Learning Nash Equilibria in Zero-Sum Markov Games: A Single-Timescale Algorithm Under Weak Reachability

Reda Ouhamma [1] **Maryam Kamgarpour** [1]

Abstract

We consider decentralized learning for zero-sum games, where players only see their payoff information and are agnostic to the opponent's actions and payoffs. Previous works demonstrated convergence to a Nash equilibrium in this setting using double timescale algorithms under strong reachability assumptions. We address the open problem of achieving an approximate Nash equilibrium efficiently with an uncoupled and single-timescale algorithm under weaker conditions. Our contribution is a rational and convergent algorithm, utilizing Tsallis-entropy regularization in a value-iteration-based approach. The algorithm learns an approximate Nash equilibrium in polynomial time, requiring only the existence of a policy pair that induces an irreducible and aperiodic Markov chain, thus considerably weakening past assumptions. Our analysis leverages negative drift inequalities and introduces novel properties of Tsallis entropy that are of independent interest.

1. Introduction

Markov games, also known as stochastic games, are a class of multi-agent decision-making problems with a rich history dating back to the foundational work of Shapley (1953). Shapley introduced two-player zero-sum stochastic games and proved the existence of a Nash equilibrium, a central solution concept where no player can improve her payoff by unilaterally deviating from her strategy. For general Markov games, Daskalakis et al. (2023) proved that finding a Nash equilibrium is PPAD-complete. Zero-sum games, a subset of Markov games, describe competitive interactions between two players with directly opposed interests. In these games, it has been shown that a Nash equilibrium can be computed efficiently, with many algorithms proposed for this purpose under known dynamics and rewards, such as variations of policy iteration or value iteration (Shapley, 1953; Hoffman and Karp, 1966; Pollatschek and Avi-Itzhak, 1969; Van Der Wal, 1978; Filar and Tolwinski, 1991).

Multi-agent reinforcement learning (MARL) refers to learning in Markov games with uncertain dynamics or rewards. Previous works in this setting focused on two key objectives: rationality and convergence (see Bowling and Veloso (2001)). Rationality requires players to converge to a best response of their opponents when those opponents' strategies tend to be stationary, while convergence requires reaching a Nash equilibrium under self-play, *i.e.* when players use the same algorithm. Recently, there has been increasing interest in developing sample-efficient MARL algorithms (Bai and Jin, 2020; Liu et al., 2021; Daskalakis et al., 2023; Cui et al., 2023). However, these works focused primarily on no-regret guarantees since computing Nash equilibria is known to be intractable in general-sum games Daskalakis et al. (2009).

For zero-sum games, extensive progress has been made on learning Nash equilibria. Past works include Q-learning variants (Littman, 1994; Zhang et al., 2021); model-based Monte Carlo estimation of value-functions (Zhang et al., 2020); optimistic gradient descent ascent (Wei et al., 2021); and policy gradient/extra-gradient algorithms (Daskalakis et al., 2020; Zhao et al., 2022; Cen et al., 2021). However, while they established finite time guarantee for convergence to a Nash equilibrium (Zhang et al., 2020; Wei et al., 2021; Zhao et al., 2022), they required stringent assumptions on the MDPs and interactions, such as irreducibility and double timescales. Irreducibility refers to the ability to attain any state from any other state under any policy of the players in finite time, and double timescales entail fixing the policies for epochs of time while continuously updating the value functions. These assumptions restrict the practical applicability of past works.

Reachability challenge. Previous analyses of zero-sum Markov games relied on a strong reachability assumption (Wei et al., 2017; 2021; Chen et al., 2021; Cai et al., 2024). Namely, there exists a positive integer L such that *for any*

[1]SYCAMORE Lab, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland.. Correspondence to: Reda Ouhamma <reda.ouhamma@epfl.ch>.

Workshop on Foundations of Reinforcement Learning and Control at the 41^{st} International Conference on Machine Learning, Vienna, Austria.

pair of policies used by the players, any state can be visited from any state in time bounded by L. Given prior knowledge of L, Wei et al. (2021) provided an algorithm that guarantees an average-iterate convergence to an ϵ approximate Nash equilibrium in $1/\epsilon^8$ time. Without knowing L, Cai et al. (2024) proposed an algorithm with a last-iterate convergence in $L^{1/\xi}/\epsilon^{9+\xi}$ time for any $\xi > 0$. To the best of our knowledge, Chen et al. (2023) is the only work relaxing this strong reachability assumption; however, they did not provide polynomial time convergence to a Nash equilibrium of the game, and rather to a biased version of the game arising from a regularization.

Timescale challenge. Most existing algorithms for learning in zero-sum games adopted double timescales (Chen et al., 2021; Wei et al., 2021). However, this entails an implicit coordination between the players to stop updating their policies simultaneously for periods of time. This limitation is also relevant for single-agent actor-critic algorithms (Xu et al., 2020a;b; Kumar et al., 2023), where Olshevsky and Gharesifard (2023) recently proposed a single-timescale algorithm; however, their analysis requires a stringent assumption of the existence of a uniform upper-bound on the mixing times *over all policies*. For zero-sum games, Chen et al. (2023) introduced a single-timescale algorithm that relaxes the above assumption, though their convergence suffers a persistent bias as explained earlier.

The objectives we aim to address in this paper can be summarized in the following question:

In zero-sum-games, can we learn an approximate Nash equilibrium in polynomial time with a single-timescale algorithm without a uniform mixing time bound or strong reachability?

Contributions. In this paper, we address payoff-based zero-sum Markov games, focusing on the fundamental problem of learning an approximate Nash equilibrium under weaker assumptions. Our contributions include:

- For a decentralized, convergent, rational, and single-timescale algorithm, we establish a polynomial sample complexity for learning a Nash equilibrium. This convergence result requires only the existence of an irreducible and aperiodic policy pair. This assumption is significantly weaker than the mixing time and reachability assumptions mentioned above.
- Our key technical contribution is introducing the Tsallis-entropy regularization for zero-sum stochastic games. This regularization allows us to:
  - Derive lower bounds on the policies, ensuring sufficient exploration (see Lemma 4.1);
  - Obtain upper bounds on mixing times, allowing us to determine when the Markov chains are close to their stationary distributions (see Lemma B.5);
  - Show a smoothness property entailing the convergence of our policies (see Lemma B.3).

2. Preliminaries

In this section, we introduce the setting of stochastic zero-sum games, define the relevant performance measure, and discuss the limitations associated with certain common assumptions in previous works.

Notations. We denote the sets of real and natural numbers by $\mathbb{R}$ and $\mathbb{N}$, respectively. The opponent of player $i \in \{1, 2\}$ is denoted by $-i$. The probability simplex over a finite space $\mathcal{X}$ is $\Delta^{\mathcal{X}}$.

2.1. Setting

Setting. We consider infinite-horizon, payoff-based, two-player zero-sum games. This is defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}^1, \mathcal{A}^2, p, \mathcal{R}^1, \mathcal{R}^2, \gamma)$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}^1$ and $\mathcal{A}^2$ are finite action spaces for players 1 and 2, respectively. The transition kernel $P : \mathcal{S} \times \mathcal{A}^1 \times \mathcal{A}^2 \mapsto \Delta_{\mathcal{S}}$ specifies the probability $P(s'|s, a^1, a^2)$ of transition from s to s' given actions a^1 and a^2. The reward functions for players 1 and 2 are denoted as $\mathcal{R}^1 : \mathcal{S} \times \mathcal{A}^1 \times \mathcal{A}^2 \mapsto \mathbb{R}$ and $\mathcal{R}^2 : \mathcal{S} \times \mathcal{A}^2 \times \mathcal{A}^1 \mapsto \mathbb{R}$, respectively, such that $\max_{s,a^1,a^2} |\mathcal{R}^1(s, a^1, a^2)| \leq 1$, and $\mathcal{R}^1(s, a^1, a^2) + \mathcal{R}^2(s, a^2, a^1) = 0$ for all s, a^1, a^2. The discount factor is denoted by γ, and satisfies $0 < \gamma < 1$. We consider a fixed initial state s_0, which Fiechter (1994) proved to be equivalent to having an initial distribution over states.

Our goal is to design an algorithm that is payoff-based and decentralized, *i.e.* players make decisions based solely on their rewards without coordinating with their opponent. Specifically, at time step k, players observe state s_k, choose actions $(a^i_k)_{i=1,2}$, observe $(\mathcal{R}^i(s_k, a^i, a^{-i}))_{i=1,2}$, and then the MDP moves to state s_{k+1}. Subsequently, the players update their strategies independently. Additionally, we assume self-play, meaning both players follow the same algorithm.

A stationary policy for player $i \in \{1, 2\}$ is a mapping π^i from S to $\Delta^{\mathcal{A}^i}$. We denote the policy pair (π^1, π^2) as π. We now define the q-function and the value function of player i:

$$q^i_\pi(s, a^i) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t \mathcal{R}^i(s_t, a^i_t, a^{-i}_t) \,|\, s_0 = s, a^i_0 = a^i\right],$$
$$v^i_\pi(s) := \mathbb{E}_{a^i\sim\pi^i(\cdot|s)}\left[q^i_\pi(s, a^i)\right],$$

where the expectation is over the randomness of the policy π and that of the transitions.

A policy pair $(\pi^1_{\text{NE}}, \pi^2_{\text{NE}})$ is a Nash equilibrium (NE) if for all $i \in \{1,2\}$, and $\pi^i \in (\Delta^{\mathcal{A}_i})^{\mathcal{S}}$:

$$v^i_{\pi^i_{\text{NE}}, \pi^{-i}_{\text{NE}}}(s) \geq v^i_{\pi^i, \pi^{-i}_{\text{NE}}}(s).$$

The above definition means that no agent can improve their value by unilaterally changing their policy. Lastly, we define the Nash Gap for a policy (π^i, π^{-i}):

$$\text{NG}(\pi^i, \pi^{-i}) := \sum_{i=1,2} \left(\max_{\tilde{\pi}^i} v^i_{(\tilde{\pi}^i, \pi^{-i})}(s_0) - v^i_{(\pi^i, \pi^{-i})}(s_0) \right).$$

From the above, it follows that the Nash Gap is zero if the policy pair constitutes a Nash equilibrium. Our objective in this paper is to develop an algorithm that learns an approximate Nash equilibrium in polynomial time without relying on strong reachability or resorting to double timescales.

2.2. Limiting assumptions in state-of-the-art

In this section, we examine two common assumptions in the literature related to the timescale and reachability objectives, discuss their limitations, and highlight the challenges of moving beyond them.

Strong reachability We begin with the common strong reachability assumption (Auer and Ortner, 2006; Chen et al., 2021), also known as the irreducible game assumption.

Assumption 2.1. (Strong Reachability) There exists a constant $L > 0$ such that:

$$\max_{s,s' \in \mathcal{S}} \max_{\pi \in (\Delta^{\mathcal{A}_i})^{\mathcal{S}} \times (\Delta^{\mathcal{A}_{-i}})^{\mathcal{S}}} T^{\pi}_{s \mapsto s'} \leq L,$$

where $T^{\pi}_{s \mapsto s'}$ is the expected time to reach state s' from state s when players follow policy π.

In particular, this implies that all the induced Markov chains are irreducible. In addition, Durrett (2019, Theorem 5.5.11) showed that the Markov chain induced by an irreducible policy π with stationary distribution μ_π satisfies

$$T^{\pi}_{s \mapsto s} = 1/\mu_\pi(s). \tag{1}$$

The above means that Assumption 2.1 implies $\mu_\pi(s) \geq 1/L$ for all states and policies. This is a prevalent assumption in reinforcement learning (Agarwal et al., 2021; Mei et al., 2020; Zhang et al., 2022), which is stringent since requiring a full support for the stationary distribution of any pair of policies is very restrictive. To illustrate, consider the MDP provided in Figure 1, and consider the policy π parameterized by $\xi \in [0,1]$ defined as:

$$\pi(1,a) = \xi \text{ and } \quad \pi(1,b) = 1 - \xi. \tag{2}$$

Then, the corresponding stationary distribution μ_π is given by:

$$\mu_\pi = \left(\frac{1}{2+\xi}, \frac{1}{2}, \frac{\xi}{4+2\xi} \right), \tag{3}$$

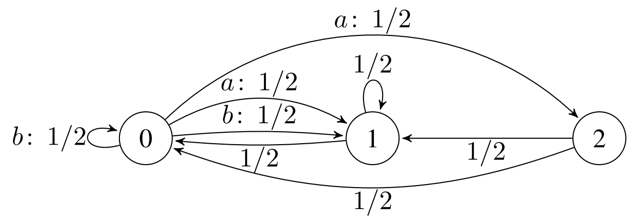

Figure 1. MDP with three states: state 0 has actions a and b, while states 2 and 3 don't have any action. The arrows indicate the possible transitions, which all have a probability of $1/2$.

which implies $\lim_{\xi \to 0} \mu_\pi(2) = 0$, invalidating Assumption 2.1 due to Equation (1) as there couldn't exist a positive L such that $\inf_s \inf_\pi \mu_\pi(s) \leq 1/L$.

