A_{\max}}{N_{k,t}\mu_{\min}\epsilon'} \ln\left(\frac{20 T_{sum}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\right) + 25\sqrt{\frac{t_{mix}A_{\max}}{N_{k,t}\mu_{\min}\epsilon'} \ln\left(\frac{20 T_{sum}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\right)} + \epsilon'\right] \\
&\quad + 2\left\|\widehat{V}_k - V'_k\right\|_\infty
\end{aligned} \tag{72}$$

$$\begin{aligned}
&\left\|\widehat{\overline{Q}}^{(m)}_{k,t} - \overline{Q}^{(m)}_{k,t}\right\|_\infty \\
&\leq Q_{\max}\left[\frac{640 t_{mix} A_{\max}}{\overline{N}_{k,t}\mu_{\min}\overline{\epsilon}'} \ln\left(\frac{20 T_{sum}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\right) + 25\sqrt{\frac{t_{mix}A_{\max}}{\overline{N}_{k,t}\mu_{\min}\overline{\epsilon}'} \ln\left(\frac{20 T_{sum}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\right)} + \overline{\epsilon}'\right] \\
&\quad + 2\left\|\widehat{V}_k - V'_k\right\|_\infty.
\end{aligned} \tag{73}$$

More specifically, the upper bounds (72)&(73) can be simplified respectively as follows when $\epsilon' = \left[\frac{650 t_{mix} A_{\max}}{N_{k,t}\mu_{\min}} \ln\left(\frac{20 T_{sum}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\right)\right]^{1/3}$ *and* $\overline{\epsilon}' = \left[\frac{650 t_{mix} A_{\max}}{\overline{N}_{k,t}\mu_{\min}} \ln\left(\frac{20 T_{sum}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\right)\right]^{1/3}$*.*

$$\left\|\widehat{Q}^{(m)}_{k,t} - Q^{(m)}_{k,t}\right\|_\infty \leq 18 Q_{\max}\left[\frac{t_{mix}A_{\max}}{N_{k,t}\mu_{\min}} \ln\left(\frac{20 T_{sum}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\right)\right]^{1/3} + 2\left\|\widehat{V}_k - V'_k\right\|_\infty, \tag{74}$$

$$\left\|\widehat{\overline{Q}}^{(m)}_{k,t} - \overline{Q}^{(m)}_{k,t}\right\|_\infty \leq 18 Q_{\max}\left[\frac{t_{mix}A_{\max}}{\overline{N}_{k,t}\mu_{\min}} \ln\left(\frac{20 T_{sum}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\right)\right]^{1/3} + 2\left\|\widehat{V}_k - V'_k\right\|_\infty. \tag{75}$$

Proof. If the initial state distribution of the minibatch $\mathcal{N}_{k,t}$ equals $\mu'_{k,t}$, i.e., $s_{k,t,0} \sim \mu'_{k,t}$, then $N_{k,t}\widehat{q}^{(m)}_{k,t}(s,a^{(m)}) := \sum_{i\in\mathcal{N}_{k,t}} g(s_i, a_i^{(1)}, a_i^{(2)}, s_{i+1})$ $(g(s_i, a_i^{(1)}, a_i^{(2)}, s_{i+1}) := \left[R_i + \gamma\widehat{V}_k(s_{i+1})\right]\mathbb{1}\{s_i = s, a_i^{(m)} = a^{(m)}\})$ and its expected value $N_{k,t}q'^{(m)}_{k,t}(s,a^{(m)})$ satisfy the following concentration bound.

$$\begin{aligned}
&\mathbb{P}_{\mu'_{k,t}}\left\{\left|N_{k,t}\widehat{q}^{(m)}_{k,t}(s,a^{(m)}) - N_{k,t}q'^{(m)}_{k,t}(s,a^{(m)})\right| \geq u\right\} \\
&\overset{(i)}{\leq} 2\exp\left[-\frac{u^2\gamma_{ps}}{8(N_{k,t}+1/\gamma_{ps})Q^2_{\max}\mu'_{k,t}(s)\pi'^{(m)}_{k,t}(a^{(m)}|s) + 40uQ_{\max}}\right] \\
&\overset{(ii)}{\leq} 2\exp\left[-\frac{u^2/(2t_{\text{mix}})}{8(N_{k,t}+2t_{\text{mix}})Q^2_{\max}\mu'_{k,t}(s)\pi'^{(m)}_{k,t}(a^{(m)}|s) + 40uQ_{\max}}\right],
\end{aligned} \tag{76}$$

where $\mathbb{P}_{\mu'_{k,t}}$ is under the initial state distribution $\mu'_{k,t}$, (i) uses Theorem 3.4 of (Paulin, 2015) and the inequalities that $\left|g(s_i, a_i^{(1)}, a_i^{(2)}, s_{i+1}) - \mathbb{E}_{s_i\sim\mu'_{k,t}} g(s_i, a_i^{(1)}, a_i^{(2)}, s_{i+1})\right| \leq 2(1+\gamma V_{\max}) = 2Q_{\max}$ (Based on Lemma 3) and that $\text{var}_{s_i\sim\mu'_{k,t}}\left[g(s_i, a_i^{(1)}, a_i^{(2)}, s_{i+1})\right] \leq$

$\mathbb{E}_{s_i\sim\mu'_{k,t}}\Big[\big[R_i+\gamma V'_k(s_{i+1})\big]^2\mathbb{1}\{s_i=s,a_i^{(m)}=a^{(m)}\}\Big]\le(1+\gamma V_{\max})^2\mu'_{k,t}(s)\pi'^{(m)}_{k,t}(a^{(m)}|s)=Q^2_{\max}\mu'_{k,t}(s)\pi'^{(m)}_{k,t}(a^{(m)}|s)$, (ii) uses Proposition 3.4 of (Paulin, 2015) which states that the pseudo spectral gap γ_{ps} has a lower bound $1/(2t_{\text{mix}})$ for any uniformly ergodic Markov chain (This condition holds for our MDP with finitely many states and actions). Then, for any initial state distribution $s_{t,k,0}\sim\xi$, Proposition 3.10 of (Paulin, 2015) implies that

$$\begin{aligned}&\mathbb{P}_\xi\big\{\big|N_{k,t}\widehat{q}^{(m)}_{k,t}(s,a^{(m)})-N_{k,t}q'^{(m)}_{k,t}(s,a^{(m)})\big|\ge u\big\}\\&\le\sqrt{\mathbb{E}_{s\sim\xi}\Big[\frac{\xi(s)}{\mu'_{k,t}(s)}\Big]\mathbb{P}_{\mu'_{k,t}}\big\{\big|N_{k,t}\widehat{q}^{(m)}_{k,t}(s,a^{(m)})-N_{k,t}q'^{(m)}_{k,t}(s,a^{(m)})\big|\ge u\big\}}\\&\overset{(i)}{\le}\sqrt{\frac{1}{\mu_{\min}}}\exp\Big[-\frac{u^2/(4t_{\text{mix}})}{8(N_{k,t}+2t_{\text{mix}})Q^2_{\max}\mu'_{k,t}(s)\pi'^{(m)}_{k,t}(a^{(m)}|s)+40uQ_{\max}}\Big],\end{aligned}\tag{77}$$

where (i) uses eq. (76) and the inequality that $\mathbb{E}_{s\sim\xi}\Big[\frac{\xi(s)}{\mu'_{k,t}(s)}\Big]\le\frac{1}{\mu_{\min}}$.

In a similar way, the following concentration bound can be proved for $N_{k,t}\widehat{\mu}_{k,t}(s):=\sum_{i\in\mathcal{N}_{k,t}}\mathbb{1}\{s_i=s\}$ and $N_{k,t}\mu'_{k,t}(s)=\mathbb{E}_{\mu'_{k,t}}\big[N_{k,t}\widehat{\mu}_{k,t}(s)\big]$.

$$\mathbb{P}_\xi\big\{\big|N_{k,t}\widehat{\mu}_{k,t}(s)-N_{k,t}\mu'_{k,t}(s)\big|\ge u\big\}\le\sqrt{\frac{1}{\mu_{\min}}}\exp\Big[-\frac{u^2/(4t_{\text{mix}})}{8(N_{k,t}+2t_{\text{mix}})\mu'_{k,t}(s)+40u}\Big],\tag{78}$$

where we use the inequalities that $\text{var}_{\mu'_{k,t}}\mathbb{1}\{s_i=s\}\le\mu'_{k,t}(s)$ and that $\big|\mathbb{1}\{s_i=s\}-\mathbb{E}_{\mu'_{k,t}}\mathbb{1}\{s_i=s\}\big|\le1$. Letting the right hand sides of eqs. (77)&(78) be upper bounded by $\delta/4$ and applying the union bound yields that, with probability at least $1-\delta/2$, the following two inequalities simultaneously hold.

$$\begin{aligned}&\big|\widehat{q}^{(m)}_{k,t}(s,a^{(m)})-q'^{(m)}_{k,t}(s,a^{(m)})\big|\\&\le\frac{160Q_{\max}}{N_{k,t}}t_{\text{mix}}\ln\Big(\frac{4}{\delta\sqrt{\mu_{\min}}}\Big)\\&\quad+\frac{6Q_{\max}}{\sqrt{N_{k,t}}}\sqrt{t_{\text{mix}}(1+2t_{\text{mix}}/N_{k,t})\mu'_{k,t}(s)\pi'^{(m)}_{k,t}(a^{(m)}|s)\ln\Big(\frac{4}{\delta\sqrt{\mu_{\min}}}\Big)},\end{aligned}\tag{79}$$

$$\begin{aligned}&\big|\widehat{\mu}_{k,t}(s)-\mu'_{k,t}(s)\big|\\&\le\frac{160}{N_{k,t}}t_{\text{mix}}\ln\Big(\frac{4}{\delta\sqrt{\mu_{\min}}}\Big)\\&\quad+\frac{6}{\sqrt{N_{k,t}}}\sqrt{t_{\text{mix}}(1+2t_{\text{mix}}/N_{k,t})\mu'_{k,t}(s)\ln\Big(\frac{4}{\delta\sqrt{\mu_{\min}}}\Big)}\overset{(i)}{\le}\frac{1}{2}\mu'_{k,t}(s),\end{aligned}\tag{80}$$

where (i) holds since $N_{k,t}\ge\frac{650t_{\text{mix}}A_{\max}}{\mu_{\min}}\ln\Big(\frac{20T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\Big)\ge650t_{\text{mix}}$ and $\mu'_{k,t}(s)\ge\mu_{\min}$.

