Finite-Time Analysis of Minimax Q-Learning

Narim Jeong, Donghwan Lee

Keywords: Minimax Q-learning, finite-time analysis, control theory, switched systems.

Summary

The goal of this paper is to present a finite-time analysis of minimax Q-learning and its smooth variant for two-player zero-sum Markov games, where the smooth variant is derived by using the Boltzmann operator. To the best of the authors' knowledge, this is the first work in the literature to provide such results. To facilitate the analysis, we introduce lower and upper comparison systems and employ switching system models. The proposed approach can not only offer a simpler and more intuitive framework for analyzing convergence but also provide deeper insights into the behavior of minimax Q-learning and its smooth variant. These novel perspectives have the potential to reveal new relationships and foster synergy between ideas in control theory and reinforcement learning.

Contribution(s)

1. This paper presents a finite-time analysis of minimax Q-learning and its smooth variant with the Boltzmann operator, which is the first work to provide such results, as far as the authors are aware.
   Context: Most of the existing literature addresses its asymptotic convergence (Littman, 2001; Zhu & Zhao, 2020) or the convergence of the modified algorithms (Diddigi et al., 2022; Fan et al., 2019). Compared to others, our method can provide stronger convergence results through the finite-time analysis. Moreover, this paper addresses the vanilla minimax Q-learning and its smooth variant, which are based on the independently and identically distributed observations and the constant step-size in the tabular domain. Although we utilize these settings to simplify the analysis, our approach can be expanded to include more complex Markovian observation models by employing the methods described in Srikant & Ying (2019) and Bhandari et al. (2018).
2. By employing the switching system model for the convergence analysis, this paper contributes new insights into the convergence analysis of minimax Q-learning and the recently developed switching system framework for the finite-time analysis of Q-learning (Lee et al., 2022).
   Context: It is noteworthy to highlight that while the switching system model introduced in Lee et al. (2022) has been used as a basis, the main analysis and proof in this work significantly differed from those in Lee et al. (2022). In addition, we present a simulation result to empirically validate our method for the convergence analysis of the minimax Q-learning and its smooth variant that makes use of the switching system model.
3. This paper suggests the theoretically straightforward convergence analysis based on the control-theoretic concepts.
   Context: On the basis of the simple analytical approach, our analysis will help to reveal new relationships and promote mutual understanding between the control theory and RL.

Finite-Time Analysis of Minimax Q-Learning

Narim Jeong, Donghwan Lee
{nrjeong,donghwan}@kaist.ac.kr

Department of Electrical Engineering, KAIST, South Korea

Abstract

The goal of this paper is to present a finite-time analysis of minimax Q-learning and its smooth variant for two-player zero-sum Markov games, where the smooth variant is derived by using the Boltzmann operator. To the best of the authors' knowledge, this is the first work in the literature to provide such results. To facilitate the analysis, we introduce lower and upper comparison systems and employ switching system models. The proposed approach can not only offer a simpler and more intuitive framework for analyzing convergence but also provide deeper insights into the behavior of minimax Q-learning and its smooth variant. These novel perspectives have the potential to reveal new relationships and foster synergy between ideas in control theory and reinforcement learning.

1 Introduction

Reinforcement learning (RL) can solve sequential decision-making problems in Markov decision processes (Sutton & Barto, 1998). Both the theoretical and practical sides of RL algorithms have seen a rise in interest recently due to their ability to outperform humans in a variety of difficult tasks (Mnih et al., 2015; Wang et al., 2016; Lillicrap et al., 2016; Heess et al., 2015; Hasselt et al., 2015; Bellemare et al., 2017; Schulman et al., 2015; 2017). One of the most fundamental and widely used RL algorithms, Q-learning (Watkins & Dayan, 1992), has been extensively studied for its convergence over decades. The primary focus of the convergence analysis has been on the asymptotic convergence (Tsitsiklis, 1994; Jaakkola et al., 1994; Borkar & Meyn, 2000; Hasselt, 2010; Melo et al., 2008; Lee & He, 2020b; Devraj & Meyn, 2017); however, recent research has concentrated on the finite-time convergence analysis (Szepesvári, 1998; Kearns & Singh, 1999; Even-Dar & Mansour, 2003; Azar et al., 2011; Beck & Srikant, 2012; Wainwright, 2019; Qu & Wierman, 2020; Li et al., 2020; Chen et al., 2021; Lee & He, 2020a), which measures how quickly the iterations approach an optimal solution. In most previous works, Q-learning dynamics was treated as nonlinear stochastic approximations (Kushner & Yin, 2003), and the contraction mapping of the Bellman operator was used for the convergence analysis (Beck & Srikant, 2012; Qu & Wierman, 2020; Chen et al., 2021; Lee & He, 2020a). Meanwhile, Lee & He (2020b) and Lee et al. (2022) presented a new viewpoint on Q-learning convergence based on the switching system models (Liberzon, 2003). This perspective captures distinctive features of Q-learning dynamics, and it served as a motivation for the development of this paper. The main results of Lee & He (2020b) and Lee et al. (2022) were reached by converting the finite-time convergence analysis into the stability analysis of the dynamic control systems.

In this paper, the Q-learning algorithm is examined for a more general Markov decision process: a two-player zero-sum Markov game (Shapley, 1953) in which two decision-making agents compete against one another. More precisely, this paper aims to provide the finite-time analysis of the minimax Q-learning (MQL) (Littman, 1994) and its smooth form with the Boltzmann operator (Sutton & Barto, 1998). In order to analyze the convergence of the MQL, the switching system models presented in Lee et al. (2022) are utilized. By establishing the upper and lower comparison systems of

the original MQL, the finite-time error bound of MQL can be examined using the control-theoretic concepts.

The main contributions of this paper can be summarized as follows:

- This paper presents a finite-time analysis of MQL and its smooth variant with the Boltzmann operator. To the authors' knowledge, this is the first work to provide such results. Note that most of the existing literature addresses its asymptotic convergence (Littman, 2001; Zhu & Zhao, 2020) or the convergence of the modified algorithms (Diddigi et al., 2022; Fan et al., 2019). Compared to others, our method can provide a stronger convergence result of the vanilla MQL through the finite-time analysis.
- By employing the switching system model for the convergence analysis, this paper contributes new insights into the convergence analysis of minimax Q-learning and the recently developed switching system framework for the finite-time analysis of Q-learning (Lee et al., 2022).
- This paper suggests the theoretically straightforward convergence analysis based on the control-theoretic concepts. Our approach will help to reveal new relationships and promote mutual understanding between the control theory and RL.

2 Preliminaries and problem formulation

Two-player zero-sum Markov game. In this paper, a two-player zero-sum Markov game (Shapley, 1953) is considered in which two agents choose actions to compete with each other. These two agents will be referred to as a user and an adversary, where the user aims to maximize the return while the adversary attempts to impede the user by minimizing the return. Here, the state-space can be denoted as $\mathcal{S} := \{1, 2, \ldots, |\mathcal{S}|\}$, the action-space of the user as $\mathcal{A} := \{1, 2, \ldots, |\mathcal{A}|\}$, and the action-space of the adversary as $\mathcal{B} := \{1, 2, \ldots, |\mathcal{B}|\}$. Moreover, $|\mathcal{S}|$ is the number of states, $|\mathcal{A}|$ is the number of user's actions, and $|\mathcal{B}|$ is the number of adversary's actions. For the alternating two-player Markov games at the k-th iteration, the user chooses an action $a_k \in \mathcal{A}$ at the state $s_k \in \mathcal{S}$ using the user's policy π without having access to the adversary's action. Then, the adversary selects an action $b_k \in \mathcal{B}$ by utilizing the user's action $a_k \in \mathcal{A}$ and the adversary's policy μ. For both agents, the state s_k moves to the next state s'_k with the state transition probability $P(s'_k|s_k, a_k, b_k)$, and this transition results in a reward $r(s_k, a_k, b_k, s'_k)$. For convenience, we assume a deterministic reward function $r(s_k, a_k, b_k, s'_k) =: r_k$. After that, they take turns choosing actions with the goal of maximizing and minimizing the cumulative discounted rewards, respectively.

It is well known that there exists an optimal policy (Littman, 1994) for both the user and the adversary. The goal of the Markov game is to discover the user's optimal policy π^* and the adversary's optimal policy μ^*:

$$(\pi^*, \mu^*) := \arg\max_{\pi\in\Theta} \min_{\mu\in\Omega} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_k \middle| \pi, \mu\right],$$

where Θ is the set of all admissible policies of the user, Ω is the set of all admissible policies of the adversary, $\mathbb{E}[\cdot|\pi, \mu]$ is an expectation subject to the policies π and μ, and $\gamma \in [0, 1)$ is the discount factor. Moreover, the optimal Q-function in two-player Markov games can be defined as

$$Q^*(s, a, b) := \max_{\pi\in\Theta} \min_{\mu\in\Omega} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_k \middle| s_0 = s, a_0 = a, b_0 = b, \pi, \mu\right],$$

which satisfies the optimal Q-Bellman equation $Q^*(s, a, b) = R(s, a, b) + \gamma \sum_{s'\in S} P(s'|s, a, b) \max_{a'\in\mathcal{A}} \min_{b'\in\mathcal{B}} Q^*(s', a', b')$ with the expected reward $R(s, a, b)$. Then the user's optimal policy can be obtained as

$$\pi^*(s) = \arg\max_{a\in\mathcal{A}} \min_{b\in\mathcal{B}} Q^*(s, a, b)$$

and the adversary's optimal policy as

$$\mu^*(s,a) = \arg\min_{b\in\mathcal{B}} Q^*(s,a,b).$$

While the user and adversary can both learn their optimal policies, this study will solely take the role of the user into account.

Switching system. A switching system (Liberzon, 2003) is a specific type of nonlinear system (Khalil, 2002) that functions across several subsystems via switching signals. Although there are various kinds of switching systems, we employ *affine switching system* in this paper:

$$x_{k+1} = A_{\sigma_k} x_k + b_{\sigma_k},$$

where $\sigma_k \in \mathcal{M} := \{1, 2, \ldots, \mathcal{M}\}$ is the switching signal, $A_{\sigma_k} \in \mathbb{R}^{n\times n}$ is the subsystem matrix, and $b_{\sigma_k} \in \mathbb{R}^n$ is the subsystem vector. A_{σ_k} and b_{σ_k} are altered according to σ_k, and σ_k can be chosen either arbitrarily or by policy. With the extra vector b_{σ_k}, it becomes challenging to make the switching system stable.

Minimax Q-learning and its smooth variant. MQL (Littman, 1994) was introduced to solve two-player zero-sum Markov games. We utilize

$$Q_{k+1}(s_k,a_k,b_k) = Q_k(s_k,a_k,b_k) + \alpha\left\{r_k + \gamma \max_{a\in\mathcal{A}}\min_{b\in\mathcal{B}} Q_k\left(s'_k,a,b\right) - Q_k(s_k,a_k,b_k)\right\} \quad (1)$$

as the MQL update equation, where $Q(s,a,b)$ is the Q-function of MQL and α is the step-size. Moreover, the smooth version of MQL with the Boltzmann operator (Sutton & Barto, 1998), which will be called *Boltzmann MQL*, can be represented as

$$\begin{aligned} Q^{bz}_{k+1}(s_k,a_k,b_k) =& Q^{bz}_k(s_k,a_k,b_k) \\ &+ \alpha\left\{r_k + \gamma h^{\omega}_{a\in\mathcal{A}}\left(h^{-\omega}_{b\in\mathcal{B}}\left(Q^{bz}_k\left(s'_k,a,b\right)\right)\right) - Q^{bz}_k(s_k,a_k,b_k)\right\} \end{aligned} \quad (2)$$

with the Q-function of the Boltzmann MQL $Q^{bz}(s,a,b)$,

$$h^{\omega}_{u\in\mathcal{U}}(v) := \frac{\sum_{u\in\mathcal{U}} v(u)\exp(v(u)\omega)}{\sum_{u\in\mathcal{U}}\exp(v(u)\omega)} \text{ and } h^{-\omega}_{u\in\mathcal{U}}(v) := \frac{\sum_{u\in\mathcal{U}} v(u)\exp(-v(u)\omega)}{\sum_{u\in\mathcal{U}}\exp(-v(u)\omega)} \quad (3)$$

for every $v \in \mathbb{R}^{|\mathcal{U}|}$. Here, $h^{\omega}_{u\in\mathcal{U}}(v)$ is a smooth approximation of the $\max$ operator, whereas $h^{-\omega}_{u\in\mathcal{U}}(v)$ is a smooth approximation of the $\min$ operator. The smooth variations have been shown to enhance the exploration and performance of the algorithm while also mitigating Q-learning's overestimation bias (Song et al., 2019; Pan et al., 2019). Furthermore, the parameter $\omega > 0$ determines the sharpness of the Boltzmann operators. A larger ω produces a sharper approximation of the $\max$ operator in $h^{\omega}_{u\in\mathcal{U}}$ and a sharper approximation of the $\min$ operator in $h^{-\omega}_{u\in\mathcal{U}}$.

Assumption 1. *The following conditions are assumed in this paper:*

(i) The state-action distribution, defined as $d(s,a,b) = p(s)\beta(a|s)\phi(b|s)$ with the stationary state distribution p and the behavior policies β and ϕ, is strictly positive $d(s,a,b) > 0$ for every $s \in \mathcal{S}, a \in \mathcal{A}$, and $b \in \mathcal{B}$.

(ii) We use the constant step-size: $\alpha \in (0,1)$.

(iii) The reward is unit-bounded: $\max_{(s,a,b,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}\times\mathcal{S}} |r(s,a,b,s')| \le 1$.

(iv) Q-functions have unit-bounded initial values: $\|Q_0\|_\infty \le 1$ and $\left\|Q^{bz}_0\right\|_\infty \le 1$.

In Assumption 1, the positive value of $d(s,a,b)$ guarantees that every state-action pair can be visited an infinite number of times, allowing the adequate exploration. Furthermore, a constant value of α eliminates the need for further assumptions in the convergence proof. Unit-bounded reward and initial values of the Q-functions are introduced to simplify the analysis without losing generality. Other general assumptions used in this paper are described in Appendix B.

Definition 1. *The following quantities are defined for convenience, which will be used often throughout this paper:*

(i) Maximum state-action distribution: $d_{\max} := \max_{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}} d(s,a,b) \in (0,1)$.

(ii) Minimum state-action distribution: $d_{\min} := \min_{(s,a,b)\in\mathcal{S}\times\mathcal{A}\times\mathcal{B}} d(s,a,b) \in (0,1)$.

(iii) Exponential decay rate: $\rho := 1 - \alpha d_{\min}(1-\gamma) \in (0,1)$.

(iv) Q-function vector:

$$Q := \begin{bmatrix} Q_{1,1} \\ \vdots \\ Q_{|\mathcal{A}|,|\mathcal{B}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|},$$

where $Q_{a,b} \in \mathbb{R}^{|\mathcal{S}|}$ is a vector that lists the Q-function for the action $a \in \mathcal{A}$ and $b \in \mathcal{B}$.