For single-agent RL, Auer et al. (2008) relaxed strong reachability by requiring an upper bound on the shortest expected time to visit any state from any other state $\max_{s,s'} \min_\pi T^{\pi}_{s \mapsto s'}$. This assumption is sufficient for regret guarantees (Auer et al., 2008; Agrawal and Jia, 2017); however, it is not clear if it suffices for learning Nash equilibria in games because each player may have the ability of blocking her opponent from visiting certain states.

Single timescale A significant challenge in single timescale algorithms is that the observed payoffs constitute a time-inhomogeneous Markov chain, which cannot be studied using reachability assumptions alone. The concept of mixing times is commonly introduced to handle this challenge, as it characterizes the convergence of Markov chains to their stationary distributions.

Definition 2.2. For a policy pair $\pi = (\pi_1, \pi_2)$, $\epsilon > 0$, and a transition matrix P, assuming that the induced Markov chain has a unique stationary distribution μ_π[1], we define the ϵ-mixing time as:

$$t_{\pi,\epsilon} = \min \left\{ k \geq 0 : \max_{s \in \mathcal{S}} \left\| P^k_\pi(s, \cdot) - \mu_\pi(\cdot) \right\|_{\text{TV}} \leq \epsilon \right\},$$

where TV is the total variation distance and P_π is the transition induced by following policy π. For a transition kernel P and an integer $k \in \mathbb{N}$, we denote by P^k_π the k-step transition kernel.

Past works with single timescale algorithms commonly assume uniformly bounded mixing-times over the space of policies (Olshevsky and Gharesifard, 2023; Chen and Zhao, 2024; Bhandari et al., 2018; Wu et al., 2020).

Assumption 2.3. (Bounded mixing-time) Given $\epsilon > 0$, there exists a mixing-time $t_{mix}(\epsilon)$ such that

$$\forall \pi \in (\Delta^{\mathcal{A}_i})^{\mathcal{S}} \times (\Delta^{\mathcal{A}_{-i}})^{\mathcal{S}}, \quad t_{\pi,\epsilon} \leq t_{mix}(\epsilon).$$

[1]For finite state spaces, μ_π exists and is unique if the Markov chain induced by π is irreducible.

Using Assumptions 2.3 and 2.1 is proven to be enough to control the estimation error of the value function in single-timescale algorithms (Chen et al., 2021; Chen and Zhao, 2024).

In practice, while it is rational to assume that the considered policies have a mixing time, it can be excessive to assume a finite bound on their supremum over the entire space of policies as in Assumption 2.3. To illustrate, consider the MDP in Figure 1, we show in Appendix E that for the policies defined of Equation 2 it holds that $\lim_{\xi \to 1} t_{\pi,\epsilon} = +\infty$, which invalidates Assumption 2.3. This assumption is more stringent in game settings as fast mixing is related to the joint policy and each agent may act to induce a joint policy with arbitrarily slow mixing.

Overcoming these assumptions is particularly difficult in MARL due to the additional complexity of cooperation required for effective state space exploration. We next present an algorithm based on self-play that overcomes these challenges.

3. Algorithm and sample complexity

In this section, we describe our algorithm, motivate its design, and present our main theoretical result.

3.1. Algorithm

We now present our algorithm, Tsallis-smoothed Best-Response Dynamics with Value Iteration (`TBRVI`). Essentially, building upon the work of Chen et al. (2023), we combine principles of value iteration and best-response dynamics. Our key algorithmic contribution is the introduction of Tsallis entropy regularization for policy updates, instead of the standard softmax smoothing (Shannon entropy) used in past works.

It is known that continuous-time best-response dynamics converges asymptotically to a Nash equilibrium in zero-sum games with access to the opponent's actions and the best response function (Leslie et al., 2020; Hofbauer and Sorin, 2006). However, we assume here discrete-time dynamics and that agents lack knowledge of their opponent's actions and the best response. Consequently, we need to estimate the q-function for each agent. To achieve this, `TBRVI` uses minimax value iteration with TD-learning of Sutton and Barto (2018) to bypass the coordination of policy updates.

Algorithm statement. The pseudo-code of `TBRVI` is presented in Algorithm 1. It takes as input the number of episodes T, the length of an episode K, and a regularization parameter η.

In Algorithm 1, the value functions are initialized to zero. At the beginning of every episode, we reset the q-functions to zero and the policies to the uniform distribution. In line 1, the policies are updated according to a smoothed best response[2]. This smoothing is achieved through the parameter β_k and the use of Tsallis entropy:

$$\mathrm{Ts}(q^i_{t,k}(s)) = \underset{w \in \Delta^{|\mathcal{A}^i|}}{\arg\max} \left\langle w, q^i_{t,k}(s,.) \right\rangle + \frac{4}{\eta} \sum_i \sqrt{w_i}, \quad (4)$$

where $\eta > 0$ is a learning rate which we omit from the notation for readability. Next, in line 1, the action of player i is sampled from policy $\pi^i_{t,k+1}(.|s_k)$. Subsequently, in line 1, the MDP transitions to state s_{k+1} according to the transition dynamics. Line 1 depicts the q-functions' updates, utilizing a simple TD-learning scheme. Finally, in line 1, the value functions are updated similarly to value-iteration.

The inner loop of `TBRVI` faces two challenges. Firstly, both policies and q-functions are time-varying (lines 1 and 1). Traditionally, a double timescale is introduced to address this issue. In our case, we update the policies at a rate that is a multiplicative constant smaller than that of the q-functions.

Secondly, the convergence of TD-learning (line 1) requires the policies to explore all actions (Sutton and Barto, 2018). This is typically addressed using Assumption 2.1 with softmax smoothing. Our main algorithmic contribution lies

---

Algorithm 1 Tsallis-smoothed Best-Response Dynamics with Value Iteration

1: **Input:** Integers K and T, real number $\eta > 0$, matrices $v^i_0 = \mathbf{0} \in \mathbb{R}^{|\mathcal{S}|}$, $q^i_{t,0} = \mathbf{0} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}^i|}$ for all t, policies $\pi^i_{t,0}(a^i|s) = 1/|\mathcal{A}^i|$ for all (s, a^i) and t.
2: **for** $t = 0, 1, \cdots, T$ **do**
3:   **for** $k = 0, 1, \cdots, K-1$ **do**
4:     Update policies: $\forall s \in \mathcal{S}, i \in \{1,2\}$: $\pi^i_{t,k+1}(s) = \pi^i_{t,k}(s) + \beta_k(\mathrm{Ts}(q^i_{t,k}(s,.)) - \pi^i_{t,k}(s))$
5:     Sample actions: $a^i_k \sim \pi^i_{t,k+1}(\cdot|s_k)$
6:     Observe: $s_{k+1} \sim p(\cdot \mid s_k, a^i_k, a^{-i}_k)$
7:     Update values, for all (s, a^i):

$$q^i_{t,k+1}(s,a^i) = q^i_{t,k}(s,a^i) + \alpha_k \Big(\mathcal{R}^i(s_k, a^i_k, a^{-i}_k) + \gamma v^i_t(s_{k+1}) - q^i_{t,k}(s_k, a^i_k) \Big) \mathbb{1}_{\{s=s_k, a^i=a^i_k\}}$$

8:   **end for**
9:   $v^i_{t+1}(s) = \pi^i_{t,K}(s)^\top q^i_{t,K}(s,.)$ for all $s \in \mathcal{S}$ and set $s_0 = s_K$
10: **end for**
11: **Output:** $\pi^i_{T,K}$

---

[2] The original best response dynamics prescribes choosing $\pi^i_{k+1}(a|s) \in \arg\max_{\pi^i} q^i_{(\pi^i, \pi^{-i}_k)}(s,a)$.

in introducing Tsallis entropy in line 1. This alteration is crucial for proving finite-time convergence to approximate Nash equilibria, in contrast to the results of Chen et al. (2023). Tsallis entropy entails more exploration of suboptimal actions compared to softmax smoothed policies, and leads to considerably faster mixing, as demonstrated in later proofs.

In Algorithm 1, lines 1 and 1 represent a payoff-based version of minimax value iteration. The conventional minimax value iteration involves running the following (Bellman) step until convergence:

$$v_{t+1}^i(s) = \max_{\pi^i \in \Delta^{|\mathcal{A}^i|}} \min_{\pi^{-i} \in \Delta^{|\mathcal{A}^{-i}|}} \Big\{ (\pi^i)^\top \Big(\mathcal{R}^i(s, a^i, a^{-i}) + \gamma \mathbb{E}[v_t^i(s_1) | s_0 = s, a^i, a^{-i}] \Big) \pi^{-i} \Big\}.$$

Once a value $v_\star^i$ is reached, the agents solve the game corresponding to the above equation, with $v_\star^i$ instead of v_t. Leveraging our q-function estimation, we decentralize minimax value iteration and combine it with TD-learning to mitigate the challenges of payoff-based feedback.

3.2. Theoretical statement

Here, we show that `TBRVI` converges in polynomial time to an approximate Nash equilibrium. This result extends the literature by establishing convergence under a considerably weaker assumption.

Assumption 3.1. There *exists* a joint policy $\pi_b = (\pi_b^i, \pi_b^{-i})$ that induces a Markov chain which is irreducible and aperiodic.

The irreducibility here is strictly weaker than Assumption 2.1 as it only concerns a single policy. The aperiodicity ensures the existence of a mixing time for π_b (Lemma 5.7 in Khodadadian et al. (2022)), and is strictly weaker than Assumption 2.3. It is needed for the single-timescale objective, as discussed in Section 2.2. Remarkably, for algorithms that are payoff-based, convergent, and single-timescale, Assumption 3.1 is the weakest in the literature.

In Algorithm 1, we choose $\alpha_k = \alpha/(k+h)$, $\beta_k = c_{\alpha,\beta}\alpha_k$ where $c_{\alpha,\beta}$ is a constant parameter, and $\alpha, h > 0$ satisfy $\alpha/h < 1$. We now present our main result, which demonstrates the polynomial time convergence to an approximate Nash equilibrium. The theorem may appear ambiguous; however, we provide Corollary 3.3 afterwards and some explanations to clearly illustrate our contributions.

Theorem 3.2 (Nash Gap bound)**.** *Assume self-play and that the agents follow Algorithm 1. Under Assumption 3.1, and if $c_{\alpha,\beta} \le \frac{c_\eta \ell_\eta^3 (1-\gamma)^2}{6272\eta^3 |\mathcal{S}| A_{\max}^4}$, then for all $K \ge k_0$:*

$$\mathbb{E}[NG(\pi_{T,K}^i, \pi_{T,K}^{-i})] \le \frac{\hat{c}_1' |\mathcal{S}| A T \eta}{(1-\gamma)^3} \left(\frac{\gamma+1}{2} \right)^{T-1} + \frac{\hat{c}_2' |\mathcal{S}|^2 A_{\max}^2 \hat{L}_\eta z_K^2 \alpha^{1/2}}{\alpha_{k_0} c_{\alpha,\beta} (1-\gamma)^5} \frac{1}{\sqrt{K}} + \frac{\hat{c}_3' \sqrt{A_{\max}}}{\eta (1-\gamma)^2},$$

where $k_0 := \min\{k \ge 0 \mid k \ge z_k\}$, $z_K = \mathcal{O}(\log(K)\eta^{4r_b})$, $\hat{L}_\eta = \mathcal{O}(\eta^{4r_b})$, $\ell_\eta = \mathcal{O}(\eta^{-2})$, $c_\eta = \Omega(\ell_\eta^2)$, $r_b := \min\{k \ge 0 : P_{\pi_b}^k(s, s') > 0, \forall (s, s')\}$, and $\{\hat{c}_j'\}_{0 \le j \le 3}$ are numerical constants.

Let us discuss the three terms on the right-hand side. The first term is a bias that arises from playing according to minimax value iteration, and its inherent approximation error. The second term reflects the combined convergence error and variance of the inner loop and scales as $1/\sqrt{K}$. Lastly, the third term is a regularization bias from using Tsallis entropy to smooth the best response.

Corollary 3.3 (Sample Complexity)**.** *Fix $\epsilon > 0$. If Assumption 3.1 holds, then the `TBRVI` algorithm with $\eta = K^{1/(24r_b+28)}$ achieves $\mathbb{E}[NG(\pi_{T,K}^i, \pi_{T,K}^{-i})] \le \epsilon$ in $\tilde{\mathcal{O}}\left(1/\epsilon^{24r_b+28}\right)$ time. That is, `TBRVI` learns an ϵ-approximate Nash equilibrium in finite time.*

This sample complexity requires only Assumption 3.1. This marks the first polynomial convergence to Nash equilibria in literature that relaxes strong reachability, and it can be applied in the settings of Wei et al. (2017; 2021); Chen et al. (2021; 2023); Cai et al. (2024).

Moreover, we show that players using `TBRVI` converge in polynomial time to the best response of opponents who play a stationary policy. This means that `TBRVI` is rational.