Similarly, it can be proved that the following two inequalities holds with probability at least $1-\delta/2$.

$$\begin{aligned}&\big|\widehat{v}_{k+1}(s)-v'_{k+1}(s)\big|\\&\le\frac{160Q_{\max}}{N_{k+1}}t_{\text{mix}}\ln\Big(\frac{4}{\delta\sqrt{\mu_{\min}}}\Big)+\frac{6Q_{\max}}{\sqrt{N_{k+1}}}\sqrt{t_{\text{mix}}(1+2t_{\text{mix}}/N_{k+1})\mu_k(s)\ln\Big(\frac{4}{\delta\sqrt{\mu_{\min}}}\Big)},\end{aligned}\tag{81}$$

$$\begin{aligned}&\big|\widehat{\mu}_k(s)-\mu_k(s)\big|\\&\le\frac{160}{N_{k+1}}t_{\text{mix}}\ln\Big(\frac{4}{\delta\sqrt{\mu_{\min}}}\Big)+\frac{6}{\sqrt{N_{k+1}}}\sqrt{t_{\text{mix}}(1+2t_{\text{mix}}/N_{k+1})\mu_k(s)\ln\Big(\frac{4}{\delta\sqrt{\mu_{\min}}}\Big)}.\end{aligned}\tag{82}$$

Hence, eqs. (79)-(82) hold with probability at least $1-\delta$. In this case, we have that

$$\big|\widehat{v}_{k+1}(s)-v_{k+1}(s)\big|$$

$$
\begin{aligned}
&\leq \left|\widehat{v}_{k+1}(s) - v'_{k+1}(s)\right| + \left|v'_{k+1}(s) - v_{k+1}(s)\right| \\
&\overset{(i)}{\leq} \frac{160 Q_{\max}}{N_{k+1}} t_{\text{mix}} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right) + \frac{6Q_{\max}}{\sqrt{N_{k+1}}}\sqrt{t_{\text{mix}}(1+2t_{\text{mix}}/N_{k+1})\mu_k(s)\ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)} \\
&\quad + \gamma \mathbb{E}_{\mu_k, \pi_k^{(1)}, \pi_k^{(2)}}\left[\left[\widehat{V}_k(s') - V'_k(s')\right]\mathbb{1}\{\widetilde{s} = s\}\right] \\
&\leq \frac{160 Q_{\max}}{N_{k+1}} t_{\text{mix}} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right) + \frac{6Q_{\max}}{\sqrt{N_{k+1}}}\sqrt{t_{\text{mix}}(1+2t_{\text{mix}}/N_{k+1})\mu_k(s)\ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)} \\
&\quad + \gamma \max_{s'}\left|\widehat{V}_k(s') - V'_k(s')\right| \mathbb{E}_{\mu_k}\mathbb{1}\{\widetilde{s} = s\} \\
&\leq \frac{160 Q_{\max}}{N_{k+1}} t_{\text{mix}} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right) + \frac{6Q_{\max}}{\sqrt{N_{k+1}}}\sqrt{t_{\text{mix}}(1+2t_{\text{mix}}/N_{k+1})\mu_k(s)\ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)} \\
&\quad + \gamma \mu_k(s)\left\|\widehat{V}_k - V'_k\right\|_\infty, \qquad (83)
\end{aligned}
$$

where (i) uses eqs. (53), (54) & (81).

$$
\begin{aligned}
&|\widehat{V}_{k+1}(s) - V'_{k+1}(s)| \\
&\overset{(i)}{=} \left|\frac{\widehat{v}_{k+1}(s)}{\widehat{\mu}_k(s)} - \frac{v_{k+1}(s)}{\mu_k(s)}\right| \\
&\leq \mu_k^{-1}(s)\left|\widehat{v}_{k+1}(s) - v_{k+1}(s)\right| + \left|\widehat{v}_{k+1}(s)\right|\left|\frac{\mu_k(s) - \widehat{\mu}_k(s)}{\mu_k(s)\widehat{\mu}_k(s)}\right| \\
&\overset{(ii)}{\leq} \mu_k^{-1}(s)\Big[\frac{160 Q_{\max}}{N_{k+1}} t_{\text{mix}} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right) + \frac{6Q_{\max}}{\sqrt{N_{k+1}}}\sqrt{t_{\text{mix}}(1+2t_{\text{mix}}/N_{k+1})\mu_k(s)\ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)} \\
&\quad + \gamma\mu_k(s)\left\|\widehat{V}_k - V'_k\right\|_\infty\Big] + 2V_{\max}\mu_k^{-1}(s)\Bigg[\frac{160}{N_{k+1}} t_{\text{mix}} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right) \\
&\quad + \frac{6}{\sqrt{N_{k+1}}}\sqrt{t_{\text{mix}}(1+2t_{\text{mix}}/N_{k+1})\mu_k(s)\ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)}\Bigg] \\
&\overset{(iii)}{\leq} \gamma\|\widehat{V}_k - V'_k\|_\infty + \frac{V_{\max}}{\mu_k(s)}\Bigg[\frac{480 t_{\text{mix}}}{N_{k+1}} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right) \\
&\quad + 18\sqrt{\frac{t_{\text{mix}}\mu_k(s)}{N_{k+1}}(1+2t_{\text{mix}}/N_{k+1})\ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)}\Bigg] \\
&\overset{(iv)}{\leq} \gamma\|\widehat{V}_k - V'_k\|_\infty + V_{\max}\left[24\frac{40 t_{\text{mix}}}{N_{k+1}\mu_{\min}} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right) + 18\sqrt{\frac{1.05 t_{\text{mix}}}{N_{k+1}\mu_{\min}} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)}\right] \\
&\overset{(v)}{\leq} \gamma\|\widehat{V}_k - V'_k\|_\infty + V_{\max}\left[24\sqrt{\frac{40 t_{\text{mix}}}{N_{k+1}\mu_{\min}} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)} + 18\sqrt{\frac{1.05 t_{\text{mix}}}{N_{k+1}\mu_{\min}} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)}\right] \\
&\leq \gamma\|\widehat{V}_k - V'_k\|_\infty + 171 V_{\max}\sqrt{\frac{t_{\text{mix}}}{N_{k+1}\mu_{\min}} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)}, \qquad (84)
\end{aligned}
$$

where (i) uses eqs. (57)&(58), (ii) uses eqs. (66), (82)&(83), (iii) uses $Q_{\max} \leq V_{\max}$, (iv) uses $\mu_k(s) \geq \mu_{\min}$ and $N_{k+1} \geq \frac{40 t_{\text{mix}}}{\mu_{\min}} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right) \geq 40 t_{\text{mix}}$, and (v) uses $N_{k+1} \geq \frac{40 t_{\text{mix}}}{\mu_{\min}} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)$. Applying the union bound to the above inequality over all $0 \leq k \leq K-1$, $s \in \mathcal{S}$ and taking maximum over $s \in \mathcal{S}$, we obtain that with probability at least $1-\delta$,

$$
\|\widehat{V}_{k+1} - V'_{k+1}\|_\infty \leq \gamma\|\widehat{V}_k - V'_k\|_\infty + 171 V_{\max}\sqrt{\frac{t_{\text{mix}}}{N_{k+1}\mu_{\min}} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)}.
$$

Iterating the above inequality yields that with probability at least $1-\delta$,

$$\begin{aligned}\|\widehat{V}_k - V'_k\|_\infty \leq & \gamma^k \|\widehat{V}_0 - V'_0\|_\infty + \frac{171 V_{\max}}{1-\gamma}\sqrt{\frac{t_{\text{mix}}}{N_{k+1}\mu_{\min}} \ln\left(\frac{4K|\mathcal{S}|}{\delta\sqrt{\mu_{\min}}}\right)} \\ \overset{(i)}{\leq} & 2V_{\max}\gamma^k + \frac{171 V_{\max}}{1-\gamma}\sqrt{\frac{t_{\text{mix}}}{N_{k+1}\mu_{\min}} \ln\left(\frac{4K|\mathcal{S}|}{\delta\sqrt{\mu_{\min}}}\right)},\end{aligned} \tag{85}$$

where (i) uses the fact that $\|V'_0\|_\infty, \|\widehat{V}_0\|_\infty \leq V_{\max}$.

Hence, with probability at least $1-2\delta$, eqs. (79), (80) & (85) hold simultaneously. In this case,

$$\begin{aligned}&\left|\widehat{Q}^{(m)}_{k,t}(s,a^{(m)}) - Q'^{(m)}_{k,t}(s,a^{(m)})\right| \overset{(i)}{=} \left|\frac{\widehat{q}^{(m)}_{k,t}(s,a^{(m)})}{\widehat{\mu}_{k,t}(s)\pi'^{(m)}_{k,t}(a^{(m)}|s)} - \frac{q^{(m)}_{k,t}(s,a^{(m)})}{\mu'_{k,t}(s)\pi'^{(m)}_{k,t}(a^{(m)}|s)}\right| \\ &\overset{(ii)}{\leq} \frac{\left|\widehat{q}^{(m)}_{k,t}(s,a^{(m)}) - q^{(m)}_{k,t}(s,a^{(m)})\right|}{\widehat{\mu}_{k,t}(s)\pi'^{(m)}_{k,t}(a^{(m)}|s)} \\ &\quad + |\mu'_{k,t}(s)\pi'^{(m)}_{k,t}(a^{(m)}|s)Q'^{(m)}_{k,t}(s,a^{(m)})| \frac{\left|\widehat{\mu}_{k,t}(s) - \mu'_{k,t}(s)\right|}{\widehat{\mu}_{k,t}(s)\mu'_{k,t}(s)\pi'^{(m)}_{k,t}(a^{(m)}|s)} \\ &\overset{(iii)}{\leq} \frac{2\left|\widehat{q}^{(m)}_{k,t}(s,a^{(m)}) - q'^{(m)}_{k,t}(s,a^{(m)})\right| + 2\left|q'^{(m)}_{k,t}(s,a^{(m)}) - q^{(m)}_{k,t}(s,a^{(m)})\right|}{\mu'_{k,t}(s)\pi'^{(m)}_{k,t}(a^{(m)}|s)} \\ &\quad + \frac{2Q_{\max}\left|\widehat{\mu}_{k,t}(s) - \mu'_{k,t}(s)\right|}{\mu'_{k,t}(s)} \\ &\overset{(iv)}{\leq} \frac{2Q_{\max}}{\mu'_{k,t}(s)\pi'^{(m)}_{k,t}(a^{(m)}|s)}\left[\frac{320}{N_{k,t}} t_{\text{mix}} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)\right. \\ &\quad \left. + \frac{12}{\sqrt{N_{k,t}}}\sqrt{t_{\text{mix}}(1+2t_{\text{mix}}/N_{k,t})\mu'_{k,t}(s)\pi'^{(m)}_{k,t}(a^{(m)}|s)\ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)}\right] + 2\left\|\widehat{V}_k - V'_k\right\|_\infty \\ &\overset{(v)}{\leq} Q_{\max}\left[\frac{640 t_{\text{mix}} A_{\max}}{N_{k,t}\mu_{\min}\epsilon'} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right) + 25\sqrt{\frac{t_{\text{mix}} A_{\max}}{N_{k,t}\mu_{\min}\epsilon'} \ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)}\right] + 2\left\|\widehat{V}_k - V'_k\right\|_\infty,\end{aligned} \tag{86}$$

where (i) uses eqs. (63)&(64), (ii) uses eq. (63), (iii) uses eq. (67) and the inequality that $\widehat{\mu}_{k,t}(s) \geq \mu'_{k,t}(s)/2$ implied by (i) of eq. (80), (iv) uses eqs. (79)&(80) and the following inequality based on eqs. (60)&(61), and (v) uses $\mu'_{k,t}(s) \geq \mu_{\min}$ (based on Assumption 1), $\pi'^{(m)}_{k,t}(a^{(m)}|s) = (1-\epsilon')\pi^{(m)}_{k,t}(a^{(m)}|s) + \epsilon'/|\mathcal{A}^{(m)}| \geq \epsilon'/A_{\max}$ and $N_{k,t} \geq 650 t_{\text{mix}}$.

$$\begin{aligned}&\left|q'^{(m)}_{k,t}(s,a^{(m)}) - q^{(m)}_{k,t}(s,a^{(m)})\right| \\ &\leq \gamma \mathbb{E}_{\mu'_{k,t},\pi'^{(1)}_{k,t},\pi'^{(2)}_{k,t}}\left[\left|\widehat{V}_k(s') - V'_k(s')\right| \mathbb{1}\{\widetilde{s} = s, \widetilde{a}^{(m)} = a^{(m)}\}\right] \\ &\leq \mu'_{k,t}(s)\pi'^{(m)}_{k,t}(s,a^{(m)})\left\|\widehat{V}_k - V'_k\right\|_\infty.\end{aligned}$$

Notice that the following inequality always holds.

$$\begin{aligned}&\left|Q'^{(m)}_{k,t}(s,a^{(m)}) - Q^{(m)}_{k,t}(s,a^{(m)})\right| \\ &\leq \sum_{a^{(\setminus m)}} \left|Q_k(s,a^{(1)},a^{(2)})\right| \left|\pi'^{(\setminus m)}_{k,t}(a^{(\setminus m)}|s) - \pi^{(\setminus m)}_{k,t}(a^{(\setminus m)}|s)\right| \\ &\overset{(i)}{\leq} Q_{\max}\epsilon' \left|\pi^{(\setminus m)}_{k,t}(a^{(\setminus m)}|s) - |\mathcal{A}^{(m)}|^{-1}\right| \leq Q_{\max}\epsilon',\end{aligned} \tag{87}$$

where (i) uses eq. (67).