From the definition of the Q-function vector, the single $Q(s,a,b)$ value can be retrieved by $Q(s,a,b) = (e_a \otimes e_b \otimes e_s)^T Q$, where $e_s \in \mathbb{R}^{|\mathcal{S}|}$, $e_a \in \mathbb{R}^{|\mathcal{A}|}$, and $e_b \in \mathbb{R}^{|\mathcal{B}|}$ are the s-th, a-th, and b-th basis vector, respectively. The k-th basis vector has a value of 1 in the k-th component, whereas the remaining values are 0.

3 Finite-time analysis of minimax Q-learning

3.1 Overview of the proposed analysis for minimax Q-learning

In Lee & He (2020b) and Lee et al. (2022), the convergence analysis of Q-learning was transformed into the stability analysis of the switching system for simplicity. Based on this, our approach also uses a similar strategy, which reduces the convergence of the MQL problem to analyzing the stability of the affine switching system. In Section 3, two straightforward comparison iterations are used, which are easier to analyze: the *upper iteration* sets upper bounds on the MQL trajectories, while the *lower iteration* sets lower bounds. Once two comparison iterations are established, the findings suggested in Lee (2024) are utilized to determine the convergence of each comparison iteration, which yields the MQL's finite-time error bound.

3.2 Upper and lower iterations of minimax Q-learning

Based on (1), the upper iteration of MQL can be represented as

$$Q^U_{k+1}(s_k,a_k,b_k) = Q^U_k(s_k,a_k,b_k) + \alpha\left\{r_k + \gamma \max_{a\in\mathcal{A}} Q^U_k\left(s'_k, a, \mu^*\left(s'_k, a\right)\right) - Q^U_k(s_k,a_k,b_k)\right\}$$

and the lower iteration of MQL as

$$Q^L_{k+1}(s_k,a_k,b_k) = Q^L_k(s_k,a_k,b_k) + \alpha\left\{r_k + \gamma \min_{b\in\mathcal{B}} Q^L_k\left(s'_k, \pi^*\left(s'_k\right), b\right) - Q^L_k(s_k,a_k,b_k)\right\},$$

which can be proved by the following propositions:

Proposition 1. *Assume that $Q^U_0(s,a,b) \geq Q_0(s,a,b)$ for every (s,a,b). Then, $Q^U_k(s,a,b) \geq Q_k(s,a,b)$ for every (s,a,b) and $k \geq 0$.*

Proposition 2. *Assume that $Q^L_0(s,a,b) \leq Q_0(s,a,b)$ for every (s,a,b). Then, $Q^L_k(s,a,b) \leq Q_k(s,a,b)$ for every (s,a,b) and $k \geq 0$.*

Proposition 1 and Proposition 2 can be proven by using an induction argument. The complete proofs can be found in Appendix C.

3.3 Finite-time error bound for the minimax Q-learning

From examining two comparison iterations, it is clear that both upper and lower iterations of MQL resemble the Q-learning iterates. In this case, the following lemma can be utilized, which shows the convergence of Q-learning by using the switching system approach.

Lemma 1 (Lee (2024)). *Assuming that $k \geq 0$ and $Q_k^q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$ is the Q-function of Q-learning with $\|Q_0^q\|_\infty \leq 1$. Then,*

$$\mathbb{E}\left[\|Q_k^q - Q^*\|_\infty\right] \leq \frac{9\alpha^{\frac{1}{2}} d_{\max}|\mathcal{S}\times\mathcal{A}|}{d_{\min}^{\frac{3}{2}}(1-\gamma)^{\frac{5}{2}}} + \frac{2|\mathcal{S}\times\mathcal{A}|^{\frac{3}{2}}}{1-\gamma}\rho^k + \frac{4\alpha\gamma d_{\max}|\mathcal{S}\times\mathcal{A}|^{\frac{2}{3}}}{1-\gamma}k\rho^{k-1}.$$

With that, we can demonstrate the convergence of MQL as follows:

Theorem 1. *For all $k \geq 0$,*

$$\begin{aligned}\mathbb{E}\left[\|Q_k - Q^*\|_2\right] \leq & \frac{27\alpha^{\frac{1}{2}} d_{\max}|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}{d_{\min}^{\frac{3}{2}}(1-\gamma)^{\frac{5}{2}}} + \frac{6|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{3}{2}}}{1-\gamma}\rho^k \\ & + \frac{12\alpha\gamma d_{\max}|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{2}{3}}}{1-\gamma}k\rho^{k-1}.\end{aligned} \tag{4}$$

Proof sketch. Using Definition 1 and the symmetric property, the finite-time error bound for the upper and lower iterations of MQL can be derived from Lemma 1. Afterward, the final conclusion can be obtained by using the relation $\mathbb{E}\left[\|Q_k - Q^*\|_2\right] \leq 2\mathbb{E}\left[\left\|Q_k^L - Q^*\right\|_2\right] + \mathbb{E}\left[\left\|Q_k^U - Q^*\right\|_2\right]$. A detailed proof is provided in Appendix D. □

By examining the right side of (4), it is apparent that the first term is a constant error that can be reduced by the smaller α. Additionally, the second and last terms diminish as k goes to infinity with $\rho \in (0,1)$ in Definition 1.

4 Finite-time analysis of Boltzmann MQL

4.1 Overview of the proposed analysis for Boltzmann MQL

In order to determine the convergence of Boltzmann MQL, we once again transfer this problem into the stability analysis of the affine switching system. However, it is difficult to verify the stability since the upper and lower iterates contain the additional terms. Therefore, we further convert them into the upper and lower comparison systems. After that, the finite-time error bounds of the upper and lower comparison systems are demonstrated, respectively. Then, the final convergence result of the Boltzmann MQL can be obtained.

4.2 Upper and lower iterations of Boltzmann MQL

Based on (2), the upper iteration of Boltzmann MQL can be established as

$$\begin{aligned}Q_{k+1}^{U,bz}(s_k,a_k,b_k) = & Q_k^{U,bz}(s_k,a_k,b_k) + \alpha\Big\{r_k \\ & + \gamma\max_{a\in\mathcal{A}} Q_k^{U,bz}\left(s_k', a, \mu^*\left(s_k', a\right)\right) - Q_k^{U,bz}(s_k,a_k,b_k) + \gamma\frac{\ln(|\mathcal{B}|)}{\omega}\Big\}\end{aligned} \tag{5}$$

and the lower iteration of Boltzmann MQL as

$$\begin{aligned}Q_{k+1}^{L,bz}(s_k,a_k,b_k) = & Q_k^{L,bz}(s_k,a_k,b_k) + \alpha\Big\{r_k \\ & + \gamma\min_{b\in\mathcal{B}} Q_k^{L,bz}\left(s_k', \pi^*\left(s_k'\right), b\right) - Q_k^{L,bz}(s_k,a_k,b_k) - \gamma\frac{\ln(|\mathcal{A}|)}{\omega}\Big\},\end{aligned} \tag{6}$$

which can be proved by the following propositions:

Proposition 3. *Assume that $Q_0^{U,bz}(s,a,b) \geq Q_0^{bz}(s,a,b)$ for every (s,a,b). Then, $Q_k^{U,bz}(s,a,b) \geq Q_k^{bz}(s,a,b)$ for every (s,a,b) and $k \geq 0$.*

Proposition 4. *Assume that $Q_0^{L,bz}(s,a,b) \leq Q_0^{bz}(s,a,b)$ for every (s,a,b). Then, $Q_k^{L,bz}(s,a,b) \leq Q_k^{bz}(s,a,b)$ for every (s,a,b) and $k \geq 0$.*

Propositions 3 and 4 can be demonstrated in a similar manner to Propositions 1 and 2, with detailed proofs available in Appendix E.

4.3 Upper and lower comparison systems

The convergence proof of the Boltzmann MQL differs slightly from the previous section due to the additional terms $\alpha\gamma\frac{\ln(|\mathcal{B}|)}{\omega}$ and $-\alpha\gamma\frac{\ln(|\mathcal{A}|)}{\omega}$ in the two comparison iterations (5) and (6). For the dynamical system representations of the upper and lower iterates of the Boltzmann MQL, the following notations are provided first:

Definition 2. *Throughout the paper, we will use the following notations:*

$$R_{a,b} := \begin{bmatrix} R(1,a,b) \\ R(2,a,b) \\ \vdots \\ R(|\mathcal{S}|,a,b) \end{bmatrix}, R := \begin{bmatrix} R_{1,1} \\ \vdots \\ R_{|\mathcal{A}|,|\mathcal{B}|} \end{bmatrix},$$

$$P_{a,b} := \begin{bmatrix} P(1|1,a,b) & P(2|1,a,b) & \cdots & P(|\mathcal{S}||1,a,b) \\ P(1|2,a,b) & P(2|2,a,b) & \cdots & P(|\mathcal{S}||2,a,b) \\ \vdots & \vdots & \ddots & \vdots \\ P(1||\mathcal{S}|,a,b) & P(2||\mathcal{S}|,a,b) & \cdots & P(|\mathcal{S}|||\mathcal{S}|,a,b) \end{bmatrix}, P := \begin{bmatrix} P_{1,1} \\ \vdots \\ P_{|\mathcal{A}|,|\mathcal{B}|} \end{bmatrix},$$

$$D_{a,b} := \begin{bmatrix} d(1,a,b) & & \\ & \ddots & \\ & & d(|S|,a,b) \end{bmatrix}, D := \begin{bmatrix} D_{1,1} & & \\ & \ddots & \\ & & D_{|\mathcal{A}|,|\mathcal{B}|} \end{bmatrix},$$

where $R_{a,b} \in \mathbb{R}^{|\mathcal{S}|}$ indicates the expected reward vector, $P_{a,b} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ represents the state transition probability matrix, and $D_{a,b} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ shows the nonsingular diagonal state-action distribution matrix. Note that $R_{a,b}$, $P_{a,b}$, and $D_{a,b}$ depend on the action pair $(a,b) \in \mathcal{A}\times\mathcal{B}$. Moreover, $R \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}$, $P \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|\times|\mathcal{S}|}$, and $D \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|\times|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}$.

Definition 3. *Let us define the greedy policies with regard to $Q \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}$ as $i(s) := \arg\max_{a\in\mathcal{A}} Q(s,a,\mu^*(s,a)) \in \mathcal{A}$ and $j(s) := \arg\min_{b\in\mathcal{B}} Q(s,\pi^*(s),b) \in \mathcal{B}$. The action transition matrix for $i(s)$ is denoted as Π_Q and for $j(s)$ as Γ_Q, where*

$$\Pi_Q := \begin{bmatrix} e_{i(1)}^T \otimes e_{\mu^*(1,i(1))}^T \otimes e_1^T \\ e_{i(2)}^T \otimes e_{\mu^*(2,i(2))}^T \otimes e_2^T \\ \vdots \\ e_{i(|\mathcal{S}|)}^T \otimes e_{\mu^*(|\mathcal{S}|,i(|\mathcal{S}|))}^T \otimes e_{|\mathcal{S}|}^T \end{bmatrix}, \Gamma_Q := \begin{bmatrix} e_{\pi^*(1)}^T \otimes e_{j(1)}^T \otimes e_1^T \\ e_{\pi^*(2)}^T \otimes e_{j(2)}^T \otimes e_2^T \\ \vdots \\ e_{\pi^*(|\mathcal{S}|)}^T \otimes e_{j(|\mathcal{S}|)}^T \otimes e_{|\mathcal{S}|}^T \end{bmatrix},$$

with $e_{i(s)}, e_{\pi^*(s)} \in \mathbb{R}^{|\mathcal{A}|}$ and $e_{\mu^*(s,i(s))}, e_{j(s)} \in \mathbb{R}^{|\mathcal{B}|}$. Then, we can express the max *and* min *operators in vector form using the Q-function vector from Definition 1 as*

$$\Pi_Q Q := \begin{bmatrix} \max_{a\in\mathcal{A}} Q(1,a,\mu^*(1,a)) \\ \max_{a\in\mathcal{A}} Q(2,a,\mu^*(2,a)) \\ \vdots \\ \max_{a\in\mathcal{A}} Q(|\mathcal{S}|,a,\mu^*(|\mathcal{S}|,a)) \end{bmatrix}, \ \Gamma_Q Q := \begin{bmatrix} \min_{b\in\mathcal{B}} Q(1,\pi^*(1),b) \\ \min_{b\in\mathcal{B}} Q(2,\pi^*(2),b) \\ \vdots \\ \min_{b\in\mathcal{B}} Q(|\mathcal{S}|,\pi^*(|\mathcal{S}|),b) \end{bmatrix}.$$

Note that $\Pi_Q, \Gamma_Q \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}$ and $\Pi_Q Q, \Gamma_Q Q \in \mathbb{R}^{|\mathcal{S}|}$. An important characteristic of Definitions 2 and 3 is that $P\Pi_Q$ and $P\Gamma_Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}||\mathcal{B}|\times|\mathcal{S}||\mathcal{A}||\mathcal{B}|}$ are the transition probability matrix under the policy with the $\max$ and $\min$ operators, respectively.

Using Definitions 2 and 3, the upper comparison system of Boltzmann MQL can be written as

$$Q_{k+1}^{U,bz} - Q^* = A_{Q_k^{U,bz}}\left(Q_k^{U,bz} - Q^*\right) + b_{Q_k^{U,bz}} + \alpha w_k^U + \alpha\gamma D\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1} \tag{7}$$

with

$$\begin{aligned} A_Q &:= I + \alpha\left(\gamma DP\Pi_Q - D\right), \\ b_Q &:= \alpha\gamma DP\left(\Pi_Q - \Pi_{Q^*}\right)Q^*, \end{aligned} \tag{8}$$

and

$$\begin{aligned} w_k^U =& (e_{a_k}\otimes e_{b_k}\otimes e_{s_k})r_k + \gamma(e_{a_k}\otimes e_{b_k}\otimes e_{s_k})\left(e_{s'_k}\right)^T \Pi_{Q_k^{U,bz}} Q_k^{U,bz} \\ &- (e_{a_k}\otimes e_{b_k}\otimes e_{s_k})(e_{a_k}\otimes e_{b_k}\otimes e_{s_k})^T Q_k^{U,bz} + \gamma(e_{a_k}\otimes e_{b_k}\otimes e_{s_k})\frac{\ln(|\mathcal{B}|)}{\omega} \\ &- \left(DR + \gamma DP\Pi_{Q_k^{U,bz}} Q_k^{U,bz} - DQ_k^{U,bz} + \gamma D\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}\right), \end{aligned} \tag{9}$$

where $\mathbf{1}$ is a column vector in which all the values are 1. Moreover, the lower comparison system of Boltzmann MQL can be written as

$$Q_{k+1}^{L,bz} - Q^* = A'_{Q_k^{L,bz}}\left(Q_k^{L,bz} - Q^*\right) + b'_{Q_k^{L,bz}} + \alpha w_k^L - \alpha\gamma D\frac{\ln(|\mathcal{A}|)}{\omega}\mathbf{1} \tag{10}$$

with

$$\begin{aligned} A'_Q &:= I + \alpha\left(\gamma DP\Gamma_Q - D\right), \\ b'_Q &:= \alpha\gamma DP\left(\Gamma_Q - \Gamma_{Q^*}\right)Q^*, \end{aligned}$$

and

$$\begin{aligned} w_k^L =& (e_{a_k}\otimes e_{b_k}\otimes e_{s_k})r_k + \gamma(e_{a_k}\otimes e_{b_k}\otimes e_{s_k})\left(e_{s'_k}\right)^T \Gamma_{Q_k^{L,bz}} Q_k^{L,bz} \\ &- (e_{a_k}\otimes e_{b_k}\otimes e_{s_k})(e_{a_k}\otimes e_{b_k}\otimes e_{s_k})^T Q_k^{L,bz} - \gamma(e_{a_k}\otimes e_{b_k}\otimes e_{s_k})\frac{\ln(|\mathcal{A}|)}{\omega} \\ &- \left(DR + \gamma DP\Gamma_{Q_k^{L,bz}} Q_k^{L,bz} - DQ_k^{L,bz} - \gamma D\frac{\ln(|\mathcal{A}|)}{\omega}\mathbf{1}\right). \end{aligned}$$

The steps taken to construct (7) and (10) are described in Appendix F. At this point, (7) and (10) can be seen as stochastic affine switching systems with an additional affine vector $b_{Q_k^{U,bz}}$ or $b'_{Q_k^{L,bz}}$ and a stochastic noise w_k^U or w_k^L. Moreover, the greedy policy changes the values of A_Q, b_Q, A'_Q, and b'_Q.