Corollary 3.4 (Rationality)**.** *If player $-i$ follows a policy π^{-i}, then for player i following `TBRVI` , for any $\epsilon > 0$, $\max_{\hat{\pi}^i} v_{(\hat{\pi}^i, \pi^{-i})}^i(s_0) - v_{(\pi_{T,K}^i, \pi^{-i})}^i(s_0) \le \epsilon$ after $\tilde{\mathcal{O}}(1/\epsilon^{24r_b+28})$ time.*

Comparison with state-of-the-art. Several relevant works have addressed learning in similar settings. For instance, Wei et al. (2021) focused on average iterate convergence and proved a $\tilde{\mathcal{O}}\left(1/\epsilon^8\right)$ sample complexity. However, this requires Assumption 2.1 and double-timescale, and an average-iterate convergence is strictly weaker than a Nash gap. For last-iterate convergence, Chen et al. (2021) showed a $\tilde{\mathcal{O}}(1/\epsilon^{5.5})$ sample complexity using a double timescale algorithm, Assumptions 2.1 and 2.3, as well as communicating the entropy of the policy to the opponent. In recent work, Cai et al. (2024) proved a last-iterate guarantee with a rate of $\tilde{\mathcal{O}}\left(L^{1/\xi}/\epsilon^{9+\xi}\right)$ for any $\xi > 0$ under strong reachability. For general Markov games, Cai et al. (2024) showed a sample path convergence implying that "for all states that players visit often enough, players learn an approximate Nash policy", this is significantly weaker than even an average-iterate convergence.

4. Proof sketch

As discussed earlier, the main limitation of prior work is their inability to learn Nash equilibria without Assumptions 2.1 and 2.3. We attribute this to the common use of Shannon entropy smoothing in the literature. Specifically, Shannon entropy results in policies that converge too rapidly. Intuitively, the weight of non-optimal actions decays exponentially fast, hindering the reachability bounds necessary for the analysis. Moreover, this rapid decay directly impacts the mixing times. For instance, the best bounds in available for mixing times, provided in Chen et al. (2023, Lemma 4), show that mixing times grow inversely proportional to the exploration level of the policies.

In our proof, we first decompose the Nash gap into a sum of functions of the value functions, q-functions, and policies, as shown in Equation (5) below.

$$\mathrm{NG}(\pi^i_{T,K}, \pi^{-i}_{T,K}) \leq C_0 \Bigg(\mathcal{V}_\pi(T,K) + 2 \underbrace{\|v^i_T + v^{-i}_T\|_\infty}_{\mathcal{V}_1} + \sum_{i=1,2} \underbrace{\|v^i_T - v^i_*\|_\infty}_{\mathcal{V}_{2,i}} + \underbrace{\frac{8\sqrt{A_{\max}}}{\eta}}_{\text{bias}} \Bigg) \quad (5)$$

where C_0 is a constant, and $\mathcal{V}_\pi(\cdot)$ is a function that we present later.

This decomposition reveals several key components influencing the Nash gap. Our main argument is to establish drift inequalities[3] for each term on the right hand side, including $\mathcal{V}_1, \mathcal{V}_{2,1}, \mathcal{V}_{2,2}$, and $\mathcal{V}_\pi(\cdot)$. We leverage the proof techniques of Chen et al. (2023), replacing the softmax with Tsallis entropy and adjusting the affected arguments accordingly. This adaptation requires developing new properties for Tsallis entropy. For instance, our proof demonstrates that Tsallis entropy enables bounding several quantities, such as the mixing time, the convergence rate of q-functions, and the convergence rate of policies. Our bounds are polynomial in η, standing in stark contrast to the exponential dependence on η observed when using Shannon entropy under Assumption 3.1.

[3]We define drift inequalities as inequalities that show a negative drift in the iterate (similar to Lyapunov drift inequalities) plus additional terms coming from the coupling with other iterates.

4.1. Tsallis entropy

First, let's provide some background on Tsallis entropy before presenting our new properties.

Tsallis entropy, introduced in Tsallis (1988), is defined as $\mathcal{H}_\alpha(\pi) = \frac{1}{1-\alpha}(1 - \sum_i \pi_i^\alpha)$, where $\alpha \in [0,1]$. It includes Shannon entropy and log-barrier potential as special cases for $\alpha \to 1$ and $\alpha \to 0$, respectively (Agarwal et al., 2017; Abernethy et al., 2015). In this paper, we adopt Tsallis entropy with $\alpha = 1/2$, a common choice in the online learning literature (Zimmert and Seldin, 2021).

The policies induced by Tsallis entropy (see Equation (4)) have an explicit expression (Zimmert and Seldin, 2021):

$$\mathrm{Ts}(q^i_{t,k}(s))(a) = 4/\left(\eta\left(q^i_{t,k}(s,a) - x_t\right)\right)^2,$$

where $x_t \in \mathbb{R}$ is defined through the normalization constraint $\sum_a \mathrm{Ts}(.)(a) = 1$. This closed-form expression allows us to study the margins of the policies of Algorithm 1. Now, let's present our two contributions characterizing Tsallis entropy, which ensure crucial properties of the induced Markov chain. Our first result concerns the policies in Algorithm 1.

Lemma 4.1 (Margins). *It holds for all $t, k \geq 0$, $i \in \{1,2\}$ and (s, a^i, a^{-i}) that*

$$\pi^i_{t,k}(a^i|s) \geq \ell_\eta, \quad \text{where } \ell_\eta = 1/\left(\sqrt{A} + \frac{\eta}{2(1-\gamma)}\right)^2.$$

Proof. This follows directly from Lemma B.1 in the appendix. ☐

This property ensures that policies are lower bounded by a function of the η coefficient. This is advantageous for two reasons: first, according to Chen et al. (2023, Lemma 4.1 (2)), we can bound the mixing time of the resulting policies; second, we can utilize Zhang et al. (2023, Lemma 4) to lower bound the components of the stationary distribution of the policies. Consequently, the policies of Algorithm 1 enjoy crucial exploration properties using only Assumption 3.1.

Our second result is the Lipschitzness of Tsallis entropy with respect to the $\|.\|_2$ norm.

Lemma 4.2 (Lipschitzness). *For all $\boldsymbol{R}$ and $\boldsymbol{R}' \in \mathbb{R}^n$, we have:*

$$\|\mathrm{Ts}(\boldsymbol{R}) - \mathrm{Ts}(\boldsymbol{R}')\|_2 \leq 2\sqrt{2}\eta\, n \|\boldsymbol{R} - \boldsymbol{R}'\|_2.$$

We utilize this result in the next section to prove a crucial smoothness result, essential for ensuring the convergence of the policies. Lemma 4.2 is also relevant in game theory, reinforcement learning, and online learning. Similar properties for the softmax function, as seen in Gao and Pavel (2017), have been highly valuable in proving convergence to Nash equilibria in the mentioned fields.

4.2. Drift inequalities

The objective here is to establish drift inequalities for the terms $\mathcal{V}_1$, $(\mathcal{V}_{2,i})_{i=1,2}$, and $\mathcal{V}_\pi$. These inequalities are crucial for proving the convergence of `TBRVI` .

Value functions. This part pertains to the value function iterates in the Nash gap decomposition, namely $\mathcal{V}_1$ and $(\mathcal{V}_{2,i})_{i=1,2}$. We refer to them as value function iterates because they help analyze the convergence of different terms related to the value functions $((v_t^i)_{i=1,2})_{t\geq 0}$. We prove drift inequalities for each of these iterates, demonstrating a negative drift plus coupling terms. The negative drift primarily stems from the contractiveness of the Bellman operator, while the coupling terms follow from the dependence with other value function iterates as well as with the policy iterate. The challenge therefore lies in handling the coupling terms.

Policy drift. Here we address the term $\mathcal{V}_\pi$ in the Nash gap decomposition. Initially, given a fixed value function pair $v = (v^1, v^2)$ and a state $s \in \mathcal{S}$, we define:

$$V_{v,s}\left(\pi^i, \pi^{-i}\right) := \sum_{i=1,2} \max_{\hat{\pi}^i \in \Delta^{|\mathcal{A}^i|}} \Bigg\{ \left(\hat{\pi}^i - \pi^i\right)^\top \mathcal{T}^i(v^i)(s)\pi^{-i} + \frac{1}{\eta}\left(\mathcal{H}\left(\hat{\pi}^i\right) - \mathcal{H}\left(\pi^i\right)\right) \Bigg\},$$

where $\mathcal{T}^i(v)(s, a^i, a^{-i}) := \mathcal{R}^i(a, a^i, a^{-i}) + \gamma\mathbb{E}\left[v(S_1) \mid S_0 = s, A_0^i = a^i, A_0^{-i} = a^{-i}\right]$. We then define:

$$\mathcal{V}_\pi := \sum_s := V_{v,s}\left(\pi^i, \pi^{-i}\right), \tag{6}$$

The function $V_{v,s}$ serves as a regularized Nash gap for the matrix game with payoffs $\mathcal{T}^i(v)(s, \cdot, \cdot)$. It is an adaptation of the Lyapunov function provided in Hofbauer and Hopkins (2005) for best response dynamics, adjusted to accommodate the Markov games setting by the payoff matrices with $\mathcal{T}^i(v)$, and Tsallis entropy replacing Shannon entropy to align with our algorithmic choices.

Also a part of our decomposition, we prove a negative drift for $\mathcal{V}_\pi$ in Lemma C.5. The analysis of the policy updates follows the steps of (Hofbauer and Hopkins, 2005) by demonstrating that V_π is strongly convex and smooth due to the newly established properties of Tsallis entropy.

***Q*-functions.** Our analysis for $\mathcal{V}_1$, $\mathcal{V}_\pi$, and $(\mathcal{V}_{2,i})_{i=1,2}$ naturally involves a q-function estimation term. To prove the convergence of q-functions, we examine the following function

$$\mathcal{V}_q(t,k) = \sum_{i=1,2} \|q_{t,k}^i - \bar{q}_{t,k}^i\|_2^2, \tag{7}$$

where $\bar{q}_k^i(s) = \mathcal{T}^i(v^i)(s)\pi_k^{-i}(s)$. Analyzing $\mathcal{V}_q$ relies heavily on our novel use of Tsallis entropy.

In particular, as mentioned earlier, the policies must appear stationary for TD-learning to converge. We use the margin property for Tsallis entropy from Lemma 4.1 to lower bound the components of the policies of `TBRVI` , and Chen et al. (2023, Lemma 4.2) to derive a bound on the mixing time of the induced Markov chains. Then, a careful choice of the episode length implies that the Markov chain is close to its stationary distribution. Finally, using a conditioning argument, we can establish a drift inequality for the term $\mathcal{V}_q(t,k)$ related to q-functions.

Decoupling. The drift inequalities for $(\mathcal{V}_{2,i})_{i=1,2}$, $\mathcal{V}_1$, $\mathcal{V}_\pi$, and $\mathcal{V}_q$ are coupled (see Lemmas C.2, C.3,C.5, and C.9, respectively), meaning they are interdependent. Consequently, this step differs from typical Lyapunov analyses, where a central Lyapunov inequality is established and can be used to prove the final bound (see McMahan and Orabona (2014, Theorem 2.7.1)). Instead, we employ a special argument to decouple the inequalities and deduce the final bound. Similar to the analysis of Chen et al. (2023), we start from a crude bound on each drift function, and then iteratively and smartly apply the drift inequalities until tighter bounds are obtained.

Bias term. This term, denoted as $\left(8\sqrt{A_{\max}}/\eta\right)$, is the final term in our Nash gap decomposition. The goal is to choose a large η to remove this regularization-induced bias. Fortunately, the convergence we achieve for the q-functions is polynomial in η (see Theorem 3.2 and Lemma 4.1). Consequently, we can optimize the final bound to remove the bias and prove a polynomial convergence to Nash equilibria. Conversely, the softmax smoothing entails margins and mixing times that are exponential in η (see (Auletta et al., 2013)), making it impossible to remove the bias term for Shannon entropy-smoothed algorithms due to exponentially slow convergence. This highlights the importance of Tsallis entropy, which results in a bias term that can be tuned successfully.

5. Conclusion

In conclusion, our work addressed learning an approximate Nash equilibrium in zero-sum Markov games using payoff-based, decentralized, and single timescale algorithms under weak state reachability assumptions. We formally defined a relaxed reachability requirement in Assumption 3.1, and we proposed the `TBRVI` algorithm. This algorithm builds upon the work of Chen et al. (2023), combining essential principles such as best response dynamics of Hofbauer and Sorin (2006); Leslie et al. (2020) and value iteration of Shapley (1953). Our key algorithm contribution is smoothing the policy updates using Tsallis entropy, a new addition to the MARL literature.

Our analysis builds upon Markov chain results of Zhang et al. (2023) and incorporates insights on mixing times from Chen et al. (2023). We demonstrate how the policy update scheme induced by Tsallis entropy enables improved exploration of the state space and ensures an accurate estimation of the q-functions with TD-learning. Finally, we proved that the TBRVI algorithm learns an approximate Nash equilibrium in polynomial time, while eliminating the prohibitive strong reachability condition as well as the need for double timescales.

For future work, we believe that our sample complexity bounds are not tight and that they can be improved using concentration inequality type analyses, similar to Wei et al. (2021); Chen et al. (2021).