Hence, combining eqs. (86)&(87), it can be seen that with probability at least $1-2\delta$, the following inequality holds

$$\begin{aligned}&\left|\widehat{Q}_{k,t}^{(m)}(s,a^{(m)})-Q_{k,t}^{(m)}(s,a^{(m)})\right|\\&\leq Q_{\max}\left[\frac{640t_{\text{mix}}A_{\max}}{N_{k,t}\mu_{\min}\epsilon'}\ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)+25\sqrt{\frac{t_{\text{mix}}A_{\max}}{N_{k,t}\mu_{\min}\epsilon'}\ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)}+\epsilon'\right]\\&\quad+2\left\|\widehat{V}_k-V_k'\right\|_\infty.\end{aligned}\tag{88}$$

Similarly, it can be proved that with probability at least $1-2\delta$,

$$\begin{aligned}&\left|\widehat{\overline{Q}}_{k,t}^{(m)}(s,a^{(m)})-\overline{Q}_{k,t}^{(m)}(s,a^{(m)})\right|\\&\leq Q_{\max}\left[\frac{640t_{\text{mix}}A_{\max}}{\overline{N}_{k,t}\mu_{\min}\overline{\epsilon}'}\ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)+25\sqrt{\frac{t_{\text{mix}}A_{\max}}{\overline{N}_{k,t}\mu_{\min}\overline{\epsilon}'}\ln\left(\frac{4}{\delta\sqrt{\mu_{\min}}}\right)}+\overline{\epsilon}'\right]\\&\quad+2\left\|\widehat{V}_k-V_k'\right\|_\infty.\end{aligned}\tag{89}$$

Finally, eqs. (71)-(73) are proved by applying a union bound to eq. (85) and to eqs. (88)&(89) over all $0\leq k\leq T-1$, $0\leq t\leq T_k-1$, $m=1,2$, $s\in\mathcal{S}$, $a^{(m)}\in\mathcal{A}^{(m)}$.

Now we consider the hyperparameter choice $\epsilon'=\left[\frac{650t_{\text{mix}}A_{\max}}{N_{k,t}\mu_{\min}}\ln\left(\frac{20T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\right)\right]^{1/3}$. First, this is a valid choice, since $N_{k,t}\geq\frac{650t_{\text{mix}}A_{\max}}{\mu_{\min}}\ln\left(\frac{20T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\right)$ implies that $\epsilon'\in[0,1]$. Then, for this ϵ', the upper bound (72) is simplified as follows.

$$\begin{aligned}&\left\|\widehat{Q}_{k,t}^{(m)}-Q_{k,t}^{(m)}\right\|_\infty\\&\leq Q_{\max}\left[\frac{640t_{\text{mix}}A_{\max}}{N_{k,t}\mu_{\min}\epsilon'}\ln\left(\frac{20T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\right)+25\sqrt{\frac{t_{\text{mix}}A_{\max}}{N_{k,t}\mu_{\min}\epsilon'}\ln\left(\frac{20T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\right)}+\epsilon'\right]\\&\quad+2\left\|\widehat{V}_k-V_k'\right\|_\infty\\&\overset{(i)}{\leq}Q_{\max}\left(\frac{640}{650}\epsilon'^2+\frac{25}{\sqrt{650}}\epsilon'\right)+2\left\|\widehat{V}_k-V_k'\right\|_\infty\\&\overset{(ii)}{\leq}2Q_{\max}\epsilon'+2\left\|\widehat{V}_k-V_k'\right\|_\infty=18Q_{\max}\left[\frac{t_{\text{mix}}A_{\max}}{N_{k,t}\mu_{\min}}\ln\left(\frac{20T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\right)\right]^{1/3}+2\left\|\widehat{V}_k-V_k'\right\|_\infty,\end{aligned}$$

where (i) uses $\frac{t_{\text{mix}}A_{\max}}{N_{k,t}\mu_{\min}}\ln\left(\frac{20T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\right)=\epsilon'^3/650$, and (ii) uses $\epsilon'\in[0,1]$. This proves eq. (74). The proof of eq. (75) is similar. □

C Properties of the Duality Gap

In this section, we prove some useful properties of the duality gap

$$D^{(\tau)}(\pi^{(1)},\pi^{(2)}):=\max_{s,\pi'^{(1)},\pi'^{(2)}}\left[V_{\pi'^{(1)},\pi^{(2)}}^{(\tau)}(s)-V_{\pi^{(1)},\pi'^{(2)}}^{(\tau)}(s)\right].$$

Lemma 5. *For any policy pair $\pi^{(1)},\pi^{(2)}$, it holds that*

$$D^{(\tau)}(\pi^{(1)},\pi^{(2)})\leq\frac{1}{1-\gamma}\max_{s,\pi'^{(1)}(s),\pi'^{(2)}(s)}\left[f_\tau\left[Q_*^{(\tau)}(s),\pi'^{(1)}(s),\pi^{(2)}(s)\right]-f_\tau\left[Q_*^{(\tau)}(s),\pi^{(1)}(s),\pi'^{(2)}(s)\right]\right].$$

This Lemma generalized the Lemma 32 of (Wei et al., 2021) to entropy-regularized Markov game.

Proof. Throughout this proof, we denote the policy pair $(\pi_{*\tau}^{(1)},\pi_{*\tau}^{(2)})$ as the Nash equilibrium and their associated V-function and Q-function are respectively denoted as $V_*^{(\tau)}$ and $Q_*^{(\tau)}$.

Note that

$$
\begin{aligned}
&V^{(\tau)}_{\pi^{(1)},\pi'^{(2)}}(s) - V^{(\tau)}_{*}(s) \\
&= \sum_{a^{(1)},a^{(2)}} \big[Q^{(\tau)}_{\pi^{(1)},\pi'^{(2)}}(s,a^{(1)},a^{(2)})\pi^{(1)}(a^{(1)}|s)\pi'^{(2)}(a^{(2)}|s) \\
&\quad - Q^{(\tau)}_{*}(s,a^{(1)},a^{(2)})\pi^{(1)}_{*\tau}(a^{(1)}|s)\pi^{(2)}_{*\tau}(a^{(2)}|s)\big] \\
&\quad + \tau\big[\mathcal{H}\big(\pi^{(1)}(s)\big) - \mathcal{H}\big(\pi'^{(2)}(s)\big) - \mathcal{H}\big(\pi^{(1)}_{*\tau}(s)\big) + \mathcal{H}\big(\pi^{(2)}_{*\tau}(s)\big)\big] \\
&= \sum_{a^{(1)},a^{(2)}} \big[Q^{(\tau)}_{\pi^{(1)},\pi'^{(2)}}(s,a^{(1)},a^{(2)}) - Q^{(\tau)}_{*}(s,a^{(1)},a^{(2)})\big]\pi^{(1)}(a^{(1)}|s)\pi'^{(2)}(a^{(2)}|s) \\
&\quad + \sum_{a^{(1)},a^{(2)}} Q^{(\tau)}_{*}(s,a^{(1)},a^{(2)})\big[\pi^{(1)}(a^{(1)}|s)\pi'^{(2)}(a^{(2)}|s) - \pi^{(1)}_{*\tau}(a^{(1)}|s)\pi^{(2)}_{*\tau}(a^{(2)}|s)\big] \\
&\quad + \tau\big[\mathcal{H}\big(\pi^{(1)}(s)\big) - \mathcal{H}\big(\pi'^{(2)}(s)\big) - \mathcal{H}\big(\pi^{(1)}_{*\tau}(s)\big) + \mathcal{H}\big(\pi^{(2)}_{*\tau}(s)\big)\big] \\
&= \gamma \sum_{a^{(1)},a^{(2)}} \Big[\pi^{(1)}(a^{(1)}|s)\pi'^{(2)}(a^{(2)}|s)\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a^{(1)},a^{(2)})}\big[V^{(\tau)}_{\pi^{(1)},\pi'^{(2)}}(s') - V^{(\tau)}_{*}(s')\big]\Big] \\
&\quad + \pi^{(1)}(s)^{\top}Q^{(\tau)}_{*}(s)\pi'^{(2)}(s) - \pi^{(1)}_{*\tau}(s)^{\top}Q^{(\tau)}_{*}(s)\pi^{(2)}_{*\tau}(s) \\
&\quad + \tau\big[\mathcal{H}\big(\pi^{(1)}(s)\big) - \mathcal{H}\big(\pi'^{(2)}(s)\big) - \mathcal{H}\big(\pi^{(1)}_{*\tau}(s)\big) + \mathcal{H}\big(\pi^{(2)}_{*\tau}(s)\big)\big] \\
&\le \gamma \max_{s'}\big[V^{(\tau)}_{\pi^{(1)},\pi'^{(2)}}(s') - V^{(\tau)}_{*}(s')\big] + \pi^{(1)}(s)^{\top}Q^{(\tau)}_{*}(s)\pi'^{(2)}(s) - \pi^{(1)}_{*\tau}(s)^{\top}Q^{(\tau)}_{*}(s)\pi^{(2)}_{*\tau}(s) \\
&\quad + \tau\big[\mathcal{H}\big(\pi^{(1)}(s)\big) - \mathcal{H}\big(\pi'^{(2)}(s)\big) - \mathcal{H}\big(\pi^{(1)}_{*\tau}(s)\big) + \mathcal{H}\big(\pi^{(2)}_{*\tau}(s)\big)\big].
\end{aligned}
$$

Applying $\max_s$ to both sides of the above inequality and rearranging it yields that

$$
\begin{aligned}
&\max_{s}\big[V^{(\tau)}_{\pi^{(1)},\pi'^{(2)}}(s) - V^{(\tau)}_{*}(s)\big] \\
&\le \frac{1}{1-\gamma}\big[\pi^{(1)}(s)^{\top}Q^{(\tau)}_{*}(s)\pi'^{(2)}(s) - \pi^{(1)}_{*\tau}(s)^{\top}Q^{(\tau)}_{*}(s)\pi^{(2)}_{*\tau}(s)\big] \\
&\quad + \frac{\tau}{1-\gamma}\big[\mathcal{H}\big(\pi^{(1)}(s)\big) - \mathcal{H}\big(\pi'^{(2)}(s)\big) - \mathcal{H}\big(\pi^{(1)}_{*\tau}(s)\big) + \mathcal{H}\big(\pi^{(2)}_{*\tau}(s)\big)\big]
\end{aligned} \tag{90}
$$

Similarly, we can obtain that

$$
\begin{aligned}
&\max_{s}\big[V^{(\tau)}_{*}(s) - V^{(\tau)}_{\pi'^{(1)},\pi^{(2)}}(s)\big] \\
&\le \frac{1}{1-\gamma}\big[\pi^{(1)}_{*\tau}(s)^{\top}Q^{(\tau)}_{*}(s)\pi^{(2)}_{*\tau}(s) - \pi'^{(1)}(s)^{\top}Q^{(\tau)}_{*}(s)\pi^{(2)}(s)\big] \\
&\quad + \frac{\tau}{1-\gamma}\big[\mathcal{H}\big(\pi^{(1)}_{*\tau}(s)\big) - \mathcal{H}\big(\pi^{(2)}_{*\tau}(s)\big) - \mathcal{H}\big(\pi'^{(1)}(s)\big) + \mathcal{H}\big(\pi^{(2)}(s)\big)\big]
\end{aligned} \tag{91}
$$