4.4 Finite-time error bound for the Boltzmann MQL

Using (7) and (10), the following theorems can be utilized to demonstrate the finite-time error bound of the upper and lower comparison systems:

Theorem 2. *For every $k \geq 0$,*

$$\begin{aligned} \mathbb{E}\left[\left\|Q_k^{U,bz} - Q^*\right\|_\infty\right] \leq& \frac{4\alpha\gamma d_{max}|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}{1-\gamma}k\rho^{k-1} \\ &+ \frac{6\sqrt{2}\alpha^{\frac{1}{2}} d_{max}(\ln(|\mathcal{B}|)+\omega)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^{\frac{3}{2}}(1-\gamma)^{\frac{5}{2}}} \\ &+ \frac{3\gamma d_{max}^2\ln(|\mathcal{B}|)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^2(1-\gamma)^2} + \frac{2|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}{1-\gamma}\rho^k. \end{aligned}$$

Theorem 3. *For every $k \geq 0$,*

$$\mathbb{E}\left[\left\|Q_k^{L,bz} - Q^*\right\|_\infty\right] \leq \frac{4\alpha\gamma d_{max}|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}{1-\gamma}k\rho^{k-1} + \frac{6\sqrt{2}\alpha^{\frac{1}{2}}d_{max}(\ln(|\mathcal{A}|)+\omega)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^{\frac{3}{2}}(1-\gamma)^{\frac{5}{2}}} + \frac{3\gamma d_{max}^2\ln(|\mathcal{A}|)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^2(1-\gamma)^2} + \frac{2|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}{1-\gamma}\rho^k.$$

The whole proofs are accessible in Appendix G. Then, the convergence of Boltzmann MQL can be demonstrated as follows:

Theorem 4. *For every $k \geq 0$,*

$$\begin{aligned}\mathbb{E}\left[\left\|Q_k^{bz} - Q^*\right\|_2\right] \leq & \frac{12\alpha\gamma d_{max}|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}{1-\gamma}k\rho^{k-1} \\ & + \frac{18\sqrt{2}\alpha^{\frac{1}{2}}d_{max}\left(\max\left(\ln(|\mathcal{A}|),\ln(|\mathcal{B}|)\right)+\omega\right)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^{\frac{3}{2}}(1-\gamma)^{\frac{5}{2}}} \\ & + \frac{9\gamma d_{max}^2\max\left(\ln(|\mathcal{A}|),\ln(|\mathcal{B}|)\right)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^2(1-\gamma)^2} + \frac{6|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}{1-\gamma}\rho^k.\end{aligned} \tag{11}$$

Proof sketch. It employs a similar idea to Theorem 1 and applies Theorems 2 and 3. The whole proof is accessible in Appendix N. □

As $k \to \infty$, the first term $O(k\rho^{k-1})$ and the last term $O(\rho^k)$ on the right side of (11) decrease exponentially. Furthermore, choosing a lower value of α and a bigger value of ω can minimize the second and third constant error terms.

5 Numerical simulation

A simulation study is provided to show the convergence of MQL and Boltzmann MQL, with the construction of their comparison systems. We consider a simple Markov game as given in Appendix O.

Figure 1 shows the simulated trajectories of the MQL and its upper and lower iterations. Furthermore, Figure 2 depicts the simulated trajectories of the Boltzmann MQL with its upper and lower comparison systems. These results empirically validate the bounding concepts for the convergence analysis of MQL and its smooth variant. Note that the trajectories of the MQL and the Boltzmann MQL do not necessarily converge to the optimal Q-function value since both the user's and the adversary's actions have an impact on them.

6 Related work

MQL was first presented by Littman (1994), which is a kind of Q-learning developed for two-player zero-sum Markov games. Based on this, Hu et al. (1998) introduced Nash Q-learning by extending MQL into multi-agent environments, and Lagoudakis & Parr (2002) investigated a value iteration of MQL and suggested least-squared policy iteration algorithm for solving the two-player Markov games. Additionally, Littman & Szepesvári (1996) demonstrated the asymptotic convergence of MQL using game theory, Bowling (2000) studied its convergence conditions, and Hu & Wellman (2003) analyzed its convergence behavior and emphasized the restrictions of the convergence assumptions. Littman et al. (2001) presented friend-or-foe Q-learning for general-sum Markov games, which outperformed Nash Q-learning in terms of convergence, and Littman (2001) examined Nash Q-learning convergence and behavior.

Although not specifically focusing on MQL, a number of noteworthy studies on Markov games provided significant insights. An approximate dynamic programming framework for two-player

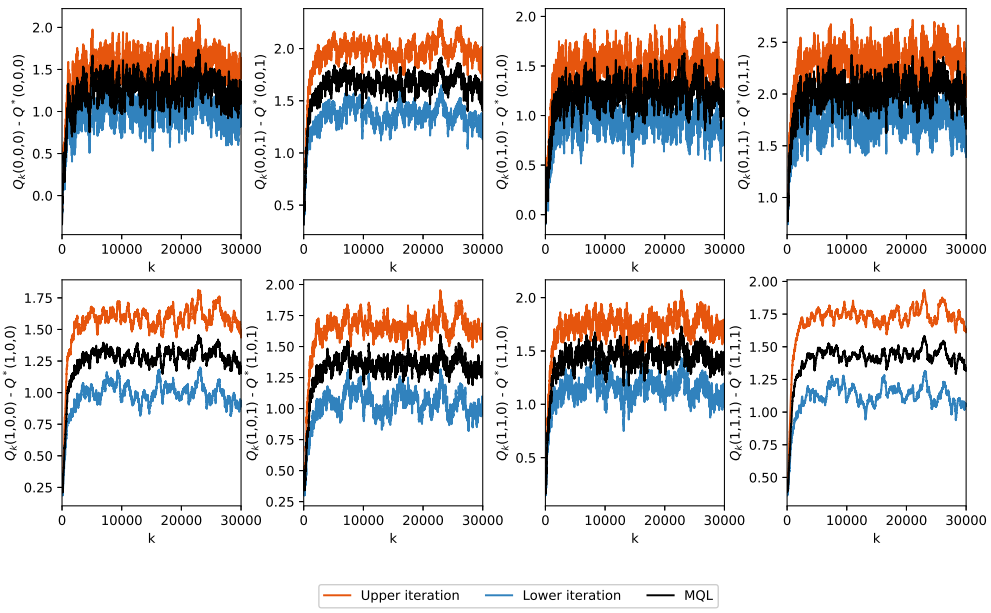

Figure 1: Trajectories of MQL and its comparison iterations

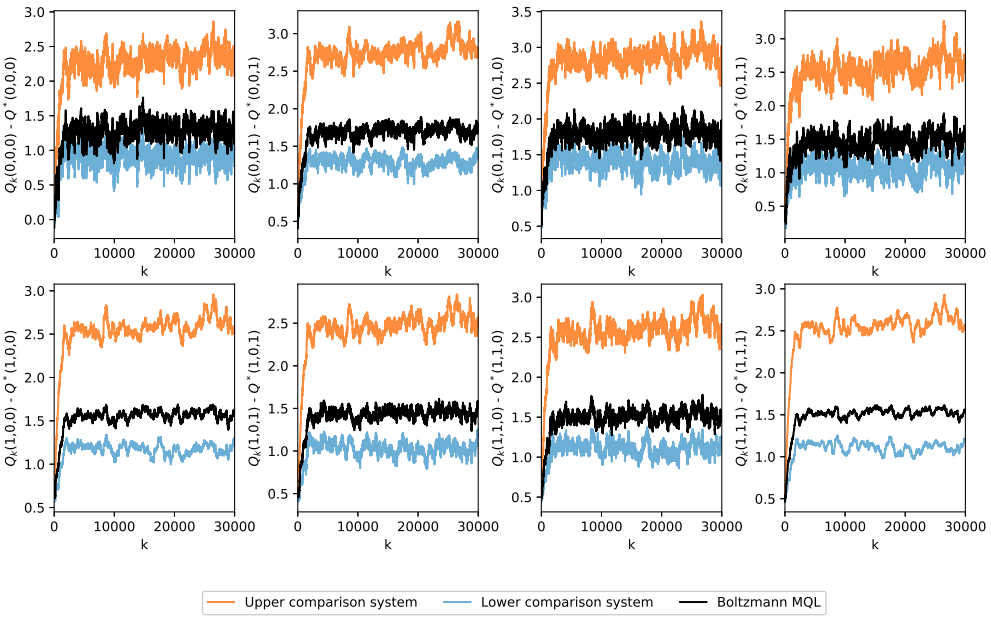

Figure 2: Trajectories of the Boltzmann MQL and its comparison systems

zero-sum Markov games was provided by Perolat et al. (2015), theoretical analyses of several generalized non-stationary RL algorithms were conducted by Pérolat et al. (2016), and the online reinforcement learning method in average-reward two-player Markov games was investigated by Wei et al. (2017). Furthermore, actor-critic algorithms designed for multi-agent Markov games were examined by Srinivasan et al. (2018) and Perolat et al. (2018), a thorough overview of the multi-agent Markov game and multi-agent RL was offered by Zhang et al. (2021), and the last-iterate convergence rate for two-player zero-sum Markov games was proposed by Cai et al. (2023) with an uncoupled, convergent, and rational algorithm.

Regarding the MQL convergence, recent research such as Fan et al. (2019) improved MQL by integrating deep Q-learning techniques and determined a finite-time error bound under mild assumptions. Zhu & Zhao (2020) also used deep Q-learning in MQL and showed asymptotic convergence in tabular settings. Diddigi et al. (2022) introduced a new generalized MQL and demonstrated its asymptotic convergence using stochastic approximation approaches. Lee (2023) developed a finite-time analysis of MQL and its value iteration by employing the switching system model.

While these prior works have made significant achievements over the years, it is important to note that most of the current literature concentrated on the asymptotic convergence (Littman, 2001; Zhu & Zhao, 2020) or the convergence of improved algorithms (Diddigi et al., 2022; Fan et al., 2019). To the best of the authors' knowledge, the finite-time convergence analysis has not yet been performed on the vanilla MQL and its smooth variant at the same time. Moreover, compared to Lee (2023), our approach can provide more intuitive and easier proofs by utilizing the switching system. Therefore, this paper can offer finite-time analysis of the MQL and its smooth variant by utilizing the switching system and the control-theoretic ideas in a more comprehensible and direct manner.

7 Conclusion

In this paper, the finite-time analysis of the MQL and its associated smooth variant for two-player zero-sum Markov games has been studied. The switching system models are used for both MQL and its smooth variant in order to conduct the analysis. By establishing upper and lower comparison systems, the finite-time analysis results for two MQL algorithms can be obtained. This method can offer more comprehensive insights into MQL and a more straightforward convergence analysis approach. Furthermore, this new perspective will reveal new relationships and encourage cooperation between ideas in the domains of control theory and reinforcement learning. For further steps in the future, it will be beneficial to get the stricter error bounds for both algorithms.

Acknowledgments

The work was supported by the Institute of Information Communications Technology Planning Evaluation (IITP) funded by the Korea government under Grant 2022-0-00469.

References

Mohammad Gheshlaghi Azar, Remi Munos, Mohammad Ghavamzadeh, and Hilbert J Kappen. Speedy Q-learning. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pp. 2411–2419, 2011.

Carolyn L Beck and Rayadurgam Srikant. Error bounds for constant step-size Q-learning. *Systems & Control letters*, 61(12):1203–1208, 2012.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pp. 449–458, 2017.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pp. 1691–1692, 2018.

Vivek S Borkar and Sean P Meyn. The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

Michael Bowling. Convergence problems of general-sum multiagent reinforcement learning. In *ICML*, pp. 89–94, 2000.

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. Uncoupled and convergent learning in two-player zero-sum markov games. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants. *arXiv preprint arXiv:2102.01567*, 2021.

Adithya M Devraj and Sean P Meyn. Zap Q-learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 2232–2241, 2017.

Raghuram Bharadwaj Diddigi, Chandramouli Kamanchi, and Shalabh Bhatnagar. A generalized minimax Q-learning algorithm for two-player zero-sum stochastic games. *IEEE Transactions on Automatic Control*, 67(9):4816–4823, 2022.

Eyal Even-Dar and Yishay Mansour. Learning rates for Q-learning. *Journal of machine learning Research*, 5(Dec):1–25, 2003.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep Q-learning. *arXiv preprint arXiv:1901.00137*, 2019.

Abhijit Gosavi. Boundedness of iterates in Q-learning. *Systems & Control letters*, 55(4):347–349, 2006.

H. V. Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *AAAI Conference on Artificial Intelligence*, 2015. URL https://api.semanticscholar.org/CorpusID:6208256.

Hado V Hasselt. Double Q-learning. In *Advances in Neural Information Processing Systems*, pp. 2613–2621, 2010.

Nicolas Heess, Jonathan J Hunt, Timothy P Lillicrap, and David Silver. Memory-based control with recurrent neural networks. *arXiv preprint arXiv:1512.04455*, 2015.

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

Junling Hu, Michael P Wellman, et al. Multiagent reinforcement learning: theoretical framework and an algorithm. In *ICML*, pp. 242–250, 1998.

Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in neural information processing systems*, pp. 703–710, 1994.

Narim Jeong and Donghwan Lee. Finite-time error analysis of soft q-learning: Switching system approach. https://arxiv.org/abs/2403.06366, 2024.

Michael J Kearns and Satinder P Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. In *Advances in neural information processing systems*, pp. 996–1002, 1999.

Hassan K Khalil. Nonlinear systems third edition (2002), 2002.

Harold Kushner and G. George Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Michail G Lagoudakis and Ronald Parr. Value function approximation in zero-sum markov games. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pp. 283–292, 2002.

Donghwan Lee. Finite-time analysis of minimax q-learning for two-player zero-sum markov games: Switching system approach. *arXiv preprint arXiv:2306.05700*, 2023.

Donghwan Lee. Final iteration convergence bound of q-learning: Switching system approach. *IEEE Transactions on Automatic Control*, 2024.

Donghwan Lee and Niao He. Periodic Q-learning. In *Learning for dynamics and control*, pp. 582–598, 2020a.