ACKNOWLEDGMENT

Reda Ouhamma and Maryam Kamgarpour are funded by an SNSF grant under NCCR Automation and an SNSF grant under #200020_207984/1

References

Jacob D Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting bandits with a new kind of smoothness. *Advances in Neural Information Processing Systems*, 28, 2015.

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pages 12–38. PMLR, 2017.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Shipra Agrawal and Randy Jia. Posterior sampling for reinforcement learning: worst-case regret bounds. *arXiv preprint arXiv:1705.07041*, 2017.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in neural information processing systems*, 19, 2006.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Vincenzo Auletta, Diodato Ferraioli, Francesco Pasquale, and Giuseppe Persiano. Mixing time and stationary expected social welfare of logit dynamics. *Theory of Computing Systems*, 53:3–40, 2013.

Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. In *International conference on machine learning*, pages 551–560. PMLR, 2020.

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

Michael Bowling and Manuela Veloso. Rational and convergent learning in stochastic games. In *International joint conference on artificial intelligence*, volume 17, pages 1021–1026. Citeseer, 2001.

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. Uncoupled and convergent learning in two-player zero-sum markov games with bandit feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34:27952–27964, 2021.

Xuyang Chen and Lin Zhao. Finite-time analysis of single-timescale actor-critic. *Advances in Neural Information Processing Systems*, 36, 2024.

Zaiwei Chen, Kaiqing Zhang, Eric Mazumdar, Asuman Ozdaglar, and Adam Wierman. A finite-sample analysis of payoff-based independent learning in zero-sum stochastic games. *arXiv preprint arXiv:2303.03100*, 2023.

Ziyi Chen, Shaocong Ma, and Yi Zhou. Sample efficient stochastic policy extragradient algorithm for zero-sum Markov game. In *International Conference on Learning Representations*, 2021.

Qiwen Cui, Kaiqing Zhang, and Simon Du. Breaking the curse of multiagents in a large state space: RL in Markov games with independent linear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2651–2652. PMLR, 2023.

Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *Communications of the ACM*, 52(2):89–97, 2009.

Constantinos Daskalakis, Dylan J Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33:5527–5540, 2020.

Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. he complexity of Markov equilibrium in stochastic games. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4180–4234. PMLR, 2023.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Claude-Nicolas Fiechter. Efficient reinforcement learning. In *Proceedings of the seventh annual conference on Computational learning theory*, pages 88–97, 1994.

Jerzy A Filar and Boleslaw Tolwinski. On the algorithm of Pollatschek and Avi-ltzhak. In *Stochastic Games And Related Topics: In Honor of Professor LS Shapley*, pages 59–70. Springer, 1991.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

Josef Hofbauer and Ed Hopkins. Learning in perturbed asymmetric games. *Games and Economic Behavior*, 52 (1):133–152, 2005.

Josef Hofbauer and Sylvain Sorin. Best response dynamics for continuous zero-sum games. *Discrete and Continuous Dynamical Systems Series B*, 6(1):215, 2006.

Alan J Hoffman and Richard M Karp. On nonterminating stochastic games. *Management Science*, 12(5):359–370, 1966.

Sajad Khodadadian, Thinh T Doan, Justin Romberg, and Siva Theja Maguluri. Finite sample analysis of two-time-scale natural actor-critic algorithm. *IEEE Transactions on Automatic Control*, 2022.

Harshat Kumar, Alec Koppel, and Alejandro Ribeiro. On the sample complexity of actor-critic method for reinforcement learning with function approximation. *Machine Learning*, pages 1–35, 2023.

David S Leslie, Steven Perkins, and Zibo Xu. Best-response dynamics in zero-sum stochastic games. *Journal of Economic Theory*, 189:105095, 2020.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. A sharp analysis of model-based reinforcement learning with self-play. In *International Conference on Machine Learning*, pages 7001–7010. PMLR, 2021.

H Brendan McMahan and Francesco Orabona. Unconstrained online linear learning in hilbert spaces: Minimax algorithms and normal approximations. In *Conference on Learning Theory*, pages 1020–1039. PMLR, 2014.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International conference on machine learning*, pages 6820–6829. PMLR, 2020.

Alex Olshevsky and Bahman Gharesifard. A small gain analysis of single timescale actor critic. *SIAM Journal on Control and Optimization*, 61(2):980–1007, 2023.

MA Pollatschek and B Avi-Itzhak. Algorithms for stochastic games with geometrical interpretation. *Management Science*, 15(7):399–415, 1969.

Muhammed Sayin, Kaiqing Zhang, David Leslie, Tamer Basar, and Asuman Ozdaglar. Decentralized Q-learning in zero-sum Markov games. *Advances in Neural Information Processing Systems*, 34:18320–18334, 2021.

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

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Constantino Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics*, 52:479–487, 1988.

J Van Der Wal. Discounted Markov games: Generalized policy iteration method. *Journal of Optimization Theory and Applications*, 25:125–138, 1978.

Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. *Advances in Neural Information Processing Systems*, 30, 2017.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive Markov games. In *Conference on learning theory*, pages 4259–4299. PMLR, 2021.

Yue Frank Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020.

Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Uncertainty in Artificial Intelligence*, pages 541–551. PMLR, 2020a.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33:4358–4369, 2020b.

Kaiqing Zhang, Sham Kakade, Tamer Basar, and Lin Yang. Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. *Advances in Neural Information Processing Systems*, 33:1166–1178, 2020.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar. Finite-sample analysis for decentralized batch multiagent reinforcement learning with networked agents. *IEEE Transactions on Automatic Control*, 66(12): 5925–5940, 2021.

Runyu Cathy Zhang, Zhaolin Ren, and Na Li. Gradient play in stochastic games: Stationary points and local geometry. *IFAC-PapersOnLine*, 55(30):73–78, 2022.

Yizhou Zhang, Guannan Qu, Pan Xu, Yiheng Lin, Zaiwei Chen, and Adam Wierman. Global convergence of localized policy iteration in networked multi-agent reinforcement learning. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(1):1–51, 2023.

Yulai Zhao, Yuandong Tian, Jason Lee, and Simon Du. Provably efficient policy optimization for two-player zero-sum Markov games. In *International Conference on Artificial Intelligence and Statistics*, pages 2736–2761. PMLR, 2022.

Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *The Journal of Machine Learning Research*, 22(1):1310–1358, 2021.

A. Notations

We dedicate this section to index all the notations used in this paper. Note that every notation is defined when it is introduced as well.

Table 1: Notations

T	:=	$K \times H$, total number of steps
A	:=	$\max\{\lvert\mathcal{A}^1\rvert, \lvert\mathcal{A}^2\rvert\}$.
v^i	:=	value function of agent $i \in \{1,2\}$, $\in \mathbb{R}^{\lvert\mathcal{S}\rvert}$
$\mathcal{R}^i(a, a^i, a^{-i})$	:=	Reward matrix for player $i (i \in \{1,2\})$
$\mathcal{T}^i(v)(s, a^i, a^{-i})$	:=	$\mathcal{R}^i(a, a^i, a^{-i}) + \gamma \mathbb{E}\left[v(s_1) \mid s_0 = s, a_0^i = a^i, a_0^{-i} = a^{-i}\right]$
$val^i(X)$	:=	$\max_{\pi^i \in \Delta^{\lvert\mathcal{A}^i\rvert}} \min_{\pi^{-i} \in \Delta^{\lvert\mathcal{A}^{-i}\rvert}} \{(\pi^i)^\top X \pi^{-i}\}$
	=	$\min_{\pi^{-i} \in \Delta^{\lvert\mathcal{A}^{-i}\rvert}} \max_{\pi^i \in \Delta^{\lvert\mathcal{A}^i\rvert}} \{(\pi^i)^\top X \pi^{-i}\}$
$\mathcal{B}^i(v)(s)$	:=	$val^i(\mathcal{T}^i(v)(s))$, the minimax Bellman operator
v^i_*	:=	The unique fixed-point of $\mathcal{B}^i(v^i)(s)$. Note that $v^i_* + v^{-i}_* = 0$.
$\mathcal{H}(w)$	:=	$4 \sum_{i \in [n]} \sqrt{w_i}$ for all $w \in \mathbb{R}^n$, Tsallis-entropy
$V_X(\pi^i, \pi^{-i})$	:=	$\sum_{i=1,2} \max_{\hat{\pi}^i \in \Delta^{\lvert\mathcal{A}^i\rvert}} \left\{ \left(\hat{\pi}^i - \pi^i\right)^\top X_i \pi^{-i} + \frac{1}{\eta}\mathcal{H}\left(\hat{\pi}^i\right) - \frac{1}{\eta}\mathcal{H}\left(\pi^i\right) \right\}$
$\mathcal{V}_q(t,k)$	:=	$\sum_{i=1,2} \lVert q^i_{t,k} - \bar{q}^i_{t,k} \rVert_2^2$
$v^i_{*,\pi^{-i}}(s)$	:=	$\max_{\hat{\pi}^i} v^i_{\hat{\pi}^i, \pi^{-i}}(s)$
$v^i_{\pi^i,*}$	:=	$\min_{\hat{\pi}^{-i}} v^i_{\pi^i, \hat{\pi}^{-i}}(s)$
$v^{-i}_{\pi^{-i},*}(s)$	:=	$\min_{\hat{\pi}^i} v^{-i}_{\pi^{-i}, \hat{\pi}^i}(s)$
$v^{-i}_{*,\pi^i}(s)$	:=	$\max_{\hat{\pi}^{-i}} v^{-i}_{\hat{\pi}^{-i}, \hat{\pi}^i}(s)$
Π_δ	:=	$\{(\pi^i, \pi^{-i}) \mid \forall i, \min_{s,a^i} \pi^i(a^i \mid s) > \delta_i\}$, where $\delta_i \in (0,1)$.

B. Supporting lemmas

In this section, we present technical results that are crucial for our proof. Particularly, we provide new results for the Tsallis entropy smoothing.

B.1. Tsallis entropy

Recall Tsallis-entropy in an n-dimensional space: $\mathcal{H}_\alpha(\pi) = \frac{1}{1-\alpha}(1 - \sum_i \pi_i^\alpha)$. In this paper, we consider Tsallis entropy with $\alpha = 1/2$, which can be equivalently written (Zimmert and Seldin, 2021): $\mathcal{H}(w) := 4\sum_{i \in [n]} \sqrt{w_i}$. In algorithm 1, we use this entropy as a regularization to the policy update to define the Tsallis smoothing:

$$\mathrm{Ts}(\boldsymbol{R}) = \underset{w \in \Delta^n}{\arg\max}\, \langle w, \boldsymbol{R} \rangle + \frac{1}{\eta}\mathcal{H}(w), \qquad \text{for all } \boldsymbol{R} \in \mathbb{R}^n$$

where η is a positive scalar. (Zimmert and Seldin, 2021) provides a closed-form expression for the Tsallis smoothing:

$$\mathrm{Ts}(\boldsymbol{R})_i = 4/\left(\eta\left(\boldsymbol{R}_i - x\right)\right)^2,$$

where $x \in \mathbb{R}$ is defined implicitly through the normalization constraint $\sum_i 4\left(\eta\left(\boldsymbol{R}_i - x\right)\right)^{-2} = 1$.

Lemma B.1 (Normalization factor)**.** *Let $\boldsymbol{R} \in \mathbb{R}^n$, and let $x \in \mathbb{R}$ be the normalization factor of $\mathrm{Ts}(\boldsymbol{R})$, then we can show that*

$$2/\eta \leq x - \max_i \boldsymbol{R}_i \leq 2\sqrt{n}/\eta$$

Proof. Let $i^\star = \arg\max \boldsymbol{R}_i$, then we have $w_{i^\star} \in [1/n, 1]$ (see remark below) and

$$x = \boldsymbol{R}_{i^\star} - 2\eta^{-1}\,\mathrm{Ts}(\boldsymbol{R})_{i^\star}^{-\frac{1}{2}}$$

Since the function $y :\to 4(\eta(\boldsymbol{R}_i - y))^{-2}$ is monotonous in the domain $I = [\boldsymbol{R}_{i^\star} + 2/\eta, \boldsymbol{R}_{i^\star} + 2\sqrt{n}/\eta]$, there exists a unique value in I, such that $\sum_i 4\left(\eta_t\left(\boldsymbol{R}_i - x\right)\right)^{-2} = 1$. □

Remark B.2. Note that there are $n+1$ candidates for the normalization factor, such that the smallest one is smaller than $\min_i \boldsymbol{R}_i$ and the second smallest one is smaller than the second smallest $(\boldsymbol{R})_j$, and so on until the n^{th} one, the $(n+1)^{\text{th}}$ candidate is larger than $\max_i \boldsymbol{R}_i$. The $(n+1)^{\text{th}}$ candidate is the normalization factor because it maximizes the objective in equation (4).

Lemma B.3 (Lipschitzness)**.** *For all $\boldsymbol{R}$ and $\boldsymbol{R}' \in \mathbb{R}^n$, we have:*

$$\|\,\mathrm{Ts}(\boldsymbol{R}) - \mathrm{Ts}(\boldsymbol{R}')\|_2 \leq 2\sqrt{2}\eta\, n\|\boldsymbol{R} - \boldsymbol{R}'\|_2$$

This means that Tsallis-smoothing is L_η-Lipschitz with respect to $\|.\|_2$.