Summing up eqs. (90)&(91) yields that

$$
\begin{aligned}
&\max_{s}\big[V^{(\tau)}_{\pi^{(1)},\pi'^{(2)}}(s) - V^{(\tau)}_{\pi'^{(1)},\pi^{(2)}}(s)\big] \\
&\le \frac{1}{1-\gamma}\max_{s}\Big[\pi^{(1)}(s)^{\top}Q^{(\tau)}_{*}(s)\pi'^{(2)}(s) - \pi'^{(1)}(s)^{\top}Q^{(\tau)}_{*}(s)\pi^{(2)}(s) \\
&\quad + \tau\big[\mathcal{H}\big(\pi^{(1)}(s)\big) - \mathcal{H}\big(\pi'^{(2)}(s)\big) - \mathcal{H}\big(\pi'^{(1)}(s)\big) + \mathcal{H}\big(\pi^{(2)}(s)\big)\big]\Big] \\
&= \frac{1}{1-\gamma}\max_{s}\Big[f_{\tau}\big[Q^{(\tau)}_{*}(s),\pi^{(1)}(s),\pi'^{(2)}(s)\big] - f_{\tau}\big[Q^{(\tau)}_{*}(s),\pi'^{(1)}(s),\pi^{(2)}(s)\big]\Big]
\end{aligned}
$$

Maximizing both sizes of the above inequality with respect to $\pi^{(1)},\pi^{(2)}$ proves this Lemma. □

Lemma 6. *$D^{(\tau)}(\pi^{(1)},\pi^{(2)})$ is $\frac{2\ln A_{\max}}{1-\gamma}$-Lipschitz continuous with regard to τ.*

Proof. The definition of the state value function $V^{(\tau)}_{\pi^{(1)},\pi^{(2)}}$ in (3) can be rewritten as follows.

$$
V^{(\tau)}_{\pi^{(1)},\pi^{(2)}}(s) = \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^{t}R_t\Big|s_0=s\Big] + \sum_{t=0}^{\infty}\gamma^{t}\tau[\mathcal{H}(\pi^{(1)}(s_t)) - \mathcal{H}(\pi^{(2)}(s_t))]. \tag{92}
$$

Hence, for any $\tau, \tau' \geq 0$,

$$\begin{aligned}
&\left|V^{(\tau')}_{\pi^{(1)},\pi^{(2)}}(s) - V^{(\tau)}_{\pi^{(1)},\pi^{(2)}}(s)\right| \\
&\leq \sum_{t=0}^{\infty} \gamma^t |\tau' - \tau| |\mathcal{H}(\pi^{(1)}(s_t)) - \mathcal{H}(\pi^{(2)}(s_t))| \\
&\overset{(i)}{\leq} \sum_{t=0}^{\infty} \gamma^t |\tau' - \tau| \ln A_{\max} \leq \frac{\ln A_{\max}}{1-\gamma} |\tau' - \tau|,
\end{aligned} \tag{93}$$

where (i) uses $0 \leq \mathcal{H}(\pi^{(m)}(s_t)) \leq \ln |\mathcal{A}^{(m)}| \leq \ln A_{\max}$. Hence, this Lemma can be proved as follows.

$$\begin{aligned}
&|D^{(\tau')}(\pi^{(1)}, \pi^{(2)}) - D^{(\tau)}(\pi^{(1)}, \pi^{(2)})| \\
&\leq \left| \max_{s,\pi'^{(1)},\pi'^{(2)}} \left[V^{(\tau')}_{\pi'^{(1)},\pi^{(2)}}(s) - V^{(\tau')}_{\pi^{(1)},\pi'^{(2)}}(s)\right] - \max_{s,\pi'^{(1)},\pi'^{(2)}} \left[V^{(\tau)}_{\pi'^{(1)},\pi^{(2)}}(s) - V^{(\tau)}_{\pi^{(1)},\pi'^{(2)}}(s)\right] \right| \\
&\leq \max_{s,\pi'^{(1)},\pi'^{(2)}} \left| \left[V^{(\tau')}_{\pi'^{(1)},\pi^{(2)}}(s) - V^{(\tau')}_{\pi^{(1)},\pi'^{(2)}}(s)\right] - \left[V^{(\tau)}_{\pi'^{(1)},\pi^{(2)}}(s) - V^{(\tau)}_{\pi^{(1)},\pi'^{(2)}}(s)\right] \right| \\
&\leq \max_{s,\pi'^{(1)},\pi'^{(2)}} \left[\left|V^{(\tau')}_{\pi'^{(1)},\pi^{(2)}}(s) - V^{(\tau)}_{\pi'^{(1)},\pi^{(2)}}(s)\right| + \left|V^{(\tau)}_{\pi^{(1)},\pi'^{(2)}}(s) - V^{(\tau')}_{\pi^{(1)},\pi'^{(2)}}(s)\right| \right] \\
&\overset{(i)}{\leq} \frac{2 \ln A_{\max}}{1-\gamma} |\tau' - \tau|,
\end{aligned}$$

where (i) uses eq. (93). □

D Proof of Theorem 1

Theorem 1 (Finite-time convergence rate)**.** *Apply Algorithm 1 to solve the entropy-regularized Markov game with $\tau \in (0, 1]$. Choose learning rate $\eta = [2(\tau + Q_{\max})]^{-1}$, initialization $\|\widehat{V}_0\|_\infty \leq V_{\max}$ and batch sizes $N_{k,t}, \overline{N}_{k,t}, N_{k+1}$ that satisfy (18) & (19). Then, the Nash equilibrium duality gap converges at the following rate with probability at least $1 - \delta$.*

$$\begin{aligned}
&D^{(\tau)}\left(\pi^{(1)}_{K-1}, \pi^{(2)}_{K-1}\right) \leq \mathcal{O}\Bigg(\frac{V_{\max} \ln A_{\max}}{1-\gamma} \sum_{k=0}^{K-1} \gamma^{K-k} (1-\eta\tau)^{T_k - 1} \\
&+ \frac{V_{\max}}{1-\gamma} \left[\frac{t_{mix} A_{\max}}{\mu_{\min}} \ln\left(\frac{T_{sum} |\mathcal{S}| A_{\max}}{\delta \mu_{\min}} \right) \right]^{2/3} \sum_{k=0}^{K-1} \gamma^{K-k-1} \sum_{t=0}^{T_k - 1} (1-\eta\tau)^{T_k - 2 - t} \left(\frac{1}{N_{k,t}^{2/3}} + \frac{V_{\max}}{\tau \overline{N}_{k,t+1}^{2/3}} \right) \\
&+ \frac{V_{\max}^3 \gamma^K}{\tau^2 (1-\gamma)^2} + \frac{t_{mix} V_{\max}^3}{\tau^2 \mu_{\min} (1-\gamma)^3} \ln\left(\frac{K|\mathcal{S}|}{\delta \mu_{\min}} \right) \sum_{k=0}^{K-1} \frac{\gamma^{K-k-1}}{N_{k+1}} \Bigg).
\end{aligned} \tag{20}$$

Proof. First, consider the SPU iterations that are defined by replacing $Q^{(m)}_{k,t}$, $\overline{Q}^{(m)}_{k,t+1}$ in (8) with $\widehat{Q}^{(m)}_{k,t}$, $\widehat{\overline{Q}}^{(m)}_{k,t+1}$, respectively, for $m = 1, 2$. We can apply Lemma 1 with the quantities being specified as follows.

- $Q := Q_k(s)$, $h_\tau(\mu, \nu) := f_\tau\big(Q_k(s); \mu, \nu\big)$,
- $\mu_t := \pi^{(1)}_{k,t}(s)$, $\nu_t := \pi^{(2)}_{k,t}(s)$, $\overline{\mu}_{t+1} := \overline{\pi}^{(1)}_{k,t+1}(s)$, $\overline{\nu}_{t+1} := \overline{\pi}^{(2)}_{k,t+1}(s)$, $\mu^*_\tau := \pi^{*(1)}_k(s)$, $\nu^*_\tau := \pi^{*(2)}_k(s)$,
- $d_m := |\mathcal{A}^{(m)}|$,
- $\delta^{(m)}_{k,t} := \widehat{Q}^{(m)}_{k,t}(s) - Q^{(m)}_{k,t}(s)$, $\overline{\delta}^{(m)}_{k,t} := \widehat{\overline{Q}}^{(m)}_{k,t+1}(s) - \overline{Q}^{(m)}_{k,t+1}(s)$,

where $\pi_k^{(1)*}(s), \pi_k^{(2)*}(s)$ is the solution policy pair to the minimax optimization problem $\min_{\pi^{(2)}(s)} \max_{\pi^{(1)}(s)} f_\tau\big(Q_k(s); \pi^{(1)}(s), \pi^{(2)}(s)\big)$.

Consequently, eqs. (28) & (29) with $t = T_k - 1$ imply the following two inequalities, respectively.

$$\begin{aligned}
&\big|f_\tau\big(Q_k(s); \overline{\pi}_{k,T_k}^{(1)}(s), \overline{\pi}_{k,T_k}^{(1)}(s)\big) - f_\tau\big(Q_k(s); \pi_k^{*(1)}(s), \pi_k^{*(2)}(s)\big)\big| \\
&= \big|V'_{k+1}(s) - f_\tau\big(Q_k(s); \pi_k^{*(1)}(s), \pi_k^{*(2)}(s)\big)\big| \\
&\leq \frac{8}{\eta}(1-\eta\tau)^{T_k-1} \ln A_{\max} + \sum_{t=0}^{T_k-1} (1-\eta\tau)^{T_k-2-t} \\
&\quad \sum_{m=1}^{2} \Big[8\eta \big\|\widehat{Q}_{k,t}^{(m)}(s) - Q_{k,t}^{(m)}(s)\big\|_\infty^2 + \frac{33}{4\tau}\big\|\widehat{\overline{Q}}_{k,t+1}^{(m)}(s) - \overline{Q}_{k,t+1}^{(m)}(s)\big\|_\infty^2\Big],
\end{aligned} \tag{94}$$

$$\begin{aligned}
&\max_{\pi^{(1)}(s),\pi^{(2)}(s)} \big[f_\tau\big(Q_k(s); \pi^{(1)}(s), \overline{\pi}_{k,T_k}^{(2)}(s)\big) - f_\tau\big(Q_k(s); \overline{\pi}_{k,T_k}^{(1)}(s), \pi^{(2)}(s)\big)\big] \\
&= \max_{\pi^{(1)}(s),\pi^{(2)}(s)} \big[f_\tau\big(Q_k(s); \pi^{(1)}(s), \pi_k^{(2)}(s)\big) - f_\tau\big(Q_k(s); \pi_k^{(1)}(s), \pi^{(2)}(s)\big)\big] \\
&\leq \frac{4}{\eta}(1-\eta\tau)^{T_k-1} \ln A_{\max} + \sum_{t=0}^{T_k-1} (1-\eta\tau)^{T_k-2-t} \\
&\quad \sum_{m=1}^{2} \Big[16\eta \big\|\widehat{Q}_{k,t}^{(m)}(s) - Q_{k,t}^{(m)}(s)\big\|_\infty^2 + \frac{12}{\tau}\big\|\widehat{\overline{Q}}_{k,t+1}^{(m)}(s) - \overline{Q}_{k,t+1}^{(m)}(s)\big\|_\infty^2\Big].
\end{aligned} \tag{95}$$

Then, define the soft Bellman operator $\mathcal{B}_\tau$ as follows.