Donghwan Lee and Niao He. A unified switching system perspective and convergence analysis of Q-learning algorithms. In *34th Conference on Neural Information Processing Systems, NeurIPS 2020*, 2020b.

Donghwan Lee, Jianghai Hu, and Niao He. A discrete-time switching system analysis of Q-learning. *SIAM Journal on Control and Optimization (accepted)*, 2022.

Donghwan Lee, Jianghai Hu, and Niao He. A discrete-time switching system analysis of Q-learning. *SIAM Journal on Control and Optimization*, 61(3):1861–1880, 2023.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous Q-learning: sharper analysis and variance reduction. *arXiv preprint arXiv:2006.03041*, 2020.

Daniel Liberzon. *Switching in systems and control*. Springer Science & Business Media, 2003.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on learning representations*, 2016.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning*, ICML'94, pp. 157–163, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN 1558603352.

Michael L Littman. Value-function reinforcement learning in Markov games. *Cognitive systems research*, 2(1):55–66, 2001.

Michael L Littman and Csaba Szepesvári. A generalized reinforcement-learning model: convergence and applications. In *ICML*, volume 96, pp. 310–318, 1996.

Michael L Littman et al. Friend-or-foe Q-learning in general-sum games. In *ICML*, volume 1, pp. 322–328, 2001.

Francisco S Melo, Sean P Meyn, and M Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning*, pp. 664–671, 2008.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Ling Pan, Qingpeng Cai, Qi Meng, Wei Chen, Longbo Huang, and Tie-Yan Liu. Reinforcement learning with dynamic boltzmann softmax updates. *arXiv preprint arXiv:1903.05926*, 2019.

Julien Perolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. Approximate dynamic programming for two-player zero-sum Markov games. In *International Conference on Machine Learning*, pp. 1321–1329, 2015.

Julien Pérolat, Bilal Piot, Bruno Scherrer, and Olivier Pietquin. On the use of non-stationary strategies for solving two-player zero-sum markov games. In *Artificial Intelligence and Statistics*, pp. 893–901, 2016.

Julien Perolat, Bilal Piot, and Olivier Pietquin. Actor-critic fictitious play in simultaneous move multistage games. In *International Conference on Artificial Intelligence and Statistics*, pp. 919–928, 2018.

Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and Q-learning. *arXiv preprint arXiv:2002.00260*, 2020.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Zhao Song, Ron Parr, and Lawrence Carin. Revisiting the softmax bellman operator: New benefits and new perspective. In *International conference on machine learning*, pp. 5916–5925. PMLR, 2019.

Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation andtd learning. In *Conference on Learning Theory*, pp. 2803–2830, 2019.

Sriram Srinivasan, Marc Lanctot, Vinicius Zambaldi, Julien Pérolat, Karl Tuyls, Rémi Munos, and Michael Bowling. Actor-critic policy optimization in partially observable multiagent environments. *Advances in neural information processing systems*, 2018.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT Press, 1998.

Csaba Szepesvári. The asymptotic convergence-rate of Q-learning. In *Advances in Neural Information Processing Systems*, pp. 1064–1070, 1998.

John N Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine learning*, 16 (3):185–202, 1994.

Martin J Wainwright. Stochastic approximation with cone-contractive operators: sharp ℓ_∞-bounds for Q-learning. *arXiv preprint arXiv:1905.06265*, 2019.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pp. 1995–2003, 2016.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

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

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: a selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pp. 321–384, 2021.

Yuanheng Zhu and Dongbin Zhao. Online minimax Q network learning for two-player zero-sum markov games. *IEEE Transactions on Neural Networks and Learning Systems*, 33(3):1228–1241, 2020.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Additional discussions

Contributions. It is noteworthy to highlight that while the switching system model introduced in Lee et al. (2022) has been used as a basis, the main analysis and proof in this work significantly differed from those in Lee et al. (2022). It is also important to note that this paper only deals with the independently and identically distributed observations and the constant step-size in the tabular domain to make the analysis easier. Nevertheless, our analysis can be expanded to include more complex Markovian observation models by employing the methods described in Srikant & Ying (2019) and Bhandari et al. (2018).

Two-player Markov games. There are two types of two-player Markov games (Shapley, 1953): alternating two-player Markov games and simultaneous two-player Markov games. Since the alternating two-player Markov games streamline the basic ideas and related procedures, this study mainly focuses on it. Note that all of the results in this paper can also be applied to the simultaneous two-player Markov games without difficulty.

Minimax Q-learning and its smooth variant. Algorithm 1 shows the MQL algorithm that is used in this study, which differs somewhat from the original MQL in Littman (1994). The MQL in Algorithm 1 uses the max operator that choose an action a in the restricted set $\mathcal{A}$, not in the set of all stochastic policies. Nevertheless, it is noteworthy that all of the analyses in this paper are still relevant to the original MQL.

Algorithm 2 describes the Boltzmann MQL algorithm, with the definition of $h^{\omega}_{a\in\mathcal{A}}$ and $h^{-\omega}_{b\in\mathcal{B}}$ in (3).

Algorithm 1 MQL

1: Initialize $Q_0 \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}$ randomly such that $\|Q_0\|_\infty \leq 1$.
2: **for** iteration $k = 0, 1, \ldots$ **do**
3: Sample $a_k \sim \beta(\cdot|s_k)$. $b_k \sim \phi(\cdot|s_k)$ and $s_k \sim p(\cdot)$
4: Sample $s'_k \sim P(\cdot|s_k, a_k, b_k)$ and $r_k = r(s_k, a_k, b_k, s'_k)$
5: Update

$$\begin{aligned} Q_{k+1}(s_k, a_k, b_k) =& Q_k(s_k, a_k, b_k) \\ &+ \alpha\left\{r_k + \gamma \max_{a\in\mathcal{A}} \min_{b\in\mathcal{B}} Q_k(s'_k, a, b) - Q_k(s_k, a_k, b_k)\right\} \end{aligned}$$

6: **end for**

Algorithm 2 Boltzmann MQL

1: Initialize $Q^{bz}_0 \in \mathbb{R}^{|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}$ randomly such that $\left\|Q^{bz}_0\right\|_\infty \leq 1$.
2: **for** iteration $k = 0, 1, \ldots$ **do**
3: Sample $a_k \sim \beta(\cdot|s_k)$. $b_k \sim \phi(\cdot|s_k)$ and $s_k \sim p(\cdot)$
4: Sample $s'_k \sim P(\cdot|s_k, a_k, b_k)$ and $r_k = r(s_k, a_k, b_k, s'_k)$
5: Update

$$\begin{aligned} Q^{bz}_{k+1}(s_k, a_k, b_k) =& Q^{bz}_k(s_k, a_k, b_k) \\ &+ \alpha\Big\{r_k + \gamma h^{\omega}_{a\in\mathcal{A}}\left(h^{-\omega}_{b\in\mathcal{B}}\left(Q^{bz}_k(s'_k, a, b)\right)\right) - Q^{bz}_k(s_k, a_k, b_k)\Big\} \end{aligned}$$

6: **end for**

B Additional assumptions

Assumption 2. *The following general assumptions are used in this paper:*

(i) Given the behavior policies β and ϕ that agents actually use to acquire experiences, $\{(s_k, a_k, b_k, s'_k)\}_{k=0}^{\infty}$ are independent and identically distributed samples.

(ii) Assuming that the state is generated from the stationary state distribution p at each iteration, the state-action distribution can be derived as

$$d(s, a, b) = p(s)\beta(a|s)\phi(b|s)$$

with $(s, a, b) \in \mathcal{S} \times \mathcal{A} \times \mathcal{B}$.

C Proof of Proposition 1 and Proposition 2

Propositions 1 and 2 can be proven by using an induction argument.

Proof of Proposition 1. Suppose the result is valid for some $k \geq 0$. Then,

$$\begin{aligned}
Q_{k+1}(s_k, a_k, b_k) =& (1-\alpha)Q_k(s_k, a_k, b_k) + \alpha\left\{r_k + \gamma \max_{a\in\mathcal{A}} \min_{b\in\mathcal{B}} Q_k\left(s'_k, a, b\right)\right\} \\
\leq& (1-\alpha)Q_k(s_k, a_k, b_k) + \alpha\left\{r_k + \gamma \max_{a\in\mathcal{A}} Q_k\left(s'_k, a, \mu^*\left(s'_k, a\right)\right)\right\} \\
\leq& (1-\alpha)Q_k^U(s_k, a_k, b_k) + \alpha\left\{r_k + \gamma \max_{a\in\mathcal{A}} Q_k^U\left(s'_k, a, \mu^*\left(s'_k, a\right)\right)\right\} \\
=& Q_{k+1}^U(s_k, a_k, b_k),
\end{aligned}$$

where the second inequality is based on the assumption $Q_k^U(s, a, b) \geq Q_k(s, a, b)$. By induction, the proof is completed. □

Proof of Proposition 2. Suppose the result is valid for some $k \geq 0$. Then,

$$\begin{aligned}
Q_{k+1}(s_k, a_k, b_k) =& (1-\alpha)Q_k(s_k, a_k, b_k) + \alpha\left\{r_k + \gamma \max_{a\in\mathcal{A}} \min_{b\in\mathcal{B}} Q_k\left(s'_k, a, b\right)\right\} \\
\geq& (1-\alpha)Q_k(s_k, a_k, b_k) + \alpha\left\{r_k + \gamma \min_{b\in\mathcal{B}} Q_k\left(s'_k, \pi^*\left(s'_k\right), b\right)\right\} \\
\geq& (1-\alpha)Q_k^L(s_k, a_k, b_k) + \alpha\left\{r_k + \gamma \min_{b\in\mathcal{B}} Q_k^L\left(s'_k, \pi^*\left(s'_k\right), b\right)\right\} \\
=& Q_{k+1}^L(s_k, a_k, b_k),
\end{aligned}$$

where the second inequality is based on the assumption $Q_k^L(s, a, b) \leq Q_k(s, a, b)$. By induction, the proof is completed. □

D Proof of Theorem 1

Proof. For every $k \geq 0$,

$$\begin{aligned}
\mathbb{E}\left[\left\|Q_k^U - Q^*\right\|_\infty\right] \leq& \frac{9\alpha^{\frac{1}{2}} d_{\max}|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}{d_{\min}^{\frac{3}{2}}(1-\gamma)^{\frac{5}{2}}} + \frac{2|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{3}{2}}}{1-\gamma}\rho^k \\
&+ \frac{4\alpha\gamma d_{\max}|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{2}{3}}}{1-\gamma} k\rho^{k-1}
\end{aligned} \tag{12}$$

can be derived from Lemma 1 with the Q-function vector in Definition 1. Similarly, one can also provide a finite-time error bound for the lower iteration of MQL because it is symmetric with respect to the upper iteration. The lower iteration of MQL shares the same bound with the right side of (12).

Then, by using the relation

$$\begin{aligned}
\mathbb{E}\left[\|Q_k - Q^*\|_2\right] =& \mathbb{E}\left[\left\|Q_k - Q_k^L + Q_k^L - Q^*\right\|_2\right] \\
\leq& \mathbb{E}\left[\left\|Q_k^L - Q^*\right\|_2\right] + \mathbb{E}\left[\left\|Q_k - Q_k^L\right\|_2\right] \\
\leq& \mathbb{E}\left[\left\|Q_k^L - Q^*\right\|_2\right] + \mathbb{E}\left[\left\|Q_k^U - Q_k^L\right\|_2\right] \\
\leq& \mathbb{E}\left[\left\|Q_k^L - Q^*\right\|_2\right] + \mathbb{E}\left[\left\|Q_k^U - Q^* + Q^* - Q_k^L\right\|_2\right] \\
\leq& \mathbb{E}\left[\left\|Q_k^L - Q^*\right\|_2\right] + \mathbb{E}\left[\left\|Q_k^U - Q^*\right\|_2\right] + \mathbb{E}\left[\left\|Q^* - Q_k^L\right\|_2\right] \\
=& 2\mathbb{E}\left[\left\|Q_k^L - Q^*\right\|_2\right] + \mathbb{E}\left[\left\|Q_k^U - Q^*\right\|_2\right],
\end{aligned}$$

where the triangle inequality accounts for the first and fourth inequalities and the fact that $Q_k^U - Q_k^L \geq Q_k - Q_k^L \geq 0$ provides the second inequality, the final conclusion can be obtained by combining (12) and the error bound for the lower iteration of MQL. □

E Proof of Proposition 3 and Proposition 4

To get the upper and lower comparison system of the Boltzmann MQL, the following proposition is utilized:

Proposition 5 (Jeong & Lee (2024)). *For any $v \in \mathbb{R}^{|\mathcal{U}|}$,*

$$\max_{u\in\mathcal{U}} v(u) - \frac{\ln(|\mathcal{U}|)}{\omega} \leq h^{\omega}_{u\in\mathcal{U}}(v) \leq \max_{u\in\mathcal{U}} v(u) \quad \textit{and} \quad \min_{u\in\mathcal{U}} v(u) \leq h^{-\omega}_{u\in\mathcal{U}}(v) \leq \min_{u\in\mathcal{U}} v(u) + \frac{\ln(|\mathcal{U}|)}{\omega},$$

where "$\leq$" represents the element-wise inequality.

Then, Propositions 3 and 4 can be proven by using an induction argument.