Proof. Let $\boldsymbol{R}$ and $\boldsymbol{R}'$ be two vectors in $\mathbb{R}^n$, and consider $\boldsymbol{R}^0, \ldots, \boldsymbol{R}^n \in \mathbb{R}^n$ such that $\boldsymbol{R}^0 = \boldsymbol{R}$, and for $j \in \{1, \ldots, n\}$, we define $\boldsymbol{R}^j := (\boldsymbol{R}'_1, \ldots, \boldsymbol{R}'_j, \boldsymbol{R}_{j+1}, \ldots, \boldsymbol{R}_n)$, we also denote $x_0, \ldots, x_n$ their respective Tsallis-normalization factor.

Let $j \in \{1, \ldots, n\}$ we have:

1) If $\boldsymbol{R}'_j \geq \boldsymbol{R}_j$: then $x_j > x_{j-1}$, this is because the function $y :\to \sum_i 4(\eta(\boldsymbol{R}_i - y))^{-2}$ is decreasing in the domain $[x_{j-1}, \max_i \boldsymbol{R}^j_i + 2\sqrt{K}/\eta]$ and it is larger than 1 in x_{j-1}.

The latter entails that for all $l \neq j, \mathrm{Ts}(\boldsymbol{R}^j)_l \leq \mathrm{Ts}(\boldsymbol{R}^{j-1})_l$, which implies that $\mathrm{Ts}(\boldsymbol{R}^{j-1})_j \geq \mathrm{Ts}(\boldsymbol{R}^{j-1})_j$ to keep the normalization condition. Then

$$\begin{aligned}\mathrm{Ts}(\boldsymbol{R}^{j-1})_j \geq \mathrm{Ts}(\boldsymbol{R}^{j-1})_j \implies& \frac{1}{\left(x_j - \boldsymbol{R}'_j\right)^2} \geq \frac{1}{\left(x_{j-1} - \boldsymbol{R}_j\right)^2},\\ \implies& x_j - x_{j-1} \leq \boldsymbol{R}'_j - \boldsymbol{R}_j,\end{aligned}$$

the last line follows because $x_j \geq \boldsymbol{R}'_j$ and $x_{j-1} \geq \boldsymbol{R}_j$ thanks to Lemma. B.1.

2) If $\boldsymbol{R}'_j \leq \boldsymbol{R}_j$: Similarly, we can prove that this entails that $x_j - x_{j-1} \geq \boldsymbol{R}'_j - \boldsymbol{R}_j$.

In 1) and 2) we have shown that for all $j \in \{1, \ldots, n\}$ we have: $x_j - x_{j-1} \leq \boldsymbol{R}'_j - \boldsymbol{R}_j$. Therefore, we can show

$$\sum_{l=1}^{n}(\boldsymbol{R}'_l - \boldsymbol{R}_l)\mathbb{1}_{\boldsymbol{R}'_l \leq \boldsymbol{R}_l} \leq x_n - x_0 \leq \sum_{l=1}^{n}(\boldsymbol{R}'_l - \boldsymbol{R}_l)\mathbb{1}_{\boldsymbol{R}'_l \geq \boldsymbol{R}_l}$$

Now that we have studied the variation of the normalization factor we are able to analyze the variation of the Tsallis-induced weights from a variation of the $\boldsymbol{R}$ vector. Let $j \in \{1, \ldots, n\}$, we have:

$$\begin{aligned}
\mathrm{Ts}(\boldsymbol{R})_j - \mathrm{Ts}(\boldsymbol{R}')_j &= \frac{4}{(\eta(\boldsymbol{R}_j - x_0))^2} - \frac{4}{(\eta(\boldsymbol{R}'_j - x_n))^2} \\
&= \frac{4}{\eta^2}\left(\frac{(\boldsymbol{R}'_j - x_n)^2 - (\boldsymbol{R}_j - x_0)^2}{(\boldsymbol{R}'_j - x_n)^2(\boldsymbol{R}_j - x_0)^2}\right) \\
&= \frac{4}{\eta^2}\left(\frac{(\boldsymbol{R}'_j - \boldsymbol{R}_j + x_0 - x_n)(\boldsymbol{R}_j - x_0 + \boldsymbol{R}'_j - x_n)}{(\boldsymbol{R}'_j - x_n)^2(\boldsymbol{R}_j - x_0)^2}\right) \\
&= \frac{4}{\eta^2}\frac{\boldsymbol{R}'_j - \boldsymbol{R}_j + x_0 - x_n}{(\boldsymbol{R}'_j - x_n)(\boldsymbol{R}_j - x_0)}\left(\frac{1}{\boldsymbol{R}'_j - x_n} + \frac{1}{\boldsymbol{R}_j - x_0}\right) \\
&\le 8\eta|\boldsymbol{R}'_j - \boldsymbol{R}_j + x_0 - x_n| \\
&\le 8\eta\|\boldsymbol{R}' - \boldsymbol{R}\|_1
\end{aligned}$$

where the penultimate inequality follows since for all j, $\frac{1}{\eta(\boldsymbol{R}'_j - x_n)} \le 1$ and $\frac{1}{\eta(\boldsymbol{R}_j - x_0)} \le 1$ because they are square-roots of probabilities.

Finally:

$$\|\,\mathrm{Ts}(\boldsymbol{R}) - \mathrm{Ts}(\boldsymbol{R}')\|_2 \le 2\eta\sqrt{2n}\|\boldsymbol{R}' - \boldsymbol{R}\|_1 \le 2\sqrt{2}\eta\, n\|\boldsymbol{R}' - \boldsymbol{R}\|_2$$

which concludes the proof. □

B.2. Markov chain tools

The following lemmas establish exploration properties under assumption 3.1. To present the results, we need additional notations. Under Assumption 3.1, there exists a joint policy π_b such that its induced Markov chain has a unique stationary distribution $\mu_b \in \Delta^{|\mathcal{S}|}$. The minimum component μ_b is denoted as $\mu_{b,\min}$, μ_b is positive thanks to Equation 1 and the irreducibility of the Markov chain induced by π_b. In addition, there exists $\rho_b \in (0, 1)$ such that $\max_{s\in\mathcal{S}} \left\|P^k_{\pi_b}(s, \cdot) - \mu_b(\cdot)\right\|_{\mathrm{TV}} \le 2\rho_b^k$ for all $k \ge 0$ (Levin and Peres, 2017), where P_{π_b} is the transition probability matrix of the Markov chain $\{S_k\}$ under π_b. In addition, thanks to lemma 4.1, it is enough to restrict our attention to policy classes of the form $\Pi_\delta := \{\pi = (\pi^i, \pi^{-i}) \mid \min_{s,a^i} \pi^i(a^i|s) > \delta_i, \min_{s,a^{-i}} \pi^{-i}(a^{-i}|s) > \delta_{-i}\}$, where $\delta_i, \delta_{-i} \in (0, 1)$.

Lemma B.4 ((Zhang et al., 2023), Lemma 4)**.** *Suppose that Assumption 3.1 is satisfied. Then we have the following results.*

1. *For any $\pi = (\pi^i, \pi^{-i}) \in \Pi_\delta$, the Markov chain $\{S_k\}$ induced by the joint policy π is irreducible and aperiodic, hence admits a unique stationary distribution $\mu_\pi \in \Delta^{|\mathcal{S}|}$.*
2. *Let $G : \mathbb{R}^{|\mathcal{S}|\mathcal{A}} \mapsto \mathbb{R}^{|\mathcal{S}|}$ be the mapping from a policy $\pi \in \Pi_\delta$ to the unique stationary distribution μ_π of the Markov chain $\{S_k\}$ induced by π. Then $G(\cdot)$ is Lipschitz continuous with respect to $\|\cdot\|_\infty$, with Lipschitz constant $\hat{L}_\delta := \frac{2\log(8|\mathcal{S}|/\rho_\delta)}{\log(1/\rho_\delta)}$, where $\rho_\delta = \rho_b^{(\delta_i\delta_{-i})^{r_b}\mu_{b,\min}}$ and $r_b := \min\{k \ge 0 \;:\; P^k_{\pi_b}(s, s') > 0,\ \forall\,(s, s')\}$.*
3. *$\mu_\delta := \inf_{\pi\in\Pi_\delta} \min_{s\in\mathcal{S}} \mu_\pi(s) > 0$.*

Lemma B.5 ((Chen et al., 2023), Lemma 4.2)**.** *Suppose that Assumption 3.1 is satisfied. Then, it holds that $\sup_{\pi\in\Pi_\delta} \max_{s\in\mathcal{S}} \|P^k_\pi(s, \cdot) - \pi_\pi(\cdot)\|_{TV} \le 2\rho_\delta^k$ for any $k \ge 0$, where $\rho_\delta = \rho_b^{(\delta_i\delta_{-i})^{r_b}\mu_{b,\min}}$ and $r_b := \min\{k \ge 0 \;:\; P^k_{\pi_b}(s, s') > 0,\ \forall\,(s, s')\}$. As a result, we have*

$$t(\pi, \lambda) := \sup_{\pi\in\Pi_\delta} t_{\pi,\lambda} \le \frac{t_{\pi_b,\lambda}}{(\delta_i\delta_{-i})^{r_b}\mu_{b,\min}}, \tag{8}$$

where we recall that $t_{\pi,\lambda}$ is the λ – mixing time of the Markov chain $\{S_k\}$ induced by π.

This enables us to see the explicit dependence of the mixing time on the margins δ_i, δ_{-i} and the mixing time of the benchmark exploration policy π_b. Note that, as the margins δ_i, δ_{-i} approach zero, the uniform mixing time in Lemma B.5 goes to infinity. This bound is generally non-vacuous, as demonstrated by a simple MDP example constructed in (Chen et al., 2023).

Given Lemma B.5, we have fast mixing for all policies in Π_δ if the margins δ_i, δ_{-i} are large, the mixing time of π_b is small, and the stationary distribution is balanced (large $\pi_{b,\min}$).

Notation For simplicity, when $\Pi_\delta = \Pi_{\ell_\eta}$, we denote $\rho_\eta := \rho_\delta$, $\mu_\eta := \mu_\delta$, and $\hat{L}_\eta := \hat{L}_\delta$. We also define $c_\eta := \mu_\eta \ell_\eta$.

B.3. Miscellaneous

Here we present two results on the value function estimates and the policies of algorithm 1.

Lemma B.6 (Bounded value function)**.** *It holds for all $t, k \geq 0$ and $i \in \{1, 2\}$ that*

1. $\|v_t^i\|_\infty \leq \frac{1}{1-\gamma}$,
2. $\|q_{t,k}^i\|_\infty \leq \frac{1}{1-\gamma}$.

Proof. We proceed by two induction arguments, first, let's show the second part of the lemma by induction over t. By our initialization it holds that $q_{0,k}^i = 0$ for all k. Assume that $\|q_{t,k}^i\|_\infty \leq \frac{1}{1-\gamma}$ and $\|v_t^i\|_\infty \leq \frac{1}{1-\gamma}$. Then we have for all (s, a^i) that

$$\begin{aligned}
|q_{t,k+1}^i(s,a^i)| &= |q_{t,k}^i(s,a^i) + \alpha_k \mathbb{1}_{\{(s,a^i)=(s_k,s_k^i)\}}(\mathcal{R}^i(s_k, a_k^i, a_k^{-i}) + \gamma v_t^i(S_{k+1}) - q_{t,k}^i(s_k, a_k^i))| \\
&\leq \alpha_k \mathbb{1}_{\{(s,a^i)=(s_k,a_k^i)\}}|\mathcal{R}^i(s_k, a_k^i, a_k^{-i}) + \gamma v_t^i(S_{k+1})| \\
&\quad + (1 - \alpha_k \mathbb{1}_{\{(s,a^i)=(s_k,a_k^i)\}})|q_{t,k}^i(s,a^i)| \\
&\leq \alpha_k \mathbb{1}_{\{(s,a^i)=(s_k,a_k^i)\}} \left(1 + \frac{\gamma}{1-\gamma}\right) + (1 - \alpha_k \mathbb{1}_{\{(s,a^i)=(s_k,a_k^i)\}})\frac{1}{1-\gamma} \\
&= \frac{1}{1-\gamma},
\end{aligned}$$

this finishes the first induction. We showed that if $\|v_t^i\|_\infty \leq \frac{1}{1-\gamma}$ then $\|q_{t,k}^i\|_\infty \leq \frac{1}{1-\gamma}$ for all $k \geq 0$.