$$\begin{aligned}
&\mathcal{B}_\tau(Q)(s, a^{(1)}, a^{(2)}) \\
&:= R(s, a^{(1)}, a^{(2)}) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a^{(1)},a^{(2)})}\Big[\max_{\pi^{(1)}(s')} \min_{\pi^{(2)}(s')} f_\tau\big(Q(s'); \pi^{(1)}(s'), \pi^{(2)}(s')\big)\Big],
\end{aligned} \tag{96}$$

where $Q : \mathcal{S} \times \mathcal{A}^{(1)} \times \mathcal{A}^{(2)} \to \mathbb{R}$ and $Q(s') \in \mathbb{R}^{|\mathcal{A}^{(1)}| \times |\mathcal{A}^{(2)}|}$ with each entry given by $[Q(s')]_{a^{(1)},a^{(2)}} := Q(s, a^{(1)}, a^{(2)})$. It can be proved that $\mathcal{B}_\tau$ is a contraction operator and has a unique fixed point $Q_*^{(\tau)}$ (Cen et al., 2021), that is,

$$\|\mathcal{B}_\tau(Q') - \mathcal{B}_\tau(Q)\|_\infty \leq \gamma \|Q' - Q\|_\infty, \tag{97}$$

$$\mathcal{B}_\tau(Q_*^{(\tau)}) = Q_*^{(\tau)}. \tag{98}$$

As a result, we obtain that

$$\begin{aligned}
&\|Q_{k+1} - Q_*^{(\tau)}\|_\infty \\
&\overset{(i)}{\leq} \|Q_{k+1} - \mathcal{B}_\tau(Q_k)\|_\infty + \|\mathcal{B}_\tau(Q_k) - \mathcal{B}_\tau(Q_*^{(\tau)})\|_\infty \\
&\overset{(ii)}{\leq} \max_{s,a^{(1)},a^{(2)}} |Q_{k+1}(s, a^{(1)}, a^{(2)}) - \mathcal{B}_\tau(Q_k)(s, a^{(1)}, a^{(2)})| + \gamma \|Q_k - Q_*^{(\tau)}\|_\infty \\
&= \max_{s,a^{(1)},a^{(2)}} \Big|R(s, a^{(1)}, a^{(2)}) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a^{(1)},a^{(2)})} V'_{k+1}(s') \\
&\quad - \Big[R(s, a^{(1)}, a^{(2)}) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a^{(1)},a^{(2)})} f_\tau\big(Q_k(s'); \pi_k^{*(1)}(s'), \pi_k^{*(2)}(s')\big)\Big]\Big| + \gamma \|Q_k - Q_*^{(\tau)}\|_\infty \\
&\leq \gamma \max_{s' \in \mathcal{S}} \big|V'_{k+1}(s') - f_\tau\big(Q_k(s'); \pi_k^{*(1)}(s'), \pi_k^{*(2)}(s')\big)\big| + \gamma \|Q_k - Q_*^{(\tau)}\|_\infty,
\end{aligned} \tag{99}$$

where (i) uses eq. (98) and (ii) uses eq. (97). Throughout the proof, suppose that eqs. (71), (74) & (75) hold simultaneously which occurs with probability at least $1 - \delta$. In this case, we have

$$\big\|\widehat{Q}_{k,t}^{(m)}(s) - Q_{k,t}^{(m)}(s)\big\|_\infty^2$$

$$
\begin{aligned}
&\overset{(i)}{\leq} 648Q_{\max}^2\Big[\frac{t_{\text{mix}}A_{\max}}{N_{k,t}\mu_{\min}}\ln\Big(\frac{20T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\Big)\Big]^{2/3}+8\big\|\widehat{V}_k-V_k'\big\|_\infty^2\\
&\overset{(ii)}{\leq} 648Q_{\max}^2\Big[\frac{t_{\text{mix}}A_{\max}}{N_{k,t}\mu_{\min}}\ln\Big(\frac{20T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\Big)\Big]^{2/3}\\
&\quad+32V_{\max}^2\gamma^{2k}+\frac{29241t_{\text{mix}}V_{\max}^2}{N_{k+1}\mu_{\min}(1-\gamma)^2}\ln\Big(\frac{4K|\mathcal{S}|}{\delta\sqrt{\mu_{\min}}}\Big),
\end{aligned}
\tag{100}
$$

where (i) uses eq. (74) and the inequality that $(a+b)^2\leq 2a^2+2b^2$ for any $a,b\geq 0$, (ii) uses eq. (71) and also $(a+b)^2\leq 2a^2+2b^2$ for any $a,b\geq 0$. Similarly, we obtain that

$$
\begin{aligned}
&\big\|\widehat{\overline{Q}}_{k,t+1}^{(m)}(s)-\overline{Q}_{k,t+1}^{(m)}(s)\big\|_\infty^2\\
&\leq 648Q_{\max}^2\Big[\frac{t_{\text{mix}}A_{\max}}{\overline{N}_{k,t+1}\mu_{\min}}\ln\Big(\frac{20T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\Big)\Big]^{2/3}\\
&\quad+32V_{\max}^2\gamma^{2k}+\frac{29241t_{\text{mix}}V_{\max}^2}{N_{k+1}\mu_{\min}(1-\gamma)^2}\ln\Big(\frac{4K|\mathcal{S}|}{\delta\sqrt{\mu_{\min}}}\Big).
\end{aligned}
\tag{101}
$$

Then, iterating eq. (99) yields that

$$
\begin{aligned}
&\|Q_{K-1}-Q_*^{(\tau)}\|_\infty\\
&\leq\gamma^{K-1}\|Q_0-Q_*^{(\tau)}\|_\infty+\sum_{k=0}^{K-2}\gamma^{K-k-1}\max_{s'\in\mathcal{S}}\big|V_{k+1}(s')-f_\tau\big(Q_k(s');\pi_k^{*(1)}(s'),\pi_k^{*(2)}(s')\big)\big|\\
&\overset{(i)}{\leq} 2\gamma^{K-1}Q_{\max}+\sum_{k=0}^{K-2}\gamma^{K-k-1}\Bigg[\frac{8}{\eta}(1-\eta\tau)^{T_k-1}\ln A_{\max}+\sum_{t=0}^{T_k-1}(1-\eta\tau)^{T_k-2-t}\\
&\quad\sum_{m=1}^{2}\Big[8\eta\big\|\widehat{Q}_{k,t}^{(m)}(s)-Q_{k,t}^{(m)}(s)\big\|_\infty^2+\frac{33}{4\tau}\big\|\widehat{\overline{Q}}_{k,t+1}^{(m)}(s)-\overline{Q}_{k,t+1}^{(m)}(s)\big\|_\infty^2\Big]\Bigg]\\
&\overset{(ii)}{\leq} 2\gamma^{K-1}Q_{\max}+\sum_{k=0}^{K-2}\gamma^{K-k-1}\Bigg[\frac{8}{\eta}(1-\eta\tau)^{T_k-1}\ln A_{\max}+\sum_{t=0}^{T_k-1}(1-\eta\tau)^{T_k-2-t}\\
&\quad\sum_{m=1}^{2}\Bigg(648Q_{\max}^2\Big[\frac{t_{\text{mix}}A_{\max}}{\mu_{\min}}\ln\Big(\frac{20T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\Big)\Big]^{2/3}\Big(\frac{8\eta}{N_{k,t}^{2/3}}+\frac{33}{4\tau\overline{N}_{k,t+1}^{2/3}}\Big)\\
&\quad+\Big(8\eta+\frac{33}{4\tau}\Big)\Big[32V_{\max}^2\gamma^{2k}+\frac{29241t_{\text{mix}}V_{\max}^2}{N_{k+1}\mu_{\min}(1-\gamma)^2}\ln\Big(\frac{4K|\mathcal{S}|}{\delta\sqrt{\mu_{\min}}}\Big)\Big]\Bigg)\Bigg]\\
&\overset{(iii)}{\leq} 2\gamma^{K-1}Q_{\max}+\frac{8}{\eta}\ln A_{\max}\sum_{k=0}^{K-2}\gamma^{K-k-1}(1-\eta\tau)^{T_k-1}\\
&\quad+324Q_{\max}^2\Big[\frac{t_{\text{mix}}A_{\max}}{\mu_{\min}}\ln\Big(\frac{20T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\Big)\Big]^{2/3}\\
&\quad\sum_{k=0}^{K-2}\gamma^{K-k-1}\sum_{t=0}^{T_k-1}(1-\eta\tau)^{T_k-2-t}\Big(\frac{32\eta}{N_{k,t}^{2/3}}+\frac{33}{\tau\overline{N}_{k,t+1}^{2/3}}\Big)+\frac{1568\gamma^{K-1}V_{\max}^2}{\eta\tau^2(1-\gamma)}\\
&\quad+\frac{1432809t_{\text{mix}}V_{\max}^2}{\eta\tau^2\mu_{\min}(1-\gamma)^2}\ln\Big(\frac{4K|\mathcal{S}|}{\delta\sqrt{\mu_{\min}}}\Big)\sum_{k=0}^{K-2}\gamma^{K-k-1}N_{k+1}^{-1}\\
&\overset{(iv)}{=}\mathcal{O}\Bigg[V_{\max}\ln A_{\max}\sum_{k=0}^{K-2}\gamma^{K-k}(1-\eta\tau)^{T_k-1}+V_{\max}\Big[\frac{t_{\text{mix}}A_{\max}}{\mu_{\min}}\ln\Big(\frac{T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\mu_{\min}}\Big)\Big]^{2/3}\\
&\quad\sum_{k=0}^{K-2}\gamma^{K-k-1}\sum_{t=0}^{T_k-1}(1-\eta\tau)^{T_k-2-t}\Big(\frac{1}{N_{k,t}^{2/3}}+\frac{V_{\max}}{\tau\overline{N}_{k,t+1}^{2/3}}\Big)
\end{aligned}
\tag{102}
$$

$$+ \frac{V_{\max}^3 \gamma^K}{\tau^2(1-\gamma)} + \frac{t_{\text{mix}} V_{\max}^3}{\tau^2 \mu_{\min}(1-\gamma)^2} \ln\Big(\frac{K|\mathcal{S}|}{\delta \mu_{\min}}\Big) \sum_{k=0}^{K-2} \gamma^{K-k-1} N_{k+1}^{-1} \Bigg] \tag{103}$$

where (i) uses eqs. (67)&(94), (ii) uses eqs. (100)&(101), (iii) uses $\eta = [2(\tau + Q_{\max})]^{-1} \leq 1/(2\tau)$, $\sum_{t=0}^{T_k-1}(1-\eta\tau)^{T_k-2-t} \leq \frac{1}{\eta\tau(1-\eta\tau)} \leq \frac{2}{\eta\tau}$ and $\sum_{k=0}^{K-2} \gamma^{K+k-1} \leq \frac{\gamma^{K-1}}{1-\gamma}$, and (iv) uses $Q_{\max} = \mathcal{O}(V_{\max})$ for $\gamma \approx 1$ and $\eta = [2(\tau + Q_{\max})]^{-1} = \mathcal{O}(Q_{\max}^{-1}) = \mathcal{O}(V_{\max}^{-1})$ (Since $Q_{\max} \geq \gamma\tau = \mathcal{O}(\tau)$).