Proof of Proposition 3. Suppose the result is valid for some $k \geq 0$. Then,

$$\begin{aligned}
& Q_{k+1}^{bz}(s_k, a_k, b_k) \\
=& (1-\alpha) Q_k^{bz}(s_k, a_k, b_k) + \alpha\left\{r_k + \gamma h^{\omega}_{a\in\mathcal{A}}\left(h^{-\omega}_{b\in\mathcal{B}}\left(Q_k^{bz}(s_k', a, b)\right)\right)\right\} \\
\leq& (1-\alpha) Q_k^{bz}(s_k, a_k, b_k) + \alpha\left\{r_k + \gamma \max_{a\in\mathcal{A}} h^{-\omega}_{b\in\mathcal{B}}\left(Q_k^{bz}(s_k', a, b)\right)\right\} \\
\leq& (1-\alpha) Q_k^{bz}(s_k, a_k, b_k) + \alpha\left\{r_k + \gamma \max_{a\in\mathcal{A}} \min_{b\in\mathcal{B}} Q_k^{bz}(s_k', a, b) + \gamma\frac{\ln(|\mathcal{B}|)}{\omega}\right\} \\
\leq& (1-\alpha) Q_k^{bz}(s_k, a_k, b_k) + \alpha\left\{r_k + \gamma \max_{a\in\mathcal{A}} Q_k^{bz}\left(s_k', a, \mu^*(s_k', a)\right) + \gamma\frac{\ln(|\mathcal{B}|)}{\omega}\right\} \\
\leq& (1-\alpha) Q_k^{U,bz}(s_k, a_k, b_k) + \alpha\left\{r_k + \gamma \max_{a\in\mathcal{A}} Q_k^{U,bz}\left(s_k', a, \mu^*(s_k', a)\right) + \gamma\frac{\ln(|\mathcal{B}|)}{\omega}\right\} \\
=& Q_{k+1}^{U,bz}(s_k, a_k, b_k),
\end{aligned}$$

where the first and second inequalities utilize Proposition 5, and the last inequality is based on the assumption $Q_k^{U,bz}(s, a, b) \geq Q_k^{bz}(s, a, b)$. By induction, the proof is completed. □

Proof of Proposition 4. Suppose the result is valid for some $k \geq 0$. Then,

$$\begin{aligned}
Q_{k+1}^{bz}(s_k, a_k, b_k) =& (1-\alpha) Q_k^{bz}(s_k, a_k, b_k) + \alpha\left\{r_k + \gamma h^{\omega}_{a\in\mathcal{A}}\left(h^{-\omega}_{b\in\mathcal{B}}\left(Q_k^{bz}(s_k', a, b)\right)\right)\right\} \\
\geq& (1-\alpha) Q_k^{bz}(s_k, a_k, b_k) + \alpha\left\{r_k + \gamma h^{\omega}_{a\in\mathcal{A}}\left(\min_{b\in\mathcal{B}} Q_k^{bz}(s_k', a, b)\right)\right\}
\end{aligned}$$

$$
\begin{aligned}
&\geq (1-\alpha) Q_k^{bz}(s_k, a_k, b_k) + \alpha \left\{ r_k + \gamma \max_{a \in \mathcal{A}} \min_{b \in \mathcal{B}} Q_k^{bz}(s_k', a, b) - \gamma \frac{\ln(|\mathcal{A}|)}{\omega} \right\} \\
&\geq (1-\alpha) Q_k^{bz}(s_k, a_k, b_k) + \alpha \left\{ r_k + \gamma \min_{b \in \mathcal{B}} Q_k^{bz}\left(s_k', \pi^*(s_k'), b\right) - \gamma \frac{\ln(|\mathcal{A}|)}{\omega} \right\} \\
&\geq (1-\alpha) Q_k^{L,bz}(s_k, a_k, b_k) + \alpha \left\{ r_k + \gamma \min_{b \in \mathcal{B}} Q_k^{L,bz}\left(s_k', \pi^*(s_k'), b\right) - \gamma \frac{\ln(|\mathcal{A}|)}{\omega} \right\} \\
&= Q_{k+1}^{L,bz}(s_k, a_k, b_k),
\end{aligned}
$$

where the first and second inequalities utilize Proposition 5, and the last inequality is based on the assumption $Q_k^{L,bz}(s,a,b) \leq Q_k^{bz}(s,a,b)$. By induction, the proof is completed. □

F Construction of the upper and lower comparison systems (7) and (10)

For the upper comparison system of Boltzmann MQL (7), we first modify the upper iteration of Boltzmann MQL (5) in a vector form as

$$
\begin{aligned}
Q_{k+1}^{U,bz} =& Q_k^{U,bz} + \alpha \Big\{ (e_{a_k} \otimes e_{b_k} \otimes e_{s_k}) r_k + \gamma (e_{a_k} \otimes e_{b_k} \otimes e_{s_k}) \left(e_{s_k'}\right)^T \Pi_{Q_k^{U,bz}} Q_k^{U,bz} \\
&- (e_{a_k} \otimes e_{b_k} \otimes e_{s_k})(e_{a_k} \otimes e_{b_k} \otimes e_{s_k})^T Q_k^{U,bz} + \gamma (e_{a_k} \otimes e_{b_k} \otimes e_{s_k}) \frac{\ln(|\mathcal{B}|)}{\omega} \Big\}.
\end{aligned} \tag{13}
$$

Then, taking the conditional expectation of (13) based on $Q_k^{U,bz}$ yields

$$
\mathbb{E}\left[Q_{k+1}^{U,bz} \middle| Q_k^{U,bz} \right] = Q_k^{U,bz} + \alpha \left(DR + \gamma DP \Pi_{Q_k^{U,bz}} Q_k^{U,bz} - D Q_k^{U,bz} + \gamma D \frac{\ln(|\mathcal{B}|)}{\omega} \mathbf{1} \right), \tag{14}
$$

where

$$
D = \mathbb{E}\left[(e_{a_k} \otimes e_{b_k} \otimes e_{s_k})(e_{a_k} \otimes e_{b_k} \otimes e_{s_k})^T \mid Q_k^{U,bz} \right],
$$

$$
DP = \mathbb{E}\left[(e_{a_k} \otimes e_{b_k} \otimes e_{s_k}) \left(e_{s_k'}\right)^T \mid Q_k^{U,bz} \right],
$$

and $\mathbf{1}$ is a column vector with all values of 1.

Through the calculation of (13) - (14) + (14), one gets

$$
Q_{k+1}^{U,bz} = Q_k^{U,bz} + \alpha \left(DR + \gamma DP \Pi_{Q_k^{U,bz}} Q_k^{U,bz} - D Q_k^{U,bz} \right) + \alpha w_k^U + \alpha \gamma D \frac{\ln(|\mathcal{B}|)}{\omega} \mathbf{1} \tag{15}
$$

with

$$
\begin{aligned}
w_k^U =& (e_{a_k} \otimes e_{b_k} \otimes e_{s_k}) r_k + \gamma (e_{a_k} \otimes e_{b_k} \otimes e_{s_k}) \left(e_{s_k'}\right)^T \Pi_{Q_k^{U,bz}} Q_k^{U,bz} \\
&- (e_{a_k} \otimes e_{b_k} \otimes e_{s_k})(e_{a_k} \otimes e_{b_k} \otimes e_{s_k})^T Q_k^{U,bz} + \gamma (e_{a_k} \otimes e_{b_k} \otimes e_{s_k}) \frac{\ln(|\mathcal{B}|)}{\omega} \\
&- \left(DR + \gamma DP \Pi_{Q_k^{U,bz}} Q_k^{U,bz} - D Q_k^{U,bz} + \gamma D \frac{\ln(|\mathcal{B}|)}{\omega} \mathbf{1} \right).
\end{aligned}
$$

By reformulating (15), the result of (7) can be obtained as follows:

$$
\begin{aligned}
Q_{k+1}^{U,bz} - Q^* =& Q_k^{U,bz} - Q^* + \alpha \left(DR + \gamma DP \Pi_{Q_k^{U,bz}} Q_k^{U,bz} - D Q_k^{U,bz} \right) + \alpha w_k^U + \alpha \gamma D \frac{\ln(|\mathcal{B}|)}{\omega} \mathbf{1} \\
=& Q_k^{U,bz} - Q^* + \alpha \left(-\gamma DP \Pi_{Q^*} Q^* + D Q^* + \gamma DP \Pi_{Q_k^{U,bz}} Q_k^{U,bz} - D Q_k^{U,bz} \right) \\
&+ \alpha w_k^U + \alpha \gamma D \frac{\ln(|\mathcal{B}|)}{\omega} \mathbf{1} + \alpha \gamma DP \left(\Pi_{Q_k^{U,bz}} Q^* - \Pi_{Q_k^{U,bz}} Q^* \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \left\{ I + \alpha \left(\gamma D P \Pi_{Q_k^{U,bz}} - D \right) \right\} \left(Q_k^{U,bz} - Q^* \right) + \alpha \gamma D P \left(\Pi_{Q_k^{U,bz}} - \Pi_{Q^*} \right) Q^* \\
&\quad + \alpha w_k^U + \alpha \gamma D \frac{\ln(|\mathcal{B}|)}{\omega} \mathbf{1} \\
&= A_{Q_k^{U,bz}} \left(Q_k^{U,bz} - Q^* \right) + b_{Q_k^{U,bz}} + \alpha w_k^U + \alpha \gamma D \frac{\ln(|\mathcal{B}|)}{\omega} \mathbf{1}
\end{aligned}
$$

with

$$
A_Q := I + \alpha \left(\gamma D P \Pi_Q - D \right) \quad \text{and} \quad b_Q := \alpha \gamma D P \left(\Pi_Q - \Pi_{Q^*} \right) Q^*,
$$

where the second equality uses the optimal Bellman equation $Q^*(s,a,b) = R(s,a,b) + \gamma \sum_{s' \in S} P(s'|s,a,b) \max_{a' \in \mathcal{A}} \min_{b' \in \mathcal{B}} Q^*(s',a',b')$ and the fact that $\max_{a \in \mathcal{A}} \min_{b \in \mathcal{B}} Q^*(s,a,b) = \max_{a \in \mathcal{A}} Q^*(s,a,\mu^*(s,a))$.

Using a similar technique, we can also get (10).

G Finite-time error bound for the upper and lower comparison systems

In order to determine the convergence of Boltzmann MQL, we once again transfer this problem into the stability analysis of the affine switching system and adjust the upper and lower comparison systems. However, since (7) and (10) contain extra affine vectors and stochastic noises compared to the MQL scenario, it is necessary to apply further comparison systems. The upper comparison system is supplemented by the additional comparison systems referred to as the *upper-upper comparison system* and the *upper-lower comparison system*. By demonstrating the convergence of the upper-upper and upper-lower comparison systems, the convergence of the upper comparison system can be proved. After doing the same procedure for the lower comparison system, the finite-time error bound of the Boltzmann MQL can eventually be found.

G.1 Finite-time error bound for the upper comparison system

In this subsection, the new comparison systems that effectively bound the upper comparison system (7) are proposed. The *upper-lower comparison system*, which gives the upper comparison system's lower bound, is taken into consideration as

$$
Q_{k+1}^{UL,bz} - Q^* = A_{Q^*} \left(Q_k^{UL,bz} - Q^* \right) + \alpha w_k^U + \alpha \gamma D \frac{\ln(|\mathcal{B}|)}{\omega} \mathbf{1}. \tag{16}
$$

Moreover, the *upper-upper comparison system*, which gives the upper comparison system's upper bound, is taken into consideration as

$$
Q_{k+1}^{UU,bz} - Q^* = A_{Q_k^{U,bz}} \left(Q_k^{UU,bz} - Q^* \right) + \alpha w_k^U + \alpha \gamma D \frac{\ln(|\mathcal{B}|)}{\omega} \mathbf{1}. \tag{17}
$$

These two additional comparison systems (16) and (17) can be proved by the following propositions:

Proposition 6. *Assume that* $Q_0^{UL,bz}(s,a,b) \leq Q_0^{U,bz}(s,a,b)$ *for every* (s,a,b). *Then,* $Q_k^{UL,bz}(s,a,b) \leq Q_k^{U,bz}(s,a,b)$ *for every* (s,a,b) *and* $k \geq 0$.

Proposition 7. *Assume that* $Q_0^{UU,bz}(s,a,b) \geq Q_0^{U,bz}(s,a,b)$ *for every* (s,a,b). *Then,* $Q_k^{UU,bz}(s,a,b) \geq Q_k^{U,bz}(s,a,b)$ *for every* (s,a,b) *and* $k \geq 0$.

The entire proofs are presented in Appendix H. Then, the convergence of the upper comparison system can be established by proving the convergence of the upper-lower and upper-upper comparison systems.

First, the convergence of the upper-lower comparison system can be shown by the following theorem, with the entire proof available in Appendix I:

Theorem 5. *For every $k \geq 0$,*

$$\mathbb{E}\left[\left\|Q_k^{UL,bz} - Q^*\right\|_2\right] \leq |\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}} \left\|Q_0^{UL,bz} - Q^*\right\|_2 \rho^k + \frac{2\sqrt{2}\alpha^{\frac{1}{2}}(\ln(|\mathcal{B}|)+\omega)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^{\frac{1}{2}}(1-\gamma)^{\frac{3}{2}}} + \frac{\gamma d_{max}\ln(|\mathcal{B}|)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}(1-\gamma)}.$$

Once the convergence of the upper-lower comparison system (16) has been established, it is also possible to confirm the convergence of the upper-upper comparison system (17). However, the dependency of the subsystem matrix on the Q-function makes it difficult to directly show the convergence of (17). In other words, in contrast to (16), the subsystem matrix $A_{Q_k^{U,bz}}$ and the state $Q_k^{UU,bz} - Q^*$ in (17) cannot be isolated if the expectation of (17) is considered. To avoid this problem, an *error system* is alternatively investigated, which can be made by subtracting the upper-lower system (16) from the upper-upper comparison system (17):

$$Q_{k+1}^{UU,bz} - Q_{k+1}^{UL,bz} = A_{Q_k^{U,bz}}\left(Q_k^{UU,bz} - Q_k^{UL,bz}\right) + B_{Q_k^{U,bz}}\left(Q_k^{UL,bz} - Q^*\right) \tag{18}$$

with

$$B_Q := A_Q - A_{Q^*} = \alpha\gamma DP\left(\Pi_Q - \Pi_{Q^*}\right). \tag{19}$$

In the error system (18), the stochastic noise (9) is eliminated, which relieves the statistical dependency problem. Moreover, $Q_k^{UL,bz} - Q^*$ can be interpreted as an external disturbance in (18).

Then, the error bound of the error system (18) can be verified by the following theorem:

Theorem 6. *For every $k \geq 0$,*

$$\begin{aligned}\mathbb{E}\left[\left\|Q_k^{UU,bz} - Q_k^{UL,bz}\right\|_\infty\right] \leq& \frac{4\alpha\gamma d_{max}|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}{1-\gamma}k\rho^{k-1} \\ &+ \frac{4\sqrt{2}\alpha^{\frac{1}{2}}\gamma d_{max}(\ln(|\mathcal{B}|)+\omega)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^{\frac{3}{2}}(1-\gamma)^{\frac{5}{2}}} \\ &+ \frac{2\gamma^2 d_{max}^2\ln(|\mathcal{B}|)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^2(1-\gamma)^2}.\end{aligned} \tag{20}$$

The complete proof is provided in Appendix L.

The main idea used to illustrate the convergence of the upper comparison system (7) is as follows:

$$\begin{aligned}\mathbb{E}\left[\left\|Q_k^{U,bz} - Q^*\right\|_\infty\right] =& \mathbb{E}\left[\left\|Q_k^{U,bz} - Q_k^{UL,bz} + Q_k^{UL,bz} - Q^*\right\|_\infty\right] \\ \leq& \mathbb{E}\left[\left\|Q_k^{U,bz} - Q_k^{UL,bz}\right\|_\infty\right] + \mathbb{E}\left[\left\|Q_k^{UL,bz} - Q^*\right\|_\infty\right] \\ \leq& \mathbb{E}\left[\left\|Q_k^{UU,bz} - Q_k^{UL,bz}\right\|_\infty\right] + \mathbb{E}\left[\left\|Q_k^{UL,bz} - Q^*\right\|_\infty\right].\end{aligned} \tag{21}$$

As $Q_k^{UL,bz} - Q^* \to 0$ and $Q_k^{UU,bz} - Q_k^{UL,bz} \to 0$ with $k \to \infty$ are shown by Theorems 5 and 6, it is also possible to have $Q_k^{U,bz} \to Q^*$, where the bound on the expected error $\mathbb{E}\left[\left\|Q_k^{U,bz} - Q^*\right\|_\infty\right]$ can be established by the following theorem:

Theorem 2 Restated. For every $k \geq 0$,

$$\mathbb{E}\left[\left\|Q_k^{U,bz} - Q^*\right\|_\infty\right] \leq \frac{4\alpha\gamma d_{max}|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}{1-\gamma}k\rho^{k-1} + \frac{6\sqrt{2}\alpha^{\frac{1}{2}}d_{max}(\ln(|\mathcal{B}|)+\omega)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^{\frac{3}{2}}(1-\gamma)^{\frac{5}{2}}}$$

$$+ \frac{3\gamma d_{max}^2 \ln(|\mathcal{B}|)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^2(1-\gamma)^2} + \frac{2|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}{1-\gamma}\rho^k.$$

Proof. The result can be achieved by combining (21), Theorems 5 and 6, followed by $\gamma \in [0,1)$, Assumption 1, Definition 1, and the property of the optimal Q-function. □

G.2 Finite-time error bound for the lower comparison system

Because of the symmetrical nature of the Boltzmann MQL's lower comparison system (10), its convergence can be easily evaluated using the findings of the previous subsection.