Second, we proceed by induction over k, assume that $\|v_t^i\|_\infty \leq \frac{1}{1-\gamma}$ then for all $s \in \mathcal{S}$

$$|v_{t+1}^i(s)| = \left|\sum_{a^i \in \mathcal{A}^i} \pi_{t,K}^i(a^i|s) q_{t,K}^i(s,a^i)\right| \leq \sum_{a^i \in \mathcal{A}^i} \pi_{t,K}^i(a^i|s) \|q_{t,K}^i\|_\infty \leq \frac{1}{1-\gamma},$$

the last line follows because the induction hypothesis entails that $\|q_{t,K}^i\|_\infty \leq \frac{1}{1-\gamma}$ (by the first induction). This concludes the second induction, we now proved that $\|v_t^i\|_\infty \leq \frac{1}{1-\gamma}$ for all $t \geq 0$.

Combining the two arguments allows us to prove that $\|q_{t,k}^i\|_\infty \leq \frac{1}{1-\gamma}$ for all $t \geq 0$. This concludes the proof for the two statements. □

Lemma B.7 (Margins (Lemma 4.1))**.** *It holds for all players $i \in \{1, 2\}$, times $t, k \geq 0$ and state-action pairs (s, a^i, a^{-i}) that*

$$\pi_{t,k}^i(a^i|s) \geq \ell_\eta := 1/\left(\sqrt{A} + \frac{\eta}{2(1-\gamma)}\right)^2,$$

we call this the margin property of the deployed policies.

Proof. We recall the policy-update equation:

$$\pi_{t+1}^i = \pi_t^i + \beta_k \left(\mathrm{Ts}\left(q_k^i\right) - \pi_t^i\right).$$

By lemma B.1 we have that $2/\eta \leq x - \max_i \boldsymbol{R}_i \leq 2\sqrt{n}/\eta$, and by LemmaB.6 we know that $\|q_{t,k}^i\|_\infty \leq \frac{1}{1-\gamma}$ therefore:

$$\forall s \in \mathcal{S}, a \in \mathcal{A} : \mathrm{Ts}\left(q_k^i\right)(a) \geq \frac{1}{(\sqrt{\mathcal{A}} + \frac{\eta}{2(1-\gamma)})^2}.$$

Next, it is easy to show by induction that $\pi^i_{t,k}(a|s) \geq \ell_\eta$ for all $s \in \mathcal{S}, a \in \mathcal{A}$. Indeed, $\pi^i_{t,0}(a|s) = 1/\mathcal{A} \geq \ell_\eta$, and for a given $k \geq 0$ if $\pi^i_{t,k}(a|s) \geq \ell_\eta$ then:

$$\begin{aligned}\pi^i_{t,k+1}(a|s) &= \pi^i_t(a|s)\ell_\eta + \beta_k \left(\mathrm{Ts}\left(q^i_k\right)_a - \pi^i_t(a|s)\right) \\ &= (1-\beta_k)\underbrace{\pi^i_t(a|s)}_{\geq \ell_\eta} + \beta_k \underbrace{\mathrm{Ts}\left(q^i_k\right)_a}_{\geq \ell_\eta} \\ &\geq (1-\beta_k)\ell_\eta + \beta_k \ell_\eta = \ell_\eta\end{aligned}$$

which concludes the proof. □

C. Sample complexity analysis

This section is dedicated to the proof of theorem 3.2. We begin the analysis by restating the Nash gap decomposition (*cf.* Equation 5):

$$\mathrm{NG}(\pi^i_{T,K}, \pi^{-i}_{T,K}) \leq C_0 \left(2\|v^i_T + v^{-i}_T\|_\infty + \sum_{i=1,2} \|v^i_T - v^i_*\|_\infty + \mathcal{V}_\pi(T,K) + \frac{8}{\eta}\sqrt{\mathcal{A}}\right),$$

where C_0 is a constant, and $\mathcal{V}_\pi(\cdot)$ is function of the policies at times T and K (*cf.* Equation 6).

We start by stating a requirement for choosing the parameters α_k and β_k. For simplicity of notation, given $k_1 \leq k_2$, we denote $\beta_{k_1,k_2} = \sum_{k=k_1}^{k_2} \beta_k$ and $\alpha_{k_1,k_2} = \sum_{k=k_1}^{k_2} \alpha_k$. For any $k \geq 0$, define $z_k = t(\ell_\eta, \beta_k)$ where ℓ_η is defined in lemma 4.1. Observe that when $\beta_k = \mathcal{O}(1/k)$, then $z_k = \mathcal{O}(\log(k))$ due to the geometric mixing property in Lemma B.5.

Condition 1. It holds that $\alpha_{k-z_k,k-1} \leq 1/4$ for all $k \geq z_k$ and $c_{\alpha,\beta} \leq \frac{c_\eta \ell_\eta^3 (1-\gamma)^2}{6272\eta^3|\mathcal{S}|\mathcal{A}^4}$. When using diminishing stepsizes $\alpha_k = \frac{\alpha}{k+h}$ and $\beta_k = \frac{\beta}{k+h}$, we additionally require $\beta > 2$.

The above condition is necessary for our proof of convergence in Theorem 3.2. Condition 1 can be explicitly satisfied as $z_k = \mathcal{O}(\log(1/k))$ with the chosen parameters. Finally, the parameter k_0 that appears in Theorem 3.2 is defined to be $\min\{k \geq 0 \mid k \geq z_k\}$.

C.1. Outer-loop

Lemma C.1 (Nash Gap in terms of value iterates)**.** *It holds for all $t \geq 0$ and $i = 1,2$ that*

$$\left\|v^i_{*,\pi^{-i}_{t,K}} - v^i_{\pi^i_{t,K},\pi^{-i}_{t,K}}\right\|_\infty \leq \frac{2}{1-\gamma}\left(2\|v^i_t + v^{-i}_t\|_\infty + 2\|v^i_t - v^i_*\|_\infty + \max_s V_{v_t,s}(\pi^i_{t,K}(s), \pi^{-i}_{t,K}(s)) + \frac{8}{\eta}\sqrt{\mathcal{A}}\right).$$

Proof. For any $t \geq 0$, $s \in \mathcal{S}$, and $i \in \{1,2\}$, we have

$$\begin{aligned}\left|v^i_{*,\pi^{-i}_{t,K}}(s) - v^i_{\pi^i_{t,K},\pi^{-i}_{t,K}}(s)\right| &= v^i_{*,\pi^{-i}_{t,K}}(s) - v^i_{\pi^i_{t,K},\pi^{-i}_{t,K}}(s) \\ &\leq v^i_{*,\pi^{-i}_{t,K}}(s) - v^i_{\pi^i_{t,K},*}(s) \\ &= -v^{-i}_{\pi^{-i}_{t,K},*}(s) - v^i_{\pi^i_{t,K},*}(s) \\ &= v^i_*(s) - v^{-i}_{\pi^{-i}_{t,K},*}(s) + v^{-i}_*(s) - v^i_{\pi^i_{t,K},*}(s) \\ &\leq \|v^{-i}_* - v^{-i}_{\pi^{-i}_{t,K},*}\|_\infty + \|v^i_* - v^i_{\pi^i_{t,K},*}\|_\infty. \end{aligned} \tag{9}$$

We now bound the two terms on the r.h.s above. For the first term, note that for any $s \in \mathcal{S}$ and $t \geq 0$, we have

$$
\begin{aligned}
0 \leq v_*^{-i}(s) - v_{\pi_{t,K}^{-i},*}^{-i}(s) =& v_{*,\pi_{t,K}^{-i}}^{i}(s) - v_*^i(s) \\
=& \max_{\pi^i} (\pi^i)^\top \mathcal{T}^i(v_{*,\pi_{t,K}^{-i}}^{i})(s)\pi_{t,K}^{-i}(s) - \max_{\pi^i}\min_{\pi^{-i}} (\pi^i)^\top \mathcal{T}^i(v_*^i)(s)\pi^{-i} \\
=& |\max_{\pi^i} (\pi^i)^\top \mathcal{T}^i(v_{*,\pi_{t,K}^{-i}}^{i})(s)\pi_{t,K}^{-i}(s) - \max_{\pi^i} (\pi^i)^\top \mathcal{T}^i(v_*^i)(s)\pi_{t,K}^{-i}(s)| \\
&+ |\max_{\pi^i} (\pi^i)^\top \mathcal{T}^i(v_*^i)(s)\pi_{t,K}^{-i}(s) - \max_{\pi^i}\min_{\pi^{-i}} (\pi^i)^\top \mathcal{T}^i(v_*^i)(s)\pi^{-i}| \\
\leq& \max_{\pi^i} |(\pi^i)^\top (\mathcal{T}^i(v_{*,\pi_{t,K}^{-i}}^{i})(s) - \mathcal{T}^i(v_*^i)(s))\pi_{t,K}^{-i}(s)| \\
&+ |\max_{\pi^i} (\pi^i)^\top \mathcal{T}^i(v_*^i)(s)\pi_{t,K}^{-i}(s) - \max_{\pi^i}\min_{\pi^{-i}} (\pi^i)^\top \mathcal{T}^i(v_t^i)(s)\pi^{-i}| \\
&+ |\max_{\pi^i}\min_{\pi^{-i}} (\pi^i)^\top \mathcal{T}^i(v_t^i)(s)\pi^{-i} - \max_{\pi^i}\min_{\pi^{-i}} (\pi^i)^\top \mathcal{T}^i(v_*^i)(s)\pi^{-i}| \\
\leq& \underbrace{\max_{\pi^i} |(\pi^i)^\top (\mathcal{T}^i(v_{*,\pi_{t,K}^{-i}}^{i})(s) - \mathcal{T}^i(v_*^i)(s))\pi_{t,K}^{-i}(s)|}_{\hat{E}_1} \\
&+ \underbrace{|\max_{\pi^i} (\pi^i)^\top \mathcal{T}^i(v_*^i)(s)\pi_{t,K}^{-i}(s) - \max_{\pi^i} (\pi^i)^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s)|}_{\hat{E}_2} \\
&+ \underbrace{\max_{\pi^i} (\pi^i)^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s) - \max_{\pi^i}\min_{\pi^{-i}} (\pi^i)^\top \mathcal{T}^i(v_t^i)(s)\pi^{-i}}_{\hat{E}_3} \\
&+ \underbrace{|\max_{\pi^i}\min_{\pi^{-i}} (\pi^i)^\top \mathcal{T}^i(v_t^i)(s)\pi^{-i} - \max_{\pi^i}\min_{\pi^{-i}} (\pi^i)^\top \mathcal{T}^i(v_*^i)(s)\pi^{-i}|}_{\hat{E}_4}. \qquad (10)
\end{aligned}
$$

We next bound the terms $\{\hat{E}_j\}_{1\leq j\leq 4}$. For any $v_1^i, v_2^i \in \mathbb{R}^{|\mathcal{S}|}$, we have for any (s, a^i, a^{-i}) that

$$
\begin{aligned}
|\mathcal{T}^i(v_1^i)(s,a^i,a^{-i}) - \mathcal{T}^i(v_2^i)(s,a^i,a^{-i})| &= \gamma |\mathbb{E}[v_1^i(s_1) - v_2^i(s_1) \mid s_0 = s, a_0^i = a^i, a_0^{-i} = a^{-i}]| \\
&\leq \gamma \|v_1^i - v_2^i\|_\infty,
\end{aligned}
$$

which implies that $\|\mathcal{T}^i(v_1^i) - \mathcal{T}^i(v_2^i)\|_\infty \leq \gamma \|v_1^i - v_2^i\|_\infty$. As a result, we have

$$
\begin{aligned}
&\hat{E}_1 \leq \|\mathcal{T}^i(v_{*,\pi_{t,K}^{-i}}^{i}) - \mathcal{T}^i(v_*^i)\|_\infty \leq \gamma \|v_{*,\pi_{t,K}^{-i}}^{i} - v_*^i\|_\infty, \\
&\hat{E}_2 \leq \|\mathcal{T}^i(v_t^i) - \mathcal{T}^i(v_*^i)\|_\infty \leq \gamma \|v_t^i - v_*^i\|_\infty, \\
&\hat{E}_4 \leq \|\mathcal{T}^i(v_t^i) - \mathcal{T}^i(v_*^i)\|_\infty \leq \gamma \|v_t^i - v_*^i\|_\infty.
\end{aligned}
$$