Using Lemma 5, the convergence rate of the duality gap in eq. (20) can be proved as follows.

$$\begin{aligned}
&D^{(\tau)}\big(\pi_{K-1}^{(1)}, \pi_{K-1}^{(2)}\big) \\
&\overset{(i)}{\leq} \frac{1}{1-\gamma} \max_{s,\pi^{(1)},\pi^{(2)}} \Big[f_\tau\big[Q_*^{(\tau)}(s), \pi^{(1)}(s), \pi_{K-1}^{(2)}(s)\big] - f_\tau\big[Q_*^{(\tau)}(s), \pi_{K-1}^{(1)}(s), \pi^{(2)}(s)\big] \Big] \\
&\leq \frac{1}{1-\gamma} \max_{s,\pi^{(1)},\pi^{(2)}} \Big[f_\tau\big(Q_{K-1}(s); \pi^{(1)}(s), \pi_{K-1}^{(2)}(s)\big) - f_\tau\big(Q_{K-1}(s); \pi_{K-1}^{(1)}(s), \pi^{(2)}(s)\big) \Big] \\
&\quad + \frac{1}{1-\gamma} \max_{s,\pi^{(1)},\pi^{(2)}} \big| f_\tau\big(Q_*^{(\tau)}(s); \pi^{(1)}(s), \pi_{K-1}^{(2)}(s)\big) - f_\tau\big(Q_{K-1}(s); \pi^{(1)}(s), \pi_{K-1}^{(2)}(s)\big) \big| \\
&\quad + \frac{1}{1-\gamma} \max_{s,\pi^{(1)},\pi^{(2)}} \big| f_\tau\big(Q_{K-1}(s); \pi_{K-1}^{(1)}(s), \pi^{(2)}(s)\big) - f_\tau\big(Q_*^{(\tau)}(s); \pi_{K-1}^{(1)}(s), \pi^{(2)}(s)\big) \big| \\
&\overset{(ii)}{\leq} \frac{4}{\eta(1-\gamma)}(1-\eta\tau)^{T_{K-1}-1} \ln A_{\max} + \frac{1}{1-\gamma} \sum_{t=0}^{T_{K-1}-1} (1-\eta\tau)^{T_{K-1}-2-t} \\
&\quad \sum_{m=1}^{2} \Big[16\eta \big\| \widehat{Q}_{K-1,t}^{(m)}(s) - Q_{K-1,t}^{(m)}(s) \big\|_\infty^2 + \frac{12}{\tau} \big\| \widehat{\overline{Q}}_{K-1,t+1}^{(m)}(s) - \overline{Q}_{K-1,t+1}^{(m)}(s) \big\|_\infty^2 \Big] \\
&\quad + \frac{1}{1-\gamma} \max_{s,\pi^{(1)}} \big| \pi^{(1)}(s)^\top [Q_*^{(\tau)}(s) - Q_{K-1}(s)] \pi_{K-1}^{(2)}(s) \big| \\
&\quad + \frac{1}{1-\gamma} \max_{s,\pi^{(2)}} \big| \pi_{K-1}^{(1)}(s)^\top [Q_{K-1}(s) - Q_*^{(\tau)}(s)] \pi^{(2)}(s) \big| \\
&\overset{(iii)}{\leq} \frac{4 \ln A_{\max}}{\eta(1-\gamma)}(1-\eta\tau)^{T_{K-1}-1} + \frac{2}{1-\gamma} \sum_{t=0}^{T_{K-1}-1} (1-\eta\tau)^{T_{K-1}-2-t} \\
&\quad \Bigg(648 Q_{\max}^2 \Big[\frac{t_{\text{mix}} A_{\max}}{\mu_{\min}} \ln\Big(\frac{20 T_{\text{sum}} |\mathcal{S}| A_{\max}}{\delta \sqrt{\mu_{\min}}} \Big) \Big]^{2/3} \Big(\frac{16\eta}{N_{K-1,t}^{2/3}} + \frac{12}{\tau \overline{N}_{K-1,t+1}^{2/3}} \Big) \\
&\quad + \Big(16\eta + \frac{12}{\tau} \Big) \Big[32 V_{\max}^2 \gamma^{2(K-1)} + \frac{29241 t_{\text{mix}} V_{\max}^2}{N_K \mu_{\min}(1-\gamma)^2} \ln\Big(\frac{4K|\mathcal{S}|}{\delta \sqrt{\mu_{\min}}} \Big) \Big] \Bigg) \\
&\quad + \frac{2}{1-\gamma} \| Q_{K-1} - Q_*^{(\tau)} \|_\infty \\
&\overset{(iv)}{\leq} \frac{4 \ln A_{\max}}{\eta(1-\gamma)}(1-\eta\tau)^{T_{K-1}-1} + \frac{1296 Q_{\max}^2}{1-\gamma} \Big[\frac{t_{\text{mix}} A_{\max}}{\mu_{\min}} \ln\Big(\frac{20 T_{\text{sum}} |\mathcal{S}| A_{\max}}{\delta \sqrt{\mu_{\min}}} \Big) \Big]^{2/3} \\
&\quad \sum_{t=0}^{T_{K-1}-1} (1-\eta\tau)^{T_{K-1}-2-t} \Big(\frac{16\eta}{N_{K-1,t}^{2/3}} + \frac{12}{\tau \overline{N}_{K-1,t+1}^{2/3}} \Big) \\
&\quad + \frac{80}{\eta\tau^2} \Big[\frac{32 V_{\max}^2}{1-\gamma} \gamma^{2(K-1)} + \frac{29241 t_{\text{mix}} V_{\max}^2}{N_K \mu_{\min}(1-\gamma)^3} \ln\Big(\frac{4K|\mathcal{S}|}{\delta \sqrt{\mu_{\min}}} \Big) \Big] \\
&\quad + \frac{4\gamma^{K-1} Q_{\max}}{1-\gamma} + \frac{16 \ln A_{\max}}{\eta(1-\gamma)} \sum_{k=0}^{K-2} \gamma^{K-k-1} (1-\eta\tau)^{T_k-1} \\
&\quad + \frac{648 Q_{\max}^2}{1-\gamma} \Big[\frac{t_{\text{mix}} A_{\max}}{\mu_{\min}} \ln\Big(\frac{20 T_{\text{sum}} |\mathcal{S}| A_{\max}}{\delta \sqrt{\mu_{\min}}} \Big) \Big]^{2/3}
\end{aligned}$$

$$
\begin{aligned}
&\sum_{k=0}^{K-2}\gamma^{K-k-1}\sum_{t=0}^{T_k-1}(1-\eta\tau)^{T_k-2-t}\Big(\frac{32\eta}{N_{k,t}^{2/3}}+\frac{33}{\tau\overline{N}_{k,t+1}^{2/3}}\Big)+\frac{3136\gamma^{K-1}V_{\max}^2}{\eta\tau^2(1-\gamma)^2}\\
&+\frac{2865618t_{\text{mix}}V_{\max}^2}{\eta\tau^2\mu_{\min}(1-\gamma)^3}\ln\Big(\frac{4K|\mathcal{S}|}{\delta\sqrt{\mu_{\min}}}\Big)\sum_{k=0}^{K-2}\gamma^{K-k-1}N_{k+1}^{-1}\\
\le\ &\frac{4\gamma^{K-1}Q_{\max}}{1-\gamma}+\frac{16\ln A_{\max}}{\eta(1-\gamma)}\sum_{k=0}^{K-1}\gamma^{K-k-1}(1-\eta\tau)^{T_k-1}\\
&+\frac{648Q_{\max}^2}{1-\gamma}\Big[\frac{t_{\text{mix}}A_{\max}}{\mu_{\min}}\ln\Big(\frac{20T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\Big)\Big]^{2/3}\\
&\sum_{k=0}^{K-1}\gamma^{K-k-1}\sum_{t=0}^{T_k-1}(1-\eta\tau)^{T_k-2-t}\Big(\frac{32\eta}{N_{k,t}^{2/3}}+\frac{33}{\tau\overline{N}_{k,t+1}^{2/3}}\Big)+\frac{5696\gamma^{K-1}V_{\max}^2}{\eta\tau^2(1-\gamma)^2}\\
&+\frac{2865618t_{\text{mix}}V_{\max}^2}{\eta\tau^2\mu_{\min}(1-\gamma)^3}\ln\Big(\frac{4K|\mathcal{S}|}{\delta\sqrt{\mu_{\min}}}\Big)\sum_{k=0}^{K-1}\gamma^{K-k-1}N_{k+1}^{-1}\\
\overset{(v)}{=}\ &\mathcal{O}\Bigg[\frac{V_{\max}\ln A_{\max}}{1-\gamma}\sum_{k=0}^{K-1}\gamma^{K-k}(1-\eta\tau)^{T_k-1}+\frac{V_{\max}}{1-\gamma}\Big[\frac{t_{\text{mix}}A_{\max}}{\mu_{\min}}\ln\Big(\frac{T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\mu_{\min}}\Big)\Big]^{2/3}\\
&\sum_{k=0}^{K-1}\gamma^{K-k-1}\sum_{t=0}^{T_k-1}(1-\eta\tau)^{T_k-2-t}\Big(\frac{1}{N_{k,t}^{2/3}}+\frac{V_{\max}}{\tau\overline{N}_{k,t+1}^{2/3}}\Big)+\frac{V_{\max}^3\gamma^K}{\tau^2(1-\gamma)^2}\\
&+\frac{t_{\text{mix}}V_{\max}^3}{\tau^2\mu_{\min}(1-\gamma)^3}\ln\Big(\frac{K|\mathcal{S}|}{\delta\mu_{\min}}\Big)\sum_{k=0}^{K-1}\gamma^{K-k-1}N_{k+1}^{-1}\Bigg],
\end{aligned}
$$

where (i) uses Lemma 5, (ii) uses eq. (95) and the definition of the function f_τ, (iii) uses eqs. (100)&(101), (iv) uses eq. (102) and $\eta=[2(\tau+Q_{\max})]^{-1}\le 1/(2\tau)$, $\sum_{t=0}^{T_k-1}(1-\eta\tau)^{T_k-2-t}\le\frac{1}{\eta\tau(1-\eta\tau)}\le\frac{2}{\eta\tau}$, and (v) uses $Q_{\max}=\mathcal{O}(V_{\max})$ for $\gamma\approx 1$ and $\eta=[2(\tau+Q_{\max})]^{-1}=\mathcal{O}(Q_{\max}^{-1})=\mathcal{O}(V_{\max}^{-1})$ (Since $Q_{\max}\ge\gamma\tau=\mathcal{O}(\tau)$).

□

E Proof of Theorem 2

Theorem 2 (Sample complexity)**.** *Implement Algorithm 1 with* $\eta=\mathcal{O}(1-\gamma)$, $\tau=\mathcal{O}\big(\frac{\epsilon(1-\gamma)}{\ln A_{\max}}\big)$, $K=\mathcal{O}\Big[\frac{1}{1-\gamma}\ln\Big(\frac{\ln A_{\max}}{\epsilon(1-\gamma)}\Big)\Big]$ *and* $T_k=1+\frac{k\ln\gamma^{-1}}{\ln(1-\eta\tau)^{-1}}$. *Choose the following adaptive batch sizes.*

$$
N_{k+1}=\widetilde{\mathcal{O}}\Big(\frac{t_{mix}(\ln^2 A_{\max})\gamma^{-\frac{k}{2}}}{\epsilon^2\mu_{\min}(1-\gamma)^8}\Big),\ N_{k,t}=\overline{N}_{k,t}\epsilon^{\frac{3}{2}}(1-\gamma)^3=\widetilde{\mathcal{O}}\Big(\frac{t_{mix}A_{\max}(1-\eta\tau)^{\frac{-3(t+1)}{5}}}{\mu_{\min}(1-\gamma)^3}\Big).
$$

Then, for any $\epsilon\le\frac{\ln A_{\max}}{1-\gamma}$, *the overall sample complexity to achieve* $D^{(0)}(\pi_{K-1}^{(1)},\pi_{K-1}^{(2)})\le\epsilon$ *is* $\widetilde{\mathcal{O}}\big(\frac{t_{mix}A_{\max}}{\mu_{\min}\epsilon^{5.5}(1-\gamma)^{13.5}}\big)$. *Please refer to (117) in Appendix E for a complete expression.*

Proof. Since $\tau=\mathcal{O}\big(\frac{\epsilon(1-\gamma)}{\ln A_{\max}}\big)$, we have

$$
V_{\max}=\frac{1+\tau\ln A_{\max}}{1-\gamma}=\frac{1}{1-\gamma}+\mathcal{O}(\epsilon)=\mathcal{O}\big((1-\gamma)^{-1}\big). \tag{104}
$$

$$
Q_{\max}=\frac{1+\gamma\tau\ln A_{\max}}{1-\gamma}=\frac{1}{1-\gamma}+\mathcal{O}(\epsilon)=\mathcal{O}\big((1-\gamma)^{-1}\big). \tag{105}
$$