Similar to Appendix G.1, The *lower-lower comparison system* can be represented as

$$Q_{k+1}^{LL,bz} - Q^* = A'_{Q_k^{L,bz}}\left(Q_k^{LL,bz} - Q^*\right) + \alpha w_k^L - \alpha\gamma D\frac{\ln(|\mathcal{A}|)}{\omega}\mathbf{1}, \tag{22}$$

and the *lower-upper comparison system* can be represented as

$$Q_{k+1}^{LU,bz} - Q^* = A'_{Q^*}\left(Q_k^{LU,bz} - Q^*\right) + \alpha w_k^L - \alpha\gamma D\frac{\ln(|\mathcal{A}|)}{\omega}\mathbf{1}, \tag{23}$$

which can be proved by the following propositions:

Proposition 8. *Assume that* $Q_0^{LL,bz}(s,a,b) \leq Q_0^{L,bz}(s,a,b)$ *for every* (s,a,b). *Then,* $Q_k^{LL,bz}(s,a,b) \leq Q_k^{L,bz}(s,a,b)$ *for every* (s,a,b) *and* $k \geq 0$.

Proposition 9. *Assume that* $Q_0^{LU,bz}(s,a,b) \geq Q_0^{L,bz}(s,a,b)$ *for every* (s,a,b). *Then,* $Q_k^{LU,bz}(s,a,b) \geq Q_k^{L,bz}(s,a,b)$ *for every* (s,a,b) *and* $k \geq 0$.

The proofs of Propositions 8 and 9 are provided in Appendix M.

As can be observed, the lower-lower comparison system (22) is similar to the upper-upper comparison system (17), and the lower-upper comparison system (23) is similar to the upper-lower comparison system (16). Therefore, the prior results can be used to demonstrate the convergence of each comparison system (22) and (23). More precisely, the findings of Appendix G.1 can be applied to demonstrate the convergence of the lower-upper comparison system (23) and the error system.

Corollary 1. *For every* $k \geq 0$,

$$\mathbb{E}\left[\left\|Q_k^{LU,bz} - Q^*\right\|_2\right] \leq |\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}\left\|Q_0^{LU,bz} - Q^*\right\|_2\rho^k + \frac{2\sqrt{2}\alpha^{\frac{1}{2}}(\ln(|\mathcal{A}|)+\omega)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^{\frac{1}{2}}(1-\gamma)^{\frac{3}{2}}} + \frac{\gamma d_{max}\ln(|\mathcal{A}|)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}(1-\gamma)}.$$

Corollary 1 illustrates the convergence result of the lower-upper comparison system (23), which can be achieved by using Theorem 5. In contrast to Theorem 5, the second and third terms on the right side of the inequality of Corollary 1 have $\ln(|\mathcal{A}|)$ rather than $\ln(|\mathcal{B}|)$. It is due to the difference between the additional terms $\alpha\gamma D\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}$ of (16) and $-\alpha\gamma D\frac{\ln(|\mathcal{A}|)}{\omega}\mathbf{1}$ of (23).

If the error system is constructed by subtracting the lower-upper comparison system (23) from the lower-lower comparison system (22), the convergence of the error system can be determined as follows:

Corollary 2. *For every* $k \geq 0$,

$$\mathbb{E}\left[\left\|Q_k^{LL,bz} - Q_k^{LU,bz}\right\|_\infty\right] \leq \frac{4\alpha\gamma d_{max}|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}{1-\gamma}k\rho^{k-1} + \frac{4\sqrt{2}\alpha^{\frac{1}{2}}\gamma d_{max}(\ln(|\mathcal{A}|)+\omega)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^{\frac{3}{2}}(1-\gamma)^{\frac{5}{2}}} + \frac{2\gamma^2 d_{max}^2\ln(|\mathcal{A}|)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^2(1-\gamma)^2}.$$

Corollary 2 illustrates the convergence result of the error system, which can be achieved by using Theorem 6. Similar to Corollary 1, the second and third terms on the right side of the inequality of Corollary 2 have $\ln(|\mathcal{A}|)$ instead of $\ln(|\mathcal{B}|)$ for the same reason.

Then, the finite-time error bound of the lower comparison system (10) can be proved by the following theorem:

Theorem 3 Restated. For every $k \geq 0$,

$$\mathbb{E}\left[\left\|Q_k^{L,bz} - Q^*\right\|_\infty\right] \leq \frac{4\alpha\gamma d_{max}|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}{1-\gamma}k\rho^{k-1} + \frac{6\sqrt{2}\alpha^{\frac{1}{2}}d_{max}(\ln(|\mathcal{A}|)+\omega)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^{\frac{3}{2}}(1-\gamma)^{\frac{5}{2}}} + \frac{3\gamma d_{max}^2\ln(|\mathcal{A}|)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^2(1-\gamma)^2} + \frac{2|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}{1-\gamma}\rho^k.$$

Proof. The result can be achieved by combining the relation

$$\begin{aligned}\mathbb{E}\left[\left\|Q_k^{L,bz} - Q^*\right\|_\infty\right] \leq& \mathbb{E}\left[\left\|Q_k^{LU,bz} - Q^*\right\|_\infty\right]\\ \leq& \mathbb{E}\left[\left\|Q_k^{LU,bz} - Q^*\right\|_\infty\right] + \mathbb{E}\left[\left\|Q_k^{LL,bz} - Q_k^{LU,bz}\right\|_\infty\right],\end{aligned}$$

Corollaries 1 and 2, followed by $\gamma \in [0,1)$, Assumption 1, Definition 1, and the property of the optimal Q-function. □

H Proof of Proposition 6 and Proposition 7

Propositions 6 and 7 can be proven by using an induction argument.

Proof of Proposition 6. Suppose the result is valid for some $k \geq 0$. Then,

$$\begin{aligned}&Q_{k+1}^{U,bz} - Q^*\\ =&A_{Q_k^{U,bz}}\left(Q_k^{U,bz} - Q^*\right) + b_{Q_k^{U,bz}} + \alpha w_k^U + \alpha\gamma D\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}\\ =&A_{Q^*}\left(Q_k^{U,bz} - Q^*\right) + \left(A_{Q_k^{U,bz}} - A_{Q^*}\right)\left(Q_k^{U,bz} - Q^*\right) + b_{Q_k^{U,bz}} + \alpha w_k^U + \alpha\gamma D\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}\\ =&A_{Q^*}\left(Q_k^{U,bz} - Q^*\right) + \alpha\gamma DP\left(\Pi_{Q_k^{U,bz}} - \Pi_{Q^*}\right)Q_k^{U,bz} + \alpha w_k^U + \alpha\gamma D\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}\\ \geq&A_{Q^*}\left(Q_k^{U,bz} - Q^*\right) + \alpha w_k^U + \alpha\gamma D\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}\\ \geq&A_{Q^*}\left(Q_k^{UL,bz} - Q^*\right) + \alpha w_k^U + \alpha\gamma D\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}\\ =&Q_{k+1}^{UL,bz} - Q^*,\end{aligned}$$

where the first inequality utilizes $\alpha\gamma DP\left(\Pi_{Q_k} - \Pi_{Q^*}\right)Q_k \geq \alpha\gamma DP\left(\Pi_{Q^*} - \Pi_{Q^*}\right)Q_k = 0$, and the last inequality is based on the assumption $Q_k^{UL,bz}(s,a,b) \leq Q_k^{U,bz}(s,a,b)$ and the fact that A_{Q^*} is a nonnegative matrix. By induction, the proof is completed. □

Proof of Proposition 7. Suppose the result is valid for some $k \geq 0$. Then,

$$\begin{aligned}Q_{k+1}^{U,bz} - Q^* =&A_{Q_k^{U,bz}}\left(Q_k^{U,bz} - Q^*\right) + b_{Q_k^{U,bz}} + \alpha w_k^U + \alpha\gamma DP\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}\\ \leq&A_{Q_k^{U,bz}}\left(Q_k^{U,bz} - Q^*\right) + \alpha w_k^U + \alpha\gamma DP\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}\end{aligned}$$

$$
\begin{aligned}
&\leq A_{Q_k^{U,bz}}\left(Q_k^{UU,bz}-Q^*\right)+\alpha w_k^U+\alpha\gamma DP\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}\\
&=Q_{k+1}^{UU,bz}-Q^*,
\end{aligned}
$$

where the first inequality utilizes the property of $b_{Q_k^{U,bz}} = \alpha\gamma DP\left(\Pi_{Q_k^{U,bz}}-\Pi_{Q^*}\right)Q^* \leq \alpha\gamma DP\left(\Pi_{Q^*}-\Pi_{Q^*}\right)Q^*=0$, and the last inequality is based on the assumption $Q_k^{UU,bz}(s,a,b)\geq Q_k^{U,bz}(s,a,b)$ and the fact that $A_{Q_k^{U,bz}}$ is a nonnegative matrix. By induction, the proof is completed. □

I Proof of Theorem 5

In order to demonstrate the convergence of the upper-lower comparison system (16), the following supplementary lemmas are required:

Lemma 2 (Lee et al. (2023))**.** *For any $v\in\mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$,*

$$\|A_v\|_\infty\leq\rho.$$

Here, A_v is the subsystem matrix in the form of (8), $\|A\|_\infty:=\max_{1\leq i\leq m}\sum_{j=1}^n|A_{ij}|$, A_{ij} is the element in the ith row and jth column of A, and ρ is specified in Definition 1. This can also be applied to every $Q\in\mathbb{R}^{|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}$ with A_Q.

Lemma 3. *For every $k\geq 0$,*

$$\mathbb{E}\left[\left(w_k^U\right)^T\left(w_k^U\right)\right]\leq\frac{8(\ln(|\mathcal{B}|)+\omega)^2}{\omega^2(1-\gamma)^2}.$$

Lemma 4. *For every $k\geq 0$,*

$$\mathbb{E}\left[w_k^U\right]=0.$$

The proofs of Lemmas 3 and 4 are presented in Appendices J and K, respectively. Then, Theorem 5 can be proved as follows:

Proof of Theorem 5. Using (16) recursively, one obtains

$$Q_k^{UL,bz}-Q^*=A_{Q^*}^k\left(Q_0^{UL,bz}-Q^*\right)+\alpha\sum_{i=0}^{k-1}A_{Q^*}^{k-i-1}w_i^U+\alpha\gamma\sum_{i=0}^{k-1}A_{Q^*}^{k-i-1}D\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}.$$

Then, considering the norm and expectation of the preceding equality yields

$$
\begin{aligned}
&\mathbb{E}\left[\left\|Q_k^{UL,bz}-Q^*\right\|_2\right]\\
=&\mathbb{E}\left[\left\|A_{Q^*}^k\left(Q_0^{UL,bz}-Q^*\right)+\alpha\sum_{i=0}^{k-1}A_{Q^*}^{k-i-1}w_i^U+\alpha\gamma\sum_{i=0}^{k-1}A_{Q^*}^{k-i-1}D\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}\right\|_2\right]\\
\leq&\mathbb{E}\left[\left\|A_{Q^*}^k\left(Q_0^{UL,bz}-Q^*\right)+\alpha\sum_{i=0}^{k-1}A_{Q^*}^{k-i-1}w_i^U\right\|_2\right]+\mathbb{E}\left[\left\|\alpha\gamma\sum_{i=0}^{k-1}A_{Q^*}^{k-i-1}D\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}\right\|_2\right].
\end{aligned}
$$

Utilizing the relation $\mathbb{E}[\|\cdot\|_2]=\mathbb{E}[\sqrt{\|\cdot\|_2^2}]\leq\sqrt{\mathbb{E}[\|\cdot\|_2^2]}$, the last inequality becomes

$$\mathbb{E}\left[\left\|Q_k^{UL,bz}-Q^*\right\|_2\right]\leq\sqrt{\mathbb{E}\left[\left\|A_{Q^*}^k\left(Q_0^{UL,bz}-Q^*\right)+\alpha\sum_{i=0}^{k-1}A_{Q^*}^{k-i-1}w_i^U\right\|_2^2\right]}$$

$$+\mathbb{E}\left[\left\|\alpha\gamma\sum_{i=0}^{k-1}A_{Q^*}^{k-i-1}D\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}\right\|_2\right].$$

Since

$$\begin{aligned}
&\mathbb{E}\left[\left\|A_{Q^*}^k\left(Q_0^{UL,bz}-Q^*\right)+\alpha\sum_{i=0}^{k-1}A_{Q^*}^{k-i-1}w_i^U\right\|_2^2\right]\\
=&\mathbb{E}\left[\left(A_{Q^*}^k\left(Q_0^{UL,bz}-Q^*\right)+\alpha\sum_{i=0}^{k-1}A_{Q^*}^{k-i-1}w_i^U\right)^T\right.\\
&\left.\cdot\left(A_{Q^*}^k\left(Q_0^{UL,bz}-Q^*\right)+\alpha\sum_{i=0}^{k-1}A_{Q^*}^{k-i-1}w_i^U\right)\right]\\
=&\mathbb{E}\left[\left(Q_0^{UL,bz}-Q^*\right)^T\left(A_{Q^*}^k\right)^TA_{Q^*}^k\left(Q_0^{UL,bz}-Q^*\right)\right]\\
&+\mathbb{E}\left[\alpha^2\sum_{i=0}^{k-1}\left(w_i^U\right)^T\left(A_{Q^*}^{k-i-1}\right)^TA_{Q^*}^{k-i-1}w_i^U\right]\\
\leq&\mathbb{E}\left[\lambda_{\max}\left(\left(A_{Q^*}^k\right)^TA_{Q^*}^k\right)\left(Q_0^{UL,bz}-Q^*\right)^T\left(Q_0^{UL,bz}-Q^*\right)\right]\\
&+\mathbb{E}\left[\alpha^2\sum_{i=0}^{k-1}\lambda_{\max}\left(\left(A_{Q^*}^{k-i-1}\right)^TA_{Q^*}^{k-i-1}\right)\left(w_i^U\right)^Tw_i^U\right]\\
=&\left\|A_{Q^*}^k\right\|_2^2\left\|Q_0^{UL,bz}-Q^*\right\|_2^2+\mathbb{E}\left[\alpha^2\sum_{i=0}^{k-1}\left\|A_{Q^*}^{k-i-1}\right\|_2^2\left(w_i^U\right)^Tw_i^U\right]
\end{aligned}$$

is satisfied by applying Lemma 4 in the second equality and utilizing $\lambda_{\max}$, which denotes the maximum eigenvalue, in the first inequality, one gets

$$\begin{aligned}
\mathbb{E}\left[\left\|Q_k^{UL,bz}-Q^*\right\|_2\right]\leq&\sqrt{\left\|A_{Q^*}^k\right\|_2^2\left\|Q_0^{UL,bz}-Q^*\right\|_2^2+\mathbb{E}\left[\alpha^2\sum_{i=0}^{k-1}\left\|A_{Q^*}^{k-i-1}\right\|_2^2\left(w_i^U\right)^Tw_i^U\right]}\\
&+\mathbb{E}\left[\left\|\alpha\gamma\sum_{i=0}^{k-1}A_{Q^*}^{k-i-1}D\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}\right\|_2\right].
\end{aligned}$$