Finally, to bound $\hat{E}_3$, observe that

$$
\begin{aligned}
\hat{E}_3 \leq& \left|\max_{\pi^i}(\pi^i)^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s) - \min_{\pi^{-i}} \pi_{t,K}^i(s)\mathcal{T}^i(v_t^i)(s)\pi^{-i}\right| \\
\leq& \left|\max_{\pi^{-i}}(\pi^{-i})^\top \mathcal{T}^{-i}(v_t^{-i})(s)\pi_{t,K}^i(s) + \min_{\pi^{-i}}(\pi^{-i})^\top \mathcal{T}^i(v_t^i)(s)^\top \pi_{t,K}^i(s)\right| \\
&+ \left|\sum_{i=1,2}\max_{\pi^i}(\pi^i)^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s)\right| \\
\leq& \left|\max_{\pi^{-i}}(\pi^{-i})^\top \mathcal{T}^{-i}(v_t^{-i})(s)\pi_{t,K}^i(s) - \max_{\pi^{-i}}(\pi^{-i})^\top [-\mathcal{T}^i(v_t^i)(s)]^\top \pi_{t,K}^i(s)\right| \\
&+ \left|\sum_{i=1,2}\max_{\pi^i}(\pi^i)^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s)\right| \\
\leq& \max_{\pi^{-i}}\left|(\pi^{-i})^\top(\mathcal{T}^{-i}(v_t^{-i})(s) + [\mathcal{T}^i(v_t^i)(s)]^\top)\pi_{t,K}^i(s)\right| + \left|\sum_{i=1,2}\max_{\pi^i}(\pi^i)^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s)\right| \\
\leq& \max_{a^i,a^{-i}}\left|\mathcal{T}^i(v_t^i)(s,a^i,a^{-i}) + \mathcal{T}^{-i}(v_t^{-i})(s,a^i,a^{-i})\right| + \left|\sum_{i=1,2}\max_{\pi^i}(\pi^i)^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s)\right| \\
\leq& \gamma\|v_t^i + v_t^{-i}\|_\infty + \left|\sum_{i=1,2}\max_{\pi^i}(\pi^i)^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s)\right|,
\end{aligned}
$$

where the last line follows because:

$$
\begin{aligned}
|\mathcal{T}^i(v_t^i)(s,a^i,a^{-i}) + \mathcal{T}^{-i}(v_t^{-i})(s,a^i,a^{-i})| =& \gamma\left|\mathbb{E}[v_t^i(S_1) + v_t^i(S_1) \mid s_0 = s, a_0^i = a^i, a_0^{-i} = a^{-i}]\right| \\
\leq& \gamma\|v_t^i + v_t^{-i}\|_\infty.
\end{aligned}
$$

In addition, we have

$$
\begin{aligned}
\left|\sum_{i=1,2}\max_{\pi^i}(\pi^i)^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s)\right| =& \left|\sum_{i=1,2}\max_{\pi^i}\left\{(\pi^i - \pi_{t,K}^i(s))^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s)\right\}\right| \\
&+ \left|\sum_{i=1,2}(\pi_{t,K}^i(s))^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s)\right| \\
\leq& \sum_{i=1,2}\max_{\pi^i}\left\{(\pi^i - \pi_{t,K}^i(s))^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s)\right. \\
&\left.+ \frac{1}{\eta}\mathcal{H}(\pi^i) - \frac{1}{\eta}\mathcal{H}(\pi_{t,K}^i(s))\right\} \\
&+ \left|\sum_{i=1,2}(\pi_{t,K}^i(s))^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s)\right| + \frac{8}{\eta}\sqrt{\mathcal{A}},
\end{aligned}
$$

then

$$\begin{aligned}\left|\sum_{i=1,2}\max_{\pi^i}(\pi^i)^\top\mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s)\right| &\leq V_{v_t,s}(\pi_{t,K}^i(s),\pi_{t,K}^{-i}(s))+\frac{8}{\eta}\sqrt{\mathcal{A}}\\&\quad+\left|\sum_{i=1,2}(\pi_{t,K}^i(s))^\top\mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s)\right|\\&\leq V_{v_t,s}(\pi_{t,K}^i(s),\pi_{t,K}^{-i}(s))+\frac{8}{\eta}\sqrt{\mathcal{A}}\\&\quad+\max_{a^i,a^{-i}}\left|\mathcal{T}^i(v_t^i)(s,a^i,a^{-i})+\mathcal{T}^{-i}(v_t^{-i})(s,a^i,a^{-i})\right|\\&\leq V_{v_t,s}(\pi_{t,K}^i(s),\pi_{t,K}^{-i}(s))+\frac{8}{\eta}\sqrt{\mathcal{A}}+\gamma\|v_t^i+v_t^{-i}\|_\infty.\end{aligned}$$

It follows that

$$\hat{E}_3\leq 2\gamma\|v_t^i+v_t^{-i}\|_\infty+\max_s V_{v_t,s}(\pi_{t,K}^i(s),\pi_{t,K}^{-i}(s))+\frac{8}{\eta}\sqrt{\mathcal{A}}. \tag{11}$$

Substituting the upper bounds we obtained for the terms $\{E_j\}_{1\leq j\leq 4}$ into Equation 10

$$\begin{aligned}\|v_*^{-i}-v_{\pi_{t,K}^{-i},*}^{-i}\|_\infty &\leq \gamma\|v_{*,\pi_{t,K}^{-i}}^i-v_*^i\|_\infty+2\gamma\|v_t^i+v_t^{-i}\|_\infty+2\gamma\|v_t^i-v_*^i\|_\infty\\&\quad+\max_s V_{v_t,s}(\pi_{t,K}^i(s),\pi_{t,K}^{-i}(s))+\frac{8}{\eta}\sqrt{\mathcal{A}}\\&=\gamma\|v_*^{-i}-v_{\pi_{t,K}^{-i},*}^{-i}\|_\infty+2\gamma\|v_t^i+v_t^{-i}\|_\infty+2\gamma\|v_t^i-v_*^i\|_\infty\\&\quad+\max_s V_{v_t,s}(\pi_{t,K}^i(s),\pi_{t,K}^{-i}(s))+\frac{8}{\eta}\sqrt{\mathcal{A}},\end{aligned}$$

which implies

$$\begin{aligned}\|v_*^{-i}-v_{\pi_{t,K}^{-i},*}^{-i}\|_\infty \leq&\frac{1}{1-\gamma}\Bigg(2\|v_t^i+v_t^{-i}\|_\infty+2\|v_t^i-v_*^i\|_\infty\\&+\max_s V_{v_t,s}(\pi_{t,K}^i(s),\pi_{t,K}^{-i}(s))+\frac{8}{\eta}\sqrt{\mathcal{A}}\Bigg).\end{aligned}$$

Similarly, we also have

$$\begin{aligned}\|v_*^{i}-v_{\pi_{t,K}^{i},*}^{i}\|_\infty \leq&\frac{1}{1-\gamma}\Bigg(2\|v_t^i+v_t^{-i}\|_\infty+2\|v_t^i-v_*^i\|_\infty\\&+\max_s V_{v_t,s}(\pi_{t,K}^i(s),\pi_{t,K}^{-i}(s))+\frac{8}{\eta}\sqrt{\mathcal{A}})\Bigg).\end{aligned}$$

Substituting the previous two inequalities into Equation 9 we get

$$\begin{aligned}\|v_{*,\pi_{t,K}^{-i}}^{i}-v_{\pi_{t,K}^{i},\pi_{t,K}^{-i}}^{i}\|_\infty \leq&\frac{2}{1-\gamma}\Bigg(2\|v_t^i+v_t^{-i}\|_\infty+2\|v_t^i-v_*^i\|_\infty\\&+\max_s V_{v_t,s}(\pi_{t,K}^i(s),\pi_{t,K}^{-i}(s))+\frac{8}{\eta}\sqrt{\mathcal{A}}\Bigg).\end{aligned}$$

□

At this point, it remains to bound the first two terms on the r.h.s in lemma C.1. The next two lemmas achieve this purpose by presenting one-step drift inequalities for the relevant terms.

Lemma C.2. *It holds for all $t \geq 0$ and $i = 1, 2$ that*

$$\|v_{t+1}^i - v_*^i\|_\infty \leq \gamma \|v_t^i - v_*^i\|_\infty + 2 \max_{s \in \mathcal{S}} V_{v_t,s}(\pi_{t,K}^i(s), \pi_{t,K}^{-i}(s)) + \frac{16}{\eta}\sqrt{\mathcal{A}} + \max_{s \in \mathcal{S}} \|\mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s) - q_{t,K}^i(s)\|_\infty + 2\gamma \|v_t^i + v_t^{-i}\|_\infty.$$

Proof. For any $i \in \{1, 2\}$, we have by the outer-loop update equation (cf. Line 1) of Algorithm 1 that

$$v_{t+1}^i(s) = \pi_{t,K}^i(s)^\top q_{t,K}^i(s) = val^i(\mathcal{T}^i(v_t^i)(s)) + \pi_{t,K}^i(s)^\top q_{t,K}^i(s) - val^i(\mathcal{T}^i(v_t^i)(s))$$

Since $val^i(\mathcal{T}^i(v_*^i)(s)) = \mathcal{B}^i(v_*^i)(s) = v_*^i(s)$, we have

$$|v_{t+1}^i(s) - v_*^i(s)| = |val^i(\mathcal{T}^i(v_t^i)(s)) - val^i(\mathcal{T}^i(v_*^i)(s))| + |\pi_{t,K}^i(s)^\top q_{t,K}^i(s) - val^i(\mathcal{T}^i(v_t^i)(s))|. \tag{12}$$

For the first term on the r.h.s of Equation 12, we have by the contraction property of the minimax Bellman operator that

$$\begin{aligned} \left|val^i(\mathcal{T}^i(v_t^i)(s)) - val^i(\mathcal{T}^i(v_*^i)(s))\right| &= \left|\mathcal{B}^i(v_t^i)(s) - \mathcal{B}^i(v_*^i)(s)\right| \\ &\leq \gamma \|v_t^i - v_*^i\|_\infty. \end{aligned}$$

For the second term on the r.h.s of Equation 12, we have

$$\begin{aligned} \left|\pi_{t,K}^i(s)^\top q_{t,K}^i(s) - val^i(\mathcal{T}^i(v_t^i)(s))\right| \leq{}& \underbrace{\left|\max_{\pi^i} (\pi^i)^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s) - \pi_{t,K}^i(s)^\top q_{t,K}^i(s)\right|}_{T_1} \\ &+ \underbrace{\left|\max_{\pi^i} (\pi^i)^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s) - val^i(\mathcal{T}^i(v_t^i)(s))\right|}_{T_2} \end{aligned}$$

For the term T_1, we have

$$\begin{aligned} T_1 \leq{}& \left|\max_{\pi^i} (\pi^i)^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s) - (\pi_{t,K}^i(s))^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s)\right| \\ &+ \left|(\pi_{t,K}^i(s))^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s) - \pi_{t,K}^i(s)^\top q_{t,K}^i(s)\right| \\ \leq{}& \max_{\pi^i} (\pi^i)^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s) - (\pi_{t,K}^i(s))^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s) \\ &+ \|\mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s) - q_{t,K}^i(s)\|_\infty \\ \leq{}& \sum_{i=1,2} \left\{\max_{\pi^i} (\pi^i - \pi_{t,K}^i(s))^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s)\right\} \\ &+ \|\mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s) - q_{t,K}^i(s)\|_\infty \\ \leq{}& \sum_{i=1,2} \left\{\max_{\pi^i} (\pi^i - \pi_{t,K}^i(s))^\top \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s) + \frac{1}{\eta}\nu(\pi^i) - \frac{1}{\eta}\nu(\pi_{t,K}^i(s))\right\} \\ &+ \frac{8}{\eta}\sqrt{\mathcal{A}} + \|\mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s) - q_{t,K}^i(s)\|_\infty \\ \leq{}& V_{v_t,s}(\pi_{t,K}^i(s), \pi_{t,K}^{-i}(s)) + \frac{8}{\eta}\sqrt{\mathcal{A}} + \|\mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s) - q_{t,K}^i(s)\|_\infty. \end{aligned}$$

Note that T_2 is exactly the term $\hat{E}_3$ we analyzed in proving Lemma C.1. Therefore, we have from Equation 11 that

$$T_2 \leq 2\gamma \|v_t^i + v_t^{-i}\|_\infty + \max_s V_{v_t,s}(\pi_{t,K}^i(s), \pi_{t,K}^{-i}(s)) + \frac{8}{\eta}\sqrt{\mathcal{A}}.$$

It follows that

$$\begin{aligned}\left|\pi^i_{t,K}(s)^\top q^i_{t,K}(s) - val^i(\mathcal{T}^i(v^i_t)(s))\right| \leq & T_1 + T_2 \\ \leq & 2\max_s V(\pi^i_{t,K}(s), \pi^{-i}_{t,K}(s)) + 2\gamma\|v^i_t + v^{-i}_t\|_\infty \\ & + \max_s \|\mathcal{T}^i(v^i_t)(s)\pi^{-i}_{t,K}(s) - q^i_{t,K}(s)\|_\infty + \frac{16}{\eta}\sqrt{\mathcal{A}}.\end{aligned}$$

Using the upper bounds we obtained for the two terms on the r.h.s of equation (12)

$$\begin{aligned}|v^i_{t+1}(s) - v^i_*(s)| \leq \gamma\|v^i_t - v^i_*\|_\infty + 2\max_{s\in\mathcal{S}} V(\pi^i_{t,K}(s), \pi^{-i}_{t,K}(s)) + \frac{16}{\eta}\sqrt{\mathcal{A}} \\ + \max_{s\in\mathcal{S}} \|\mathcal{T}^i(v^i_t)(s)\pi^{-i}_{t,K}(s) - q^i_{t,K}(s)\|_\infty + 2\gamma\|v^i_t + v^{-i}_t\|_\infty.\end{aligned}$$