Since $T_k=1+\frac{k\ln\gamma^{-1}}{\ln(1-\eta\tau)^{-1}}$, we have

$$
T_{\text{sum}}=\sum_{k=0}^{K-1}\Big(1+\frac{k\ln\gamma^{-1}}{\ln(1-\eta\tau)^{-1}}\Big)
$$

$$
\begin{aligned}
&= K + \frac{K(K-1)}{2}\mathcal{O}\Big(\frac{1-\gamma}{\eta\tau}\Big) \\
&= K + \frac{K(K-1)}{2}\mathcal{O}\Big(\frac{\ln A_{\max}}{\epsilon(1-\gamma)}\Big) \\
&= \mathcal{O}\Big(\frac{K^2 \ln A_{\max}}{\epsilon(1-\gamma)}\Big) \\
&= \mathcal{O}\Big[\frac{\ln A_{\max}}{\epsilon(1-\gamma)^3}\ln^2\Big(\frac{\ln A_{\max}}{\epsilon(1-\gamma)}\Big)\Big]. \qquad (106)
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\ln\Big(\frac{T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\mu_{\min}}\Big) &= \mathcal{O}\Big[\ln\Big(\frac{|\mathcal{S}|A_{\max}\ln A_{\max}}{\delta\mu_{\min}\epsilon(1-\gamma)^3}\Big) + \ln\Big(2\ln\Big(\frac{\ln A_{\max}}{\epsilon(1-\gamma)}\Big)\Big)\Big] \\
&= \mathcal{O}\Big[\ln\Big(\frac{|\mathcal{S}|A_{\max}}{\epsilon\delta\mu_{\min}(1-\gamma)}\Big)\Big]. \qquad (107)
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\ln\Big(\frac{K|\mathcal{S}|}{\delta\mu_{\min}}\Big) &= \mathcal{O}\Big[\ln\Big(\Big(\frac{|\mathcal{S}|}{\delta\mu_{\min}(1-\gamma)}\Big)\ln\Big(\frac{\ln A_{\max}}{\epsilon(1-\gamma)}\Big)\Big)\Big] \\
&= \mathcal{O}\Big[\ln\Big(\frac{|\mathcal{S}|\ln(\epsilon^{-1}\ln A_{\max})}{\delta\mu_{\min}(1-\gamma)}\Big)\Big] + \mathcal{O}\Big[\ln\Big(\Big(\frac{|\mathcal{S}|}{\delta\mu_{\min}(1-\gamma)}\Big)\ln\Big(\frac{1}{1-\gamma}\Big)\Big)\Big] \\
&= \mathcal{O}\Big[\ln\Big(\frac{|\mathcal{S}|\ln(\epsilon^{-1}\ln A_{\max})}{\delta\mu_{\min}(1-\gamma)}\Big)\Big] + \mathcal{O}\Big[\ln\Big(\frac{|\mathcal{S}|}{\delta\mu_{\min}(1-\gamma)}\Big)\Big] \\
&= \mathcal{O}\Big[\ln\Big(\frac{|\mathcal{S}|\ln(\epsilon^{-1}\ln A_{\max})}{\delta\mu_{\min}(1-\gamma)}\Big)\Big] \qquad (108)
\end{aligned}
$$

where (i) uses $\epsilon \le \frac{\ln A_{\max}}{1-\gamma}$. With the above equalities, we will prove below that the hyperparameter choices in Theorem 2 are valid and satisfy the conditions of Theorem 1 with proper constants hidden in $\mathcal{O}(\cdot)$.

$$
\eta = [2(\tau + Q_{\max})]^{-1} = \mathcal{O}(1-\gamma) \qquad (109)
$$

$$
\tau = \mathcal{O}\Big(\frac{\epsilon(1-\gamma)}{\ln A_{\max}}\Big) \in (0,1] \qquad (110)
$$

$$
K = \frac{1}{\ln\gamma^{-1}}\ln\Big(\frac{\ln^{5/2} A_{\max}}{\epsilon^5(1-\gamma)^{7.5}}\Big) = \mathcal{O}\Big[\frac{1}{1-\gamma}\ln\Big(\frac{\ln A_{\max}}{\epsilon(1-\gamma)}\Big)\Big] \ge 1 \qquad (111)
$$

$$
T_k = 1 + \frac{k\ln\gamma^{-1}}{\ln(1-\eta\tau)^{-1}} \ge 1 \qquad (112)
$$

$$
\begin{aligned}
N_{k+1} &= \mathcal{O}\Big[\frac{t_{\text{mix}}(\ln^2 A_{\max})\gamma^{-k/2}}{\epsilon^2\mu_{\min}(1-\gamma)^8}\ln\Big(\frac{|\mathcal{S}|\ln(\epsilon^{-1}\ln A_{\max})}{\delta\mu_{\min}(1-\gamma)}\Big)\Big] \\
&\ge \frac{650 t_{\text{mix}}}{\mu_{\min}}\ln\Big(\frac{4}{\delta\sqrt{\mu_{\min}}}\Big) = \mathcal{O}\Big[\frac{t_{\text{mix}}}{\mu_{\min}}\ln\Big(\frac{1}{\delta\mu_{\min}}\Big)\Big] \qquad (113)
\end{aligned}
$$

$$
\begin{aligned}
N_{k,t} &= \mathcal{O}\Big[\frac{t_{\text{mix}}A_{\max}(1-\eta\tau)^{-3(t+1)/5}}{\mu_{\min}(1-\gamma)^3}\ln\Big(\frac{|\mathcal{S}|A_{\max}}{\epsilon\delta\mu_{\min}(1-\gamma)}\Big)\Big] \\
&\ge \frac{650 t_{\text{mix}}A_{\max}}{\mu_{\min}}\ln\Big(\frac{20T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\Big) = \mathcal{O}\Big[\frac{t_{\text{mix}}A_{\max}}{\mu_{\min}}\ln\Big(\frac{|\mathcal{S}|A_{\max}}{\epsilon\delta\mu_{\min}(1-\gamma)}\Big)\Big] \qquad (114)
\end{aligned}
$$

$$
\begin{aligned}
\overline{N}_{k,t} &= \mathcal{O}\Big[\frac{t_{\text{mix}}A_{\max}(\ln^{3/2} A_{\max})(1-\eta\tau)^{-3(t+1)/5}}{\mu_{\min}\epsilon^{3/2}(1-\gamma)^6}\ln\Big(\frac{|\mathcal{S}|A_{\max}}{\epsilon\delta\mu_{\min}(1-\gamma)}\Big)\Big] \\
&\ge \frac{650 t_{\text{mix}}A_{\max}}{\mu_{\min}}\ln\Big(\frac{20T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\sqrt{\mu_{\min}}}\Big) = \mathcal{O}\Big[\frac{t_{\text{mix}}A_{\max}}{\mu_{\min}}\ln\Big(\frac{|\mathcal{S}|A_{\max}}{\epsilon\delta\mu_{\min}(1-\gamma)}\Big)\Big], \qquad (115)
\end{aligned}
$$

With these hyperparameters and eqs. (107)&(108), the duality gap bound (20) can be simplified as follows,

$$
D^{(\tau)}(\pi^{(1)}_{K-1}, \pi^{(2)}_{K-1})
$$

$$
\begin{aligned}
&\leq \mathcal{O}\Bigg[\frac{V_{\max}\ln A_{\max}}{1-\gamma}\sum_{k=0}^{K-1}\gamma^{K-k}(1-\eta\tau)^{T_k-1}+\frac{V_{\max}}{1-\gamma}\Big[\frac{t_{\text{mix}}A_{\max}}{\mu_{\min}}\ln\Big(\frac{T_{\text{sum}}|\mathcal{S}|A_{\max}}{\delta\mu_{\min}}\Big)\Big]^{2/3}\\
&\quad\sum_{k=0}^{K-1}\gamma^{K-k-1}\sum_{t=0}^{T_k-1}(1-\eta\tau)^{T_k-2-t}\Big(\frac{1}{N_{k,t}^{2/3}}+\frac{V_{\max}}{\tau\overline{N}_{k,t+1}^{2/3}}\Big)+\frac{V_{\max}^3\gamma^K}{\tau^2(1-\gamma)^2}\\
&\quad+\frac{t_{\text{mix}}V_{\max}^3}{\tau^2\mu_{\min}(1-\gamma)^3}\ln\Big(\frac{K|\mathcal{S}|}{\delta\mu_{\min}}\Big)\sum_{k=0}^{K-1}\frac{\gamma^{K-k-1}}{N_{k+1}}\Bigg]\\
&\overset{(i)}{\leq}\mathcal{O}\Bigg[\frac{\ln A_{\max}}{(1-\gamma)^2}\sum_{k=0}^{K-1}\gamma^{K-k}\gamma^k\\
&\quad+\frac{1}{(1-\gamma)^2}\Big[\frac{t_{\text{mix}}A_{\max}}{\mu_{\min}}\ln\Big(\frac{|\mathcal{S}|A_{\max}}{\epsilon\delta\mu_{\min}(1-\gamma)}\Big)\Big]^{2/3}\sum_{k=0}^{K-1}\gamma^{K-k-1}\sum_{t=0}^{T_k-1}(1-\eta\tau)^{T_k-2-t}\\
&\quad\Bigg(\mathcal{O}\Big[\frac{t_{\text{mix}}A_{\max}}{\mu_{\min}(1-\gamma)^3}\ln\Big(\frac{|\mathcal{S}|A_{\max}}{\epsilon\delta\mu_{\min}(1-\gamma)}\Big)\Big]^{-2/3}(1-\eta\tau)^{2(t+1)/5}\\
&\quad+\mathcal{O}\Big(\frac{\ln A_{\max}}{\epsilon(1-\gamma)^2}\Big)\mathcal{O}\Big[\frac{t_{\text{mix}}A_{\max}(\ln^{3/2}A_{\max})}{\mu_{\min}\epsilon^{3/2}(1-\gamma)^6}\ln\Big(\frac{|\mathcal{S}|A_{\max}}{\epsilon\delta\mu_{\min}(1-\gamma)}\Big)\Big]^{-2/3}(1-\eta\tau)^{2(t+1)/5}\Bigg)\\
&\quad+\frac{\gamma^K(\ln^2A_{\max})}{\epsilon^2(1-\gamma)^7}+\frac{t_{\text{mix}}(\ln^2A_{\max})}{\epsilon^2\mu_{\min}(1-\gamma)^8}\ln\Big(\frac{|\mathcal{S}|\ln(\epsilon^{-1}\ln A_{\max})}{\delta\mu_{\min}(1-\gamma)}\Big)\\
&\quad\sum_{k=0}^{K-1}\gamma^{K-k-1}\mathcal{O}\Big[\frac{t_{\text{mix}}(\ln^2A_{\max})\gamma^{-k/2}}{\epsilon^2\mu_{\min}(1-\gamma)^8}\ln\Big(\frac{|\mathcal{S}|\ln(\epsilon^{-1}\ln A_{\max})}{\delta\mu_{\min}(1-\gamma)}\Big)\Big]^{-1}\Bigg]\\
&\leq\mathcal{O}\Bigg[\frac{\ln A_{\max}}{(1-\gamma)^2}K\gamma^K+\sum_{k=0}^{K-1}\gamma^{K-k-1}\sum_{t=0}^{T_k-1}(1-\eta\tau)^{T_k-1-3(t+1)/5}\\
&\quad+\frac{\gamma^K(\ln^2A_{\max})}{\epsilon^2(1-\gamma)^7}+\sum_{k=0}^{K-1}\gamma^{K-k/2-1}\Bigg]\\
&\overset{(ii)}{\leq}\mathcal{O}\Bigg[\frac{\ln A_{\max}}{(1-\gamma)^2}K\gamma^K+\sum_{k=0}^{K-1}\gamma^{K-1}\frac{(1-\eta\tau)^{-3/5}[(1-\eta\tau)^{-3T_k/5}-1]}{(1-\eta\tau)^{-3/5}-1}\\
&\quad+\frac{\gamma^K(\ln^2A_{\max})}{\epsilon^2(1-\gamma)^7}+\gamma^{K-1}\frac{\gamma^{-K/2}-1}{\gamma^{-1/2}-1}\Bigg]\\
&\overset{(iii)}{\leq}\mathcal{O}\Bigg[\frac{\ln A_{\max}}{(1-\gamma)^2}K\gamma^K+\sum_{k=0}^{K-1}\frac{\gamma^{K-1-3k/5}}{1-(1-\eta\tau)^{3/5}}+\frac{\gamma^K(\ln^2A_{\max})}{\epsilon^2(1-\gamma)^7}+\frac{\gamma^{(K-1)/2}}{1-\gamma^{1/2}}\Bigg]\\
&\overset{(iv)}{\leq}\mathcal{O}\Bigg[\frac{\ln A_{\max}}{(1-\gamma)^2}K\gamma^K+\frac{\gamma^{K-1}(\gamma^{-3K/5}-1)}{\eta\tau(\gamma^{-3/5}-1)}+\frac{\gamma^K(\ln^2A_{\max})}{\epsilon^2(1-\gamma)^7}+\frac{\gamma^{K/2}}{1-\gamma}\Bigg]\\
&\overset{(v)}{\leq}\mathcal{O}\Bigg[\frac{\gamma^K\ln A_{\max}}{(1-\gamma)^3}\ln\Big(\frac{\ln A_{\max}}{\epsilon(1-\gamma)}\Big)+\frac{\gamma^{2K/5}\ln A_{\max}}{\epsilon(1-\gamma)^3}+\frac{\gamma^K(\ln^2A_{\max})}{\epsilon^2(1-\gamma)^7}+\frac{\gamma^{K/2}}{1-\gamma}\Bigg]\\
&\overset{(vi)}{\leq}\mathcal{O}\Bigg[\epsilon+\epsilon^3\sqrt{\frac{1-\gamma}{\ln A_{\max}}}+\frac{\epsilon^{5/2}(1-\gamma)^{2.75}}{\ln^{5/4}A_{\max}}\Bigg]=\mathcal{O}(\epsilon)\overset{(vii)}{\leq}c\epsilon
\end{aligned}
\tag{116}
$$