Moreover, combining the prior inequality with $\|\cdot\|_2\leq|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}\|\cdot\|_\infty$ and Lemma 2 produces

$$\begin{aligned}
&\mathbb{E}\left[\left\|Q_k^{UL,bz}-Q^*\right\|_2\right]\\
\leq&\sqrt{|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|\left\|A_{Q^*}^k\right\|_\infty^2\left\|Q_0^{UL,bz}-Q^*\right\|_2^2+\mathbb{E}\left[\alpha^2\sum_{i=0}^{k-1}|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|\left\|A_{Q^*}^{k-i-1}\right\|_\infty^2\left(w_i^U\right)^Tw_i^U\right]}\\
&+\frac{\alpha\gamma d_{max}\ln(|\mathcal{B}|)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega}\sum_{i=0}^{k-1}\left\|A_{Q^*}^{k-i-1}\right\|_\infty\\
\leq&\sqrt{|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|\left\|Q_0^{UL,bz}-Q^*\right\|_2^2\rho^{2k}+\alpha^2|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|\sum_{i=0}^{k-1}\rho^{2(k-i-1)}\mathbb{E}\left[\left(w_i^U\right)^Tw_i^U\right]}\\
&+\frac{\alpha\gamma d_{max}\ln(|\mathcal{B}|)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega}\sum_{i=0}^{k-1}\rho^{k-i-1}.
\end{aligned}$$

Applying Lemma 3 and the relation $\sum_{i=0}^{k-1}\rho^i \leq \sum_{i=0}^{\infty}\rho^i \leq \frac{1}{1-\rho}$, $\sum_{i=0}^{k-1}\rho^{2i} \leq \sum_{i=0}^{\infty}\rho^{2i} \leq \frac{1}{1-\rho^2} \leq \frac{1}{1-\rho}$ with Definition 1 to the last inequality, one obtains

$$\mathbb{E}\left[\left\|Q_k^{UL,bz}-Q^*\right\|_2\right] \leq \sqrt{|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|\left\|Q_0^{UL,bz}-Q^*\right\|_2^2\rho^{2k}+\frac{8\alpha(\ln(|\mathcal{B}|)+\omega)^2|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}{\omega^2 d_{min}(1-\gamma)^3}} + \frac{\gamma d_{max}\ln(|\mathcal{B}|)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}(1-\gamma)}.$$

The final conclusion can be reached by utilizing the subadditivity of the square root function in the aforementioned inequality. □

J Proof of Lemma 3

In order to derive Lemma 3, the following lemma is presented first:

Lemma 5. *For every $k \geq 0$,*

$$\left\|Q_k^{U,bz}\right\|_\infty \leq \frac{1}{1-\gamma}\left(1+\gamma\frac{\ln(|\mathcal{B}|)}{\omega}\right).$$

Proof. From (5), the upper comparison system of Boltzmann MQL can be represented as

$$\begin{aligned} Q_{i+1}^{U,bz} =& Q_i^{U,bz} + \alpha\Big\{(e_{a_i}\otimes e_{b_i}\otimes e_{s_i})r_i + \gamma(e_{a_i}\otimes e_{b_i}\otimes e_{s_i})\left(e_{s_i'}\right)^T\Pi_{Q_k^{U,bz}}Q_k^{U,bz} \\ &- (e_{a_i}\otimes e_{b_i}\otimes e_{s_i})(e_{a_i}\otimes e_{b_i}\otimes e_{s_i})^T Q_i^{U,bz} + \gamma(e_{a_i}\otimes e_{b_i}\otimes e_{s_i})\frac{\ln(|\mathcal{B}|)}{\omega}\Big\}. \end{aligned} \tag{24}$$

Then, taking the norm on both sides of (24) with $i=0$ results in

$$\begin{aligned} \left\|Q_1^{U,bz}\right\|_\infty \leq& (1-\alpha)\left\|Q_0^{U,bz}\right\|_\infty + \alpha\left\{\|r_0\|_\infty + \gamma\left\|\Pi_{Q_0^{U,bz}}Q_0^{U,bz}\right\|_\infty + \gamma\left\|\frac{\ln(|\mathcal{B}|)}{\omega}\right\|_\infty\right\} \\ \leq& (1-\alpha)+\alpha+\alpha\gamma+\alpha\gamma\frac{\ln(|\mathcal{B}|)}{\omega} \\ =& 1+\alpha\gamma+\alpha\gamma\frac{\ln(|\mathcal{B}|)}{\omega} \\ \leq& (1+\gamma)+\gamma\frac{\ln(|\mathcal{B}|)}{\omega}, \end{aligned}$$

from which the second and last inequalities are obtained from Assumption 1.

Next, suppose the following result is valid for some $i=k-1\geq 0$ to utilize an induction argument:

$$\left\|Q_k^{U,bz}\right\|_\infty \leq \left(1+\gamma+\cdots+\gamma^k\right)+\left(\gamma+\cdots+\gamma^k\right)\frac{\ln(|\mathcal{B}|)}{\omega}. \tag{25}$$

Using $i=k$ to obtain the norm on (24) yields

$$\begin{aligned} \left\|Q_{k+1}^{U,bz}\right\|_\infty \leq& (1-\alpha)\left\|Q_k^{U,bz}\right\|_\infty + \alpha\left\{\|r_k\|_\infty + \gamma\left\|\Pi_{Q_k^{U,bz}}Q_k^{U,bz}\right\|_\infty + \gamma\left\|\frac{\ln(|\mathcal{B}|)}{\omega}\right\|_\infty\right\} \\ \leq& (1-\alpha)\left\{\left(1+\gamma+\cdots+\gamma^k\right)+\left(\gamma+\cdots+\gamma^k\right)\frac{\ln(|\mathcal{B}|)}{\omega}\right\} \\ &+\alpha\gamma\left\{\left(1+\gamma+\cdots+\gamma^k\right)+\left(\gamma+\cdots+\gamma^k\right)\frac{\ln(|\mathcal{B}|)}{\omega}\right\}+\alpha+\alpha\gamma\frac{\ln(|\mathcal{B}|)}{\omega} \end{aligned}$$

$$
\begin{aligned}
&=(1-\alpha)\left\{\left(1+\gamma+\cdots+\gamma^k\right)+\left(\gamma+\cdots+\gamma^k\right)\frac{\ln(|\mathcal{B}|)}{\omega}\right\}\\
&\quad+\alpha\left\{\left(1+\gamma+\cdots+\gamma^k\right)+\left(\gamma+\cdots+\gamma^k\right)\frac{\ln(|\mathcal{B}|)}{\omega}\right\}+\alpha\left(\gamma^{k+1}+\gamma^{k+1}\frac{\ln(|\mathcal{B}|)}{\omega}\right)\\
&=\left\{\left(1+\gamma+\cdots+\gamma^k\right)+\left(\gamma+\cdots+\gamma^k\right)\frac{\ln(|\mathcal{B}|)}{\omega}\right\}+\alpha\left(\gamma^{k+1}+\gamma^{k+1}\frac{\ln(|\mathcal{B}|)}{\omega}\right)\\
&\leq\left(1+\gamma+\cdots+\gamma^k+\gamma^{k+1}\right)+\left(\gamma+\cdots+\gamma^k+\gamma^{k+1}\right)\frac{\ln(|\mathcal{B}|)}{\omega},
\end{aligned}
$$

with Assumption 1 and (25) in the second and last inequalities.

This leads to

$$
\begin{aligned}
\left\|Q_k^{U,bz}\right\|_\infty &\leq\left(1+\gamma+\cdots+\gamma^k\right)+\left(\gamma+\cdots+\gamma^k\right)\frac{\ln(|\mathcal{B}|)}{\omega}\\
&=\left(1+\gamma+\cdots+\gamma^k\right)+\gamma\left(1+\gamma+\cdots+\gamma^{k-1}\right)\frac{\ln(|\mathcal{B}|)}{\omega}\\
&\leq\sum_{i=0}^{\infty}\gamma^i+\gamma\frac{\ln(|\mathcal{B}|)}{\omega}\sum_{i=0}^{\infty}\gamma^i\\
&\leq\frac{1}{1-\gamma}\left(1+\gamma\frac{\ln(|\mathcal{B}|)}{\omega}\right),
\end{aligned}
$$

which represents the final result. □

Now, the demonstration of Lemma 3 is ready.

Proof of Lemma 3.

$$
\begin{aligned}
&\mathbb{E}\left[\left(w_k^U\right)^T\left(w_k^U\right)\right]\\
=&\mathbb{E}\left[\left\|w_k^U\right\|_2^2\right]\\
=&\mathbb{E}\Bigg[\Bigg\|(e_{a_k}\otimes e_{b_k}\otimes e_{s_k})r_k+\gamma(e_{a_k}\otimes e_{b_k}\otimes e_{s_k})\left(e_{s'_k}\right)^T\Pi_{Q_k^{U,bz}}Q_k^{U,bz}\\
&-(e_{a_k}\otimes e_{b_k}\otimes e_{s_k})(e_{a_k}\otimes e_{b_k}\otimes e_{s_k})^TQ_k^{U,bz}+\gamma(e_{a_k}\otimes e_{b_k}\otimes e_{s_k})\frac{\ln(|\mathcal{B}|)}{\omega}\\
&-\left(DR+\gamma DP\Pi_{Q_k^{U,bz}}Q_k^{U,bz}-DQ_k^{U,bz}+\gamma D\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}\right)\Bigg\|_2^2\Bigg]\\
=&\mathbb{E}\left[\mathbb{E}\left[\left\|r_k+\gamma\left(e_{s'_k}\right)^T\Pi_{Q_k^{U,bz}}Q_k^{U,bz}-(e_{a_k}\otimes e_{b_k}\otimes e_{s_k})^TQ_k^{U,bz}+\gamma\frac{\ln(|\mathcal{B}|)}{\omega}\right\|_2^2\middle|Q_k^{U,bz}\right]\right]\\
&-\mathbb{E}\Bigg[\mathbb{E}\Bigg[2\left(r_k+\gamma\left(e_{s'_k}\right)^T\Pi_{Q_k^{U,bz}}Q_k^{U,bz}-(e_{a_k}\otimes e_{b_k}\otimes e_{s_k})^TQ_k^{U,bz}+\gamma\frac{\ln(|\mathcal{B}|)}{\omega}\right)\\
&\cdot\left(DR+\gamma DP\Pi_{Q_k^{U,bz}}Q_k^{U,bz}-DQ_k^{U,bz}+\gamma D\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}\right)\Bigg|Q_k^{U,bz}\Bigg]\Bigg]\\
&+\mathbb{E}\left[\mathbb{E}\left[\left\|DR+\gamma DP\Pi_{Q_k^{U,bz}}Q_k^{U,bz}-DQ_k^{U,bz}+\gamma D\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}\right\|_2^2\middle|Q_k^{U,bz}\right]\right]\\
=&\mathbb{E}\left[\left\|r_k+\gamma\left(e_{s'_k}\right)^T\Pi_{Q_k^{U,bz}}Q_k^{U,bz}-(e_{a_k}\otimes e_{b_k}\otimes e_{s_k})^TQ_k^{U,bz}+\gamma\frac{\ln(|\mathcal{B}|)}{\omega}\right\|_2^2\right]\\
&-\left\|DR+\gamma DP\Pi_{Q_k^{U,bz}}Q_k^{U,bz}-DQ_k^{U,bz}+\gamma D\frac{\ln(|\mathcal{B}|)}{\omega}\mathbf{1}\right\|_2^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq \mathbb{E}\left[\left\| r_k + \gamma \left(e_{s'_k}\right)^T \Pi_{Q_k^{U,bz}} Q_k^{U,bz} - (e_{a_k} \otimes e_{b_k} \otimes e_{s_k})^T Q_k^{U,bz} + \gamma \frac{\ln(|\mathcal{B}|)}{\omega} \right\|_2^2\right] \\
&\leq 4\mathbb{E}\left[\|r_k\|_2^2\right] + 4\gamma^2 \mathbb{E}\left[\left\| \left(e_{s'_k}\right)^T \Pi_{Q_k^{U,bz}} Q_k^{U,bz} \right\|_2^2\right] + 4\mathbb{E}\left[\left\| (e_{a_k} \otimes e_{b_k} \otimes e_{s_k})^T Q_k^{U,bz} \right\|_2^2\right] \\
&\quad + 4\gamma^2 \mathbb{E}\left[\left\| \frac{\ln(|\mathcal{B}|)}{\omega} \right\|_2^2\right] \\
&\leq 4 + 4\gamma^2 \left\{ \frac{1}{1-\gamma} \left(1 + \gamma \frac{\ln(|\mathcal{B}|)}{\omega}\right) \right\}^2 + 4 \left\{ \frac{1}{1-\gamma} \left(1 + \gamma \frac{\ln(|\mathcal{B}|)}{\omega}\right) \right\}^2 + 4\gamma^2 \left(\frac{\ln(|\mathcal{B}|)}{\omega}\right)^2 \\
&\leq \frac{8(\ln(|\mathcal{B}|) + \omega)^2}{\omega^2 (1-\gamma)^2},
\end{aligned}
$$

where the second inequality arises from $\|a+b+c+d\|_2^2 \leq 4\|a\|_2^2 + 4\|b\|_2^2 + 4\|c\|_2^2 + 4\|d\|_2^2$ for any a, b, c, d by using the Cauchy-Schwarz inequality, and the third inequality derives from Assumption 1 and Lemma 5. □

K Proof of Lemma 4

Proof. Considering the expectation on (9), one obtains

$$
\begin{aligned}
\mathbb{E}\left[w_k^U\right] =& \mathbb{E}\Big[(e_{a_k} \otimes e_{b_k} \otimes e_{s_k}) r_k + \gamma (e_{a_k} \otimes e_{b_k} \otimes e_{s_k}) \left(e_{s'_k}\right)^T \Pi_{Q_k^{U,bz}} Q_k^{U,bz} \\
&- (e_{a_k} \otimes e_{b_k} \otimes e_{s_k})(e_{a_k} \otimes e_{b_k} \otimes e_{s_k})^T Q_k^{U,bz} + \gamma (e_{a_k} \otimes e_{b_k} \otimes e_{s_k}) \frac{\ln(|\mathcal{B}|)}{\omega} \\
&- \left(DR + \gamma DP \Pi_{Q_k^{U,bz}} Q_k^{U,bz} - DQ_k^{U,bz} + \gamma D \frac{\ln(|\mathcal{B}|)}{\omega} \mathbf{1} \right) \Big] \\
=& DR + \gamma DP \Pi_{Q_k^{U,bz}} Q_k^{U,bz} - DQ_k^{U,bz} + \gamma D \frac{\ln(|\mathcal{B}|)}{\omega} \mathbf{1} \\
&- \left\{ DR + \gamma DP \Pi_{Q_k^{U,bz}} Q_k^{U,bz} - DQ_k^{U,bz} + \gamma D \frac{\ln(|\mathcal{B}|)}{\omega} \mathbf{1} \right\} \\
=& 0.
\end{aligned}
$$

□

L Proof of Theorem 6

To prove the convergence of the error system (18), the following additional lemma is needed:

Lemma 6 (Gosavi (2006))**.** *The value of Q^* is bounded by*

$$
\|Q^*\|_\infty \leq \frac{1}{1-\gamma}.
$$

This allows us to prove Theorem 6.