Since the r.h.s of the previous inequality does not depend on s, we have for any $i \in \{1, 2\}$ that

$$\begin{aligned}\|v^i_{t+1} - v^i_*\|_\infty \leq \gamma\|v^i_t - v^i_*\|_\infty + 2\max_{s\in\mathcal{S}} V(\pi^i_{t,K}(s), \pi^{-i}_{t,K}(s)) + \frac{16}{\eta}\sqrt{\mathcal{A}} \\ + \max_{s\in\mathcal{S}} \|\mathcal{T}^i(v^i_t)(s)\pi^{-i}_{t,K}(s) - q^i_{t,K}(s)\|_\infty + 2\gamma\|v^i_t + v^{-i}_t\|_\infty.\end{aligned}$$

□

Lemma C.3. *It holds for all $t \geq 0$ that*

$$\|v^i_{t+1} + v^{-i}_{t+1}\|_\infty \leq \gamma\|v^i_t + v^{-i}_t\|_\infty + \sum_{i=1,2} \max_{s\in\mathcal{S}} \|q^i_{t,K}(s) - \mathcal{T}^i(v^i_t)(s)\pi^{-i}_{t,K}(s)\|_\infty.$$

Proof. Using the outer-loop update equation (line 1 of algorithm 1)

$$\begin{aligned}\left|\sum_{i=1,2} v^i_{t+1}(s)\right| &= \sum_{i=1,2} \pi^i_{t,K}(s)^\top q^i_{t,K}(s) \\ &= \left|\sum_{i=1,2} \pi^i_{t,K}(s)^\top (q^i_{t,K}(s) - \mathcal{T}^i(v^i_t)(s)\pi^{-i}_{t,K}(s))\right| \\ &\quad + \left|\sum_{i=1,2} \pi^i_{t,K}(s)\mathcal{T}^i(v^i_t)(s)\pi^{-i}_{t,K}(s)\right| \\ &\leq \sum_{i=1,2} \max_{s\in\mathcal{S}} \|q^i_{t,K}(s) - \mathcal{T}^i(v^i_t)(s)\pi^{-i}_{t,K}(s)\|_\infty \\ &\quad + \max_{(s,a^i,a^{-i})} \left|\mathcal{T}^i(v^i_t)(s, a^i, a^{-i}) + \mathcal{T}^{-i}(v^{-i}_t)(s, a^i, a^{-i})\right| \\ &\leq \sum_{i=1,2} \max_{s\in\mathcal{S}} \|q^i_{t,K}(s) - \mathcal{T}^i(v^i_t)(s)\pi^{-i}_{t,K}(s)\|_\infty + \gamma\|v^i_t + v^{-i}_t\|_\infty.\end{aligned}$$

Since the r.h.s of the previous inequality does not depend on s, we have

$$\|v^i_{t+1} + v^{-i}_{t+1}\|_\infty \leq \gamma\|v^i_t + v^{-i}_t\|_\infty + \sum_{i=1,2} \max_{s\in\mathcal{S}} \|q^i_{t,K}(s) - \mathcal{T}^i(v^i_t)(s)\pi^{-i}_{t,K}(s)\|_\infty.$$

□

C.2. Inner-loop

C.2.1. Analyzing the policy update

For readability purposes, and in this section only, we denote $X^i = \mathcal{T}^i(v)(s,.,.) \in \mathbb{R}^{|\mathcal{A}^i|\times|\mathcal{A}^{-i}|}$ and we define

$$V_X\left(\pi^i,\pi^{-i}\right) = \sum_{i=1,2} \max_{\hat{\pi}^i\in\Delta^{|\mathcal{A}^i|}} \left\{\left(\hat{\pi}^i-\pi^i\right)^\top X_i\pi^{-i} + \frac{1}{\eta}\mathcal{H}\left(\hat{\pi}^i\right) - \frac{1}{\eta}\mathcal{H}\left(\pi^i\right)\right\}$$

this is introduced to simplify the notation of $V_{v,s}(\pi^i,\pi^{-i})$ defined in Equation 25. Indeed, it is easy to see that $V_{v,s}(\cdot,\cdot) := V_{\left(\mathcal{T}^i(v^i)(s),\mathcal{T}^{-i}(v^{-i})(s)\right)}(\cdot,\cdot)$.

We use $\nabla_1 V_X(\cdot,\cdot)$ (respectively, $\nabla_2 V_X(\cdot,\cdot)$) to denote the gradient concerning the first (respectively, second) argument. The following lemma establishes the strong convexity and the smoothness of $V_X(\pi^i,\pi^{-i})$.

Lemma C.4. *The function $V_X(\cdot,\cdot)$ has the following properties.*

1. *For all $\pi^{-i}\in\Delta^{|\mathcal{A}^{-i}|}$, $V_X(\pi^i,\pi^{-i})$ as a function of π^i is $1/\eta$ – strongly convex with respect to $\|\cdot\|_2$.*
2. *For any $\delta_i>0$ and $\pi^{-i}\in\Delta^{|\mathcal{A}^{-i}|}$, the function $V_X(\cdot,\pi^{-i})$ is $L_{\eta,i}$ – smooth on $\{\pi^i\in\Delta^{|\mathcal{A}^i|} \mid \min_{a^i}\pi^i(a^i)\geq\delta_i\}$ with respect to $\|\cdot\|_2$, where $L_{\eta,i} = \left(2\sqrt{2}\,\eta\,\sigma^2_{\max}(X_{-i})\,|\mathcal{A}^{-i}| + \frac{1}{\eta\delta_i^{3/2}}\right)$.*
3. *It holds for any (π^i,π^{-i}) that*

$$\begin{aligned}\langle\nabla_1 V_X(\pi^i,\pi^{-i}),\mathrm{Ts}(X_i\pi^{-i})-\pi^i\rangle + &\langle\nabla_2 V_X(\pi^i,\pi^{-i}),\mathrm{Ts}(X_{-i}\pi^i)-\pi^{-i}\rangle\\ &\leq -\frac{7}{8}V_X(\pi^i,\pi^{-i}) + 16\eta\|X_i+X_{-i}^\top\|_2^2.\end{aligned}$$

4. *For any $u^i\in\mathbb{R}^{|\mathcal{A}^i|}, u^{-i}\in\mathbb{R}^{|\mathcal{A}^{-i}|}$, we have for all $(\pi^i,\pi^{-i})\in\{\pi^i\in\Delta^{|\mathcal{A}^i|},\pi^{-i}\in\Delta^{|\mathcal{A}^{-i}|} \mid \min_{a^i}\pi^i(a^i)\geq\delta_i,\min_{a^{-i}}\pi^{-i}(a^{-i})\geq\delta_{-i}\}$ (where $\delta_i,\delta_{-i}>0$) that*

$$\begin{aligned}&\langle\nabla_1 V_X(\pi^i,\pi^{-i}),\mathrm{Ts}(u^i)-\mathrm{Ts}(X_i\pi^{-i})\rangle + \langle\nabla_2 V_X(\pi^i,\pi^{-i}),\mathrm{Ts}(u^{-i})-\mathrm{Ts}(X_{-i}\pi^i)\rangle\\ &\leq\left(\frac{1}{\eta\delta_i^{3/2}}+\frac{1}{\eta\delta_{-i}^{3/2}}+\|X_i\|_2+\|X_{-i}\|_2\right)\Bigg[2\bar{c}\eta V_X(\pi^i,\pi^{-i})\\ &\quad+\frac{8\eta^2|\mathcal{A}^i|^2}{\bar{c}}\|u^i-X_i\pi^{-i}\|_2^2+\frac{8\eta^2|\mathcal{A}^{-i}|^2}{\bar{c}}\|u^{-i}-X_i\pi^i\|_2^2\Bigg]\end{aligned}$$

where $\bar{c}$ is any positive real number.

Proof. First, observe that

$$\arg\max_{\hat{\pi}^i\in\Delta^{|\mathcal{A}^i|}}\left\{\left(\hat{\pi}^i\right)^\top X_i\pi^{-i}+\frac{1}{\eta}\mathcal{H}\left(\hat{\pi}^i\right)\right\} = \mathrm{Ts}_\eta\left(X_i\pi^{-i}\right).$$

Therefore, the function $V_X(\cdot,\cdot)$ can be equivalently written as

$$\begin{aligned}V_X\left(\pi^i,\pi^{-i}\right) = \sum_{i=1,2}\Bigg[&\left(\mathrm{Ts}_\eta\left(X_i\pi^{-i}\right)\right)^\top X_i\pi^{-i}+\frac{1}{\eta}\mathcal{H}\left(\mathrm{Ts}_\eta\left(X_i\pi^{-i}\right)\right)\\ &-\left(\pi^i\right)^\top X_i\pi^{-i}-\frac{1}{\eta}\mathcal{H}\left(\pi^i\right)\Bigg]\end{aligned}$$

1) Convexity: first, it is easy to show that the negative Tsallis–entropy $-\mathcal{H}(\cdot)$ is 1–strongly convex with respect to $\|\|_2$. Indeed, $\nabla^2(-\mathcal{H}(\pi)) = \mathrm{diag}((1/\pi_a^{3/2})_{a\in\mathcal{A}})$ then $\nabla^2(-\mathcal{H}(\pi))\geq I_{\mathcal{A}}$, *i.e.* it is 1–strongly convex.

Second, we have that

$$\left(\mathrm{Ts}_\eta\left(X_{-i}\pi^i\right)\right)^\top X_{-i}\pi^i + \frac{1}{\eta}\mathcal{H}\left(\mathrm{Ts}\left(X_{-i}\pi^i\right)\right) = \max_{\hat{\pi}^{-i}\in\Delta^{|\mathcal{A}-i|}}\left\{\left(\hat{\pi}^{-i}\right)^\top X_{-i}\pi^i + \frac{1}{\eta}\mathcal{H}\left(\hat{\pi}^{-i}\right)\right\},$$

which is convex as a maximum of convex functions.

Therefore, the function $V_X\left(\cdot,\pi^{-i}\right)$ is $1/\eta$ - strongly convex with respect to $\|\cdot\|_2$ uniformly for all π^{-i}.

2) Smoothness: first, from the Hessian $\nabla^2(\mathcal{H}(\pi)) = \mathrm{diag}((1/\pi_a^{3/2})_{a\in\mathcal{A}})$, it is clear that the negative Tsallis entropy is $1/\delta_i^{3/2}$– smooth on $\{\pi^i \in \Delta^{|\mathcal{A}^i|} \mid \min_{a^i}\pi^i(a^i) \geq \delta_i\}$ with respect to $\|\cdot\|_2$.

Second, we have

$$\begin{aligned}
\nabla_{\pi^i}\Big(\mathrm{Ts}_\eta\left(X_{-i}\pi^i\right)^\top X_{-i}\pi^i + \frac{1}{\eta}\mathcal{H}\left(\mathrm{Ts}_\eta\left(X_{-i}\pi^i\right)\right)\Big)& \\
&= \nabla_{\pi^i}\max_{\hat{\pi}^{-i}\in\Delta^{|\mathcal{A}-i|}}\left\{\left(\hat{\pi}^{-i}\right)^\top X_{-i}\pi^i + \frac{1}{\eta}\mathcal{H}\left(\hat{\pi}^{-i}\right)\right\} \\
&= X_{-i}^\top\,\mathrm{Ts}_\eta\left(X_{-i}\pi^i\right).
\end{aligned}$$

where the first line follows using Danskin's theorem. Since there aren't readily usable formulas for the gradient of the Tsallis weights, we will use the standard characterization of smoothness. Let $\pi_1,\pi_2\in\Delta^{|\mathcal{A}^i|}$, we have by lemma B.3 we have:

$$\begin{aligned}
X_{-i}^\top\,\mathrm{Ts}_\eta\left(X_{-i}\pi_1\right) - X_{-i}^\top\,\mathrm{Ts}_\eta\left(X_{-i}\pi_2\right) &= X_{-i}^\top\left(\mathrm{Ts}_\eta\left(X_{-i}\pi_1\right) - \mathrm{Ts}_\eta\left(X_{-i}\pi_2\right)\right) \\
&\leq \sigma_{\max}(X_{-i})\|\,\mathrm{Ts}_\eta\left(X_{-i}\pi_1\right) - \mathrm{Ts}_\eta\left(X_{-i}\pi_2\right)\|_2 \\
&\leq \sigma_{\max}(X_{-i})2\sqrt{2}\eta|\mathcal{A}^{-i}|\|X_{-i}(\pi_1-\pi_2)\|_2 \\
&\leq 2\sqrt{2}\,\eta\,\sigma_{\max}^2(X_{-i})\,|\mathcal{A}^{-i}|\,\|\pi_1-\pi_2\|_2.
\end{aligned}$$

Since we showed before that

$$\nabla_{\pi^i}\left(\mathrm{Ts}_\eta\left(X_{-i}\pi^i\right)^\top X_{-i}\pi^i + \frac{1}{\eta}\mathcal{H}\left(\mathrm{Ts}_\eta\left(X_{-i}\pi^i\right)\right)\right) = X_{-i}^\top\,\mathrm{Ts}_\eta\left(X_{-i}\pi^i\right),$$

we deduce that the function $\mathrm{Ts}_\eta\left(X_{-i}\pi^i\right)^\top X_{-i}\pi^i + \frac{1}{\eta}\mathcal{H}\left(\mathrm{Ts}