where (i), (ii) and (iii) use $(1-\eta\tau)^{T_k-1}=\gamma^k$ based on eq. (112), (i) also uses eqs. (104), (107)-(110) & (113)-(115), (iv) uses $1-(1-\eta\tau)^{3/5}=\mathcal{O}(\eta\tau)$, $1-\gamma^{1/2}=\mathcal{O}(1-\gamma)$, (v) uses eqs. (109)-(111), (vi) uses $\mathcal{O}\Big[\frac{\gamma^K\ln A_{\max}}{(1-\gamma)^3}\ln\Big(\frac{\ln A_{\max}}{\epsilon(1-\gamma)}\Big)\Big]\leq\mathcal{O}\Big[\frac{\gamma^K(\ln^2A_{\max})}{\epsilon^2(1-\gamma)^7}\Big]$ and $\gamma^K=\frac{\epsilon^5(1-\gamma)^{7.5}}{\ln^{5/2}A_{\max}}$ implied by eq. (111), and the numeric constant $c>0$ in (vii) exists. By replacing ϵ with ϵ/c, we obtain that

$D_\tau\big(\pi^{(1)}_{K-1}, \pi^{(2)}_{K-1}\big) \leq \epsilon$ which means $(\pi^{(1)}_{K-1}, \pi^{(2)}_{K-1})$ is an ϵ-Nash equilibrium policy pair, and the orders of all the hyperparameter choices remain the same. In this case, the overall sample complexity is given by

$$
\begin{aligned}
&\sum_{k=0}^{K-1}\Big[N_{k+1} + 2\sum_{t=0}^{T_k-1}(N_{k,t} + \overline{N}_{k,t+1})\Big] \\
&\overset{(i)}{=} \mathcal{O}\Big[\frac{t_{\text{mix}}(\ln^2 A_{\max})}{\epsilon^2\mu_{\min}(1-\gamma)^8}\ln\Big(\frac{|\mathcal{S}|\ln(\epsilon^{-1}\ln A_{\max})}{\delta\mu_{\min}(1-\gamma)}\Big)\Big]\sum_{k=0}^{K-1}\gamma^{-k/2} \\
&\quad + \mathcal{O}\Big[\frac{t_{\text{mix}}A_{\max}(\ln^{3/2} A_{\max})}{\mu_{\min}\epsilon^{3/2}(1-\gamma)^6}\ln\Big(\frac{|\mathcal{S}|A_{\max}}{\epsilon\delta\mu_{\min}(1-\gamma)}\Big)\Big]\sum_{k=0}^{K-1}\sum_{t=0}^{T_k-1}(1-\eta\tau)^{-3(t+1)/5} \\
&= \mathcal{O}\Big[\frac{t_{\text{mix}}(\ln^2 A_{\max})}{\epsilon^2\mu_{\min}(1-\gamma)^8}\ln\Big(\frac{|\mathcal{S}|\ln(\epsilon^{-1}\ln A_{\max})}{\delta\mu_{\min}(1-\gamma)}\Big)\Big]\frac{\gamma^{-K/2}-1}{\gamma^{-1/2}-1} \\
&\quad + \mathcal{O}\Big[\frac{t_{\text{mix}}A_{\max}(\ln^{3/2} A_{\max})}{\mu_{\min}\epsilon^{3/2}(1-\gamma)^6}\ln\Big(\frac{|\mathcal{S}|A_{\max}}{\epsilon\delta\mu_{\min}(1-\gamma)}\Big)\Big]\sum_{k=0}^{K-1}\frac{[(1-\eta\tau)^{-3T_k/5}-1]}{1-(1-\eta\tau)^{3/5}} \\
&\overset{(ii)}{\leq} \mathcal{O}\Big[\frac{t_{\text{mix}}(\ln^2 A_{\max})}{\epsilon^2\mu_{\min}(1-\gamma)^8}\ln\Big(\frac{|\mathcal{S}|\ln(\epsilon^{-1}\ln A_{\max})}{\delta\mu_{\min}(1-\gamma)}\Big)\Big]\gamma^{-K/2} \\
&\quad + \mathcal{O}\Big[\frac{t_{\text{mix}}A_{\max}(\ln^{3/2} A_{\max})}{\mu_{\min}\epsilon^{3/2}(1-\gamma)^6}\ln\Big(\frac{|\mathcal{S}|A_{\max}}{\epsilon\delta\mu_{\min}(1-\gamma)}\Big)\frac{\ln A_{\max}}{\epsilon(1-\gamma)^2}\Big]\sum_{k=0}^{K-1}\gamma^{-3k/5} \\
&\leq \mathcal{O}\Big[\frac{t_{\text{mix}}(\ln^2 A_{\max})}{\epsilon^2\mu_{\min}(1-\gamma)^8}\ln\Big(\frac{|\mathcal{S}|\ln(\epsilon^{-1}\ln A_{\max})}{\delta\mu_{\min}(1-\gamma)}\Big)\Big]\gamma^{-K/2} \\
&\quad + \mathcal{O}\Big[\frac{t_{\text{mix}}A_{\max}(\ln^{5/2} A_{\max})}{\mu_{\min}\epsilon^{5/2}(1-\gamma)^8}\ln\Big(\frac{|\mathcal{S}|A_{\max}}{\epsilon\delta\mu_{\min}(1-\gamma)}\Big)\Big]\frac{\gamma^{-3K/5}}{\gamma^{-3/5}-1} \\
&\overset{(iii)}{\leq} \mathcal{O}\Big[\frac{t_{\text{mix}}(\ln^2 A_{\max})}{\epsilon^2\mu_{\min}(1-\gamma)^8}\ln\Big(\frac{|\mathcal{S}|\ln(\epsilon^{-1}\ln A_{\max})}{\delta\mu_{\min}(1-\gamma)}\Big)\Big]\frac{\ln^{5/4} A_{\max}}{\epsilon^{5/2}(1-\gamma)^{3.75}} \\
&\quad + \mathcal{O}\Big[\frac{t_{\text{mix}}A_{\max}(\ln^{5/2} A_{\max})}{\mu_{\min}\epsilon^{5/2}(1-\gamma)^8}\ln\Big(\frac{|\mathcal{S}|A_{\max}}{\epsilon\delta\mu_{\min}(1-\gamma)}\Big)\Big]\frac{\ln^{3/2} A_{\max}}{\epsilon^3(1-\gamma)^{5.5}} \\
&\overset{(iv)}{=} \mathcal{O}\Big[\frac{t_{\text{mix}}A_{\max}(\ln^4 A_{\max})}{\mu_{\min}\epsilon^{5.5}(1-\gamma)^{13.5}}\ln\Big(\frac{|\mathcal{S}|A_{\max}}{\epsilon\delta\mu_{\min}(1-\gamma)}\Big)\Big] \qquad (117) \\
&= \widetilde{\mathcal{O}}\Big(\frac{t_{\text{mix}}A_{\max}}{\mu_{\min}\epsilon^{5.5}(1-\gamma)^{13.5}}\Big)
\end{aligned}
$$

where (i) uses $N_{k,t} \leq \overline{N}_{k,t}$ and eqs. (113)&(115), (ii) uses $\gamma^{-1/2}-1 = \mathcal{O}(1-\gamma)$, $1-(1-\eta\tau)^{3/5} = \mathcal{O}(\eta\tau) = \mathcal{O}\big(\frac{\epsilon(1-\gamma)^2}{\ln A_{\max}}\big)$ and $(1-\eta\tau)^{T_k-1} = \gamma^k$, (iii) uses $\gamma^{-3/5}-1 = \mathcal{O}(1-\gamma)$ and $\gamma^K = \frac{\epsilon^5(1-\gamma)^{7.5}}{\ln^{5/2} A_{\max}}$ implied by the hyperparameter choice (111), and (iv) uses the fact that the second term of (iii) is larger than the first term of (iii). □

F Explanation of Table 1

In this section, we explain Table (1) with more details.

In the column "Model-free", a ✓means that an algorithm does not need any prior knowledge of the environment, including reward mapping and transition kernel.

In the column "Private update", a ✓means that the algorithm updates do not involve the opponent's sensitive information, including action and policy.

In the column "Symmetric update", a ✓means that both players perform symmetric updates.

In the column "Data type", "Markovian" means that the algorithm uses samples queried from the dependent Markov decision process. The "–" sign means no stochastic samples are used.

The sample complexity is defined as the total number of samples required to achieve an ϵ-duality gap for the Markov game. A "–" sign in this column means that such type of sample complexity is unavailable in that paper or its setting is different from ours. We use $\widetilde{\mathcal{O}}$ to hide all the logarithm factors.

We next explain how the sample complexity of (Wei et al., 2021) is calculated. Corollary 4 of (Wei et al., 2021) shows that the iterate-average duality gap is smaller than ϵ with the hyper-parameter choices: number of iterations $T = \widetilde{\mathcal{O}}\Big[\frac{|\mathcal{S}|^2}{\eta^2(1-\gamma)^4\epsilon^2}\Big]$ and number of samples queried per iteration $L = \widetilde{\mathcal{O}}\Big[\frac{A_{\max}^3|\mathcal{S}|^6}{(1-\gamma)^{13}\mu_{\min}\eta^3\epsilon^6}\Big]$ (we replace ξ with ϵ). Hence, the overall sample complexity is

$$LT = \widetilde{\mathcal{O}}\Big[\frac{A_{\max}^3|\mathcal{S}|^8}{(1-\gamma)^{17}\mu_{\min}\eta^5\epsilon^8}\Big] \overset{(i)}{\geq} \widetilde{\mathcal{O}}\Big[\frac{A_{\max}^3|\mathcal{S}|^{10.5}}{(1-\gamma)^{29.5}\mu_{\min}\epsilon^8}\Big],$$

where (i) uses the choice of learning rate $\eta \leq \mathcal{O}\Big[\sqrt{\frac{(1-\gamma)^5}{|\mathcal{S}|}}\Big]$ required by their Algorithm 1.