Proof of Theorem 6. Taking into account the norm of (18) with Lemma 2, one gets

$$
\begin{aligned}
\left\| Q_{i+1}^{UU,bz} - Q_{i+1}^{UL,bz} \right\|_\infty &\leq \|A_{Q_i}\|_\infty \left\| Q_i^{UU,bz} - Q_i^{UL,bz} \right\|_\infty + \|B_{Q_i}\|_\infty \left\| Q_k^{UL,bz} - Q^* \right\|_\infty \\
&\leq \rho \left\| Q_i^{UU,bz} - Q_i^{UL,bz} \right\|_\infty + 2\alpha\gamma d_{max} \left\| Q_k^{UL,bz} - Q^* \right\|_\infty .
\end{aligned}
$$

Then, considering the expectation of the prior inequality results in

$$\begin{aligned}\mathbb{E}\left[\left\|Q_{i+1}^{UU,bz}-Q_{i+1}^{UL,bz}\right\|_\infty\right] \leq & \rho\mathbb{E}\left[\left\|Q_{i}^{UU,bz}-Q_{i}^{UL,bz}\right\|_\infty\right]+2\alpha\gamma d_{max}\mathbb{E}\left[\left\|Q_{k}^{UL,bz}-Q^*\right\|_2\right] \\ \leq & \rho\mathbb{E}\left[\left\|Q_{i}^{UU,bz}-Q_{i}^{UL,bz}\right\|_\infty\right] \\ & +2\alpha\gamma d_{max}|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}\left\|Q_0^{UL,bz}-Q^*\right\|_2\rho^i \\ & +\frac{4\sqrt{2}\alpha^{\frac{3}{2}}\gamma d_{max}(\ln(|\mathcal{B}|)+\omega)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^{\frac{1}{2}}(1-\gamma)^{\frac{3}{2}}} \\ & +\frac{2\alpha\gamma^2 d_{max}^2\ln(|\mathcal{B}|)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}(1-\gamma)},\end{aligned}$$

where the last inequality derives from Theorem 5.

By unrolling the aforementioned inequality from $i=0$ to $k-1$ and assuming $Q_0^{UU,bz}=Q_0^{UL,bz}$, one obtains

$$\begin{aligned}\mathbb{E}\left[\left\|Q_{k}^{UU,bz}-Q_{k}^{UL,bz}\right\|_\infty\right] \leq & 2\alpha\gamma d_{max}|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}\left\|Q_0^{UL,bz}-Q^*\right\|_2 k\rho^{k-1} \\ & +\frac{4\sqrt{2}\alpha^{\frac{3}{2}}\gamma d_{max}(\ln(|\mathcal{B}|)+\omega)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^{\frac{1}{2}}(1-\gamma)^{\frac{3}{2}}}\sum_{i=0}^{k-1}\rho^i \\ & +\frac{2\alpha\gamma^2 d_{max}^2\ln(|\mathcal{B}|)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}(1-\gamma)}\sum_{i=0}^{k-1}\rho^i.\end{aligned}$$

Utilizing the relation

$$\begin{aligned}\left\|Q_0^{UL,bz}-Q^*\right\|_2 \leq & |\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}\left\|Q_0^{UL,bz}-Q^*\right\|_\infty \\ \leq & |\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}\left(\left\|Q_0^{UL,bz}\right\|_\infty+\|Q^*\|_\infty\right) \\ \leq & |\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}\left(1+\frac{1}{1-\gamma}\right) \\ \leq & |\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}\frac{2}{1-\gamma}\end{aligned}$$

with Assumption 1 and Lemma 6 and $\sum_{i=0}^{k-1}\rho^i \leq \sum_{i=0}^{\infty}\rho^i \leq \frac{1}{1-\rho}$ with Definition 1, the last inequality becomes

$$\begin{aligned}\mathbb{E}\left[\left\|Q_{k}^{UU,bz}-Q_{k}^{UL,bz}\right\|_\infty\right] \leq & \frac{4\alpha\gamma d_{max}|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|}{1-\gamma}k\rho^{k-1} \\ & +\frac{4\sqrt{2}\alpha^{\frac{1}{2}}\gamma d_{max}(\ln(|\mathcal{B}|)+\omega)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^{\frac{3}{2}}(1-\gamma)^{\frac{5}{2}}} \\ & +\frac{2\gamma^2 d_{max}^2\ln(|\mathcal{B}|)|\mathcal{S}\times\mathcal{A}\times\mathcal{B}|^{\frac{1}{2}}}{\omega d_{min}^2(1-\gamma)^2},\end{aligned}$$

which completes the proof. □

M Proof of Proposition 8 and Proposition 9

Propositions 8 and 9 can be proven using an induction argument.

Proof of Proposition 8. Suppose the result is valid for some $k \geq 0$. Then,

$$\begin{aligned}
Q_{k+1}^{L,bz} - Q^* =& A'_{Q_k^{L,bz}} \left(Q_k^{L,bz} - Q^*\right) + b'_{Q_k^{L,bz}} + \alpha w_k^L - \alpha\gamma D \frac{\ln(|\mathcal{A}|)}{\omega} \mathbf{1} \\
\geq& A'_{Q_k^{L,bz}} \left(Q_k^{L,bz} - Q^*\right) + \alpha w_k^L - \alpha\gamma D \frac{\ln(|\mathcal{A}|)}{\omega} \mathbf{1} \\
\geq& A'_{Q_k^{L,bz}} \left(Q_k^{LL,bz} - Q^*\right) + \alpha w_k^L - \alpha\gamma D \frac{\ln(|\mathcal{A}|)}{\omega} \mathbf{1} \\
=& Q_{k+1}^{LL,bz} - Q^*,
\end{aligned}$$

where the first inequality utilizes the property of $b'_{Q_k^{L,bz}} = \alpha\gamma DP \left(\Gamma_{Q_k^{L,bz}} - \Gamma_{Q^*}\right) Q^* \geq \alpha\gamma DP \left(\Gamma_{Q^*} - \Gamma_{Q^*}\right) Q^* = 0$, and the last inequality is based on the assumption $Q_k^{LL,bz}(s,a,b) \leq Q_k^{L,bz}(s,a,b)$ and the fact that $A'_{Q_k^{L,bz}}$ is a nonnegative matrix. By induction, the proof is completed. □

Proof of Proposition 9. Suppose the result is valid for some $k \geq 0$. Then,

$$\begin{aligned}
& Q_{k+1}^{L,bz} - Q^* \\
=& A'_{Q_k^{L,bz}} (Q_k^{L,bz} - Q^*) + b'_{Q_k^{L,bz}} + \alpha w_k^L - \alpha\gamma D \frac{\ln(|\mathcal{A}|)}{\omega} \mathbf{1} \\
=& A'_{Q^*} (Q_k^{L,bz} - Q^*) + (A'_{Q_k^{L,bz}} - A'_{Q^*})(Q_k^{L,bz} - Q^*) + b'_{Q_k^{L,bz}} + \alpha w_k^L - \alpha\gamma D \frac{\ln(|\mathcal{A}|)}{\omega} \mathbf{1} \\
=& A'_{Q^*} (Q_k^{L,bz} - Q^*) + \alpha\gamma DP \left(\Gamma_{Q_k^{L,bz}} - \Gamma_{Q^*}\right) Q_k^{L,bz} + \alpha w_k^L - \alpha\gamma D \frac{\ln(|\mathcal{A}|)}{\omega} \mathbf{1} \\
\leq& A'_{Q^*} (Q_k^{L,bz} - Q^*) + \alpha w_k^L - \alpha\gamma D \frac{\ln(|\mathcal{A}|)}{\omega} \mathbf{1} \\
\leq& A'_{Q^*} (Q_k^{LU,bz} - Q^*) + \alpha w_k^L - \alpha\gamma D \frac{\ln(|\mathcal{A}|)}{\omega} \mathbf{1} \\
=& Q_{k+1}^{LU,bz} - Q^*,
\end{aligned}$$

where the first inequality utilizes $\alpha\gamma DP \left(\Gamma_{Q_k} - \Gamma_{Q^*}\right) Q_k \leq \alpha\gamma DP \left(\Gamma_{Q^*} - \Gamma_{Q^*}\right) Q_k = 0$, and the last inequality is based on the assumption $Q_k^{LU,bz}(s,a,b) \geq Q_k^{L,bz}(s,a,b)$ and the fact that A'_{Q^*} is a nonnegative matrix. By induction, the proof is completed. □

N Proof of Theorem 4

Proof. The final conclusion can be obtained by using the relation

$$\begin{aligned}
\mathbb{E}\left[\left\|Q_k^{bz} - Q^*\right\|_2\right] =& \mathbb{E}\left[\left\|Q_k^{bz} - Q_k^{L,bz} + Q_k^{L,bz} - Q^*\right\|_2\right] \\
\leq& \mathbb{E}\left[\left\|Q_k^{L,bz} - Q^*\right\|_2\right] + \mathbb{E}\left[\left\|Q_k^{bz} - Q_k^{L,bz}\right\|_2\right] \\
\leq& \mathbb{E}\left[\left\|Q_k^{L,bz} - Q^*\right\|_2\right] + \mathbb{E}\left[\left\|Q_k^{U,bz} - Q_k^{L,bz}\right\|_2\right] \\
\leq& \mathbb{E}\left[\left\|Q_k^{L,bz} - Q^*\right\|_2\right] + \mathbb{E}\left[\left\|Q_k^{U,bz} - Q^* + Q^* - Q_k^{L,bz}\right\|_2\right] \\
\leq& \mathbb{E}\left[\left\|Q_k^{L,bz} - Q^*\right\|_2\right] + \mathbb{E}\left[\left\|Q_k^{U,bz} - Q^*\right\|_2\right] + \mathbb{E}\left[\left\|Q^* - Q_k^{L,bz}\right\|_2\right] \\
=& 2\mathbb{E}\left[\left\|Q_k^{L,bz} - Q^*\right\|_2\right] + \mathbb{E}\left[\left\|Q_k^{U,bz} - Q^*\right\|_2\right]
\end{aligned}$$

with Theorems 2 and 3, where the triangle inequality accounts for the first and fourth inequalities and the fact that $Q_k^{U,bz} - Q_k^{L,bz} \geq Q_k^{bz} - Q_k^{L,bz} \geq 0$ provides the second inequality. □

O Numerical simulation settings

To show the simulation results for the convergence of MQL and Boltzmann MQL, we consider a Markov game with $\mathcal{S} = \{1, 2\}$, $\mathcal{A} = \{1, 2\}$, $\mathcal{B} = \{1, 2\}$, $\alpha = 0.1$, $\gamma = 0.9$, and $\omega = 100$. We set the state transition probability matrices as

$$P_{1,1} = \begin{bmatrix} 0.8 & 0.2 \\ 0.7 & 0.3 \end{bmatrix}, P_{1,2} = \begin{bmatrix} 0.2 & 0.8 \\ 0.3 & 0.7 \end{bmatrix},$$

$$P_{2,1} = \begin{bmatrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{bmatrix}, P_{2,2} = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix},$$

the reward function as

$$r(\cdot, 1, 1, \cdot) = \begin{bmatrix} r(1,1,1,1) \\ r(1,1,1,2) \\ r(2,1,1,1) \\ r(2,1,1,2) \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ -0.3 \\ 0.3 \end{bmatrix}, r(\cdot, 1, 2, \cdot) = \begin{bmatrix} r(1,1,2,1) \\ r(1,1,2,2) \\ r(2,1,2,1) \\ r(2,1,2,2) \end{bmatrix} = \begin{bmatrix} -0.7 \\ 0.7 \\ 0.2 \\ -0.2 \end{bmatrix},$$

$$r(\cdot, 2, 1, \cdot) = \begin{bmatrix} r(1,2,1,1) \\ r(1,2,1,2) \\ r(2,2,1,1) \\ r(2,2,1,2) \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ 0.3 \\ -0.3 \end{bmatrix}, r(\cdot, 2, 2, \cdot) = \begin{bmatrix} r(1,2,2,1) \\ r(1,2,2,2) \\ r(2,2,2,1) \\ r(2,2,2,2) \end{bmatrix} = \begin{bmatrix} 0.7 \\ -0.7 \\ -0.2 \\ 0.2 \end{bmatrix},$$

and the behavior policies as a uniform distribution. The graphs in Section 5 display the averages of the three simulation runs for each algorithm with different random initializations.

P Numerical simulation of Boltzmann MQL with its entire comparison systems

Figure 3 shows the simulated trajectories of the Boltzmann MQL with its entire comparison systems under the same simulation settings as in Appendix O. This result also provides empirical support for the bounding concepts used in the convergence analysis of Boltzmann MQL.

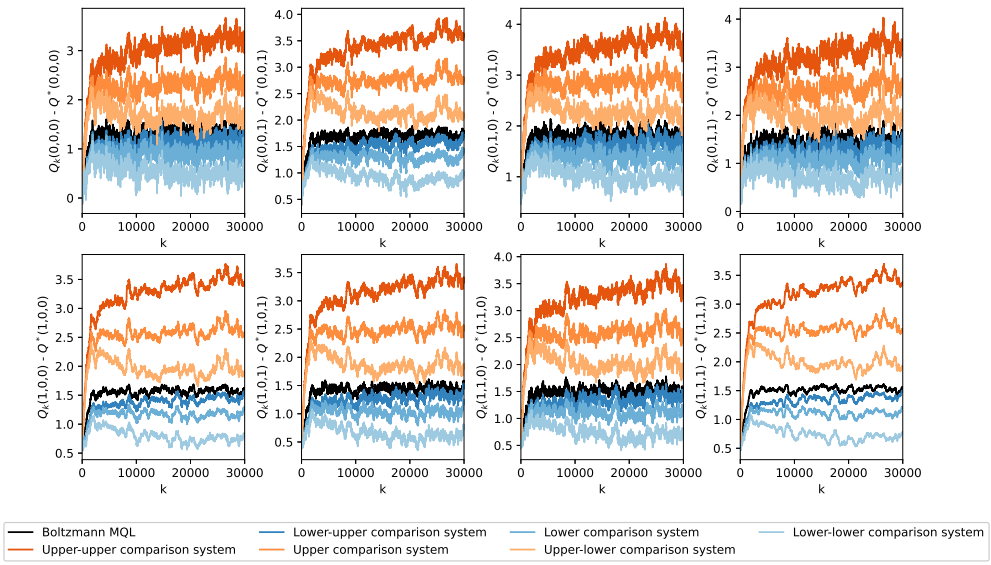

Figure 3: Trajectories of the Boltzmann MQL and its entire comparison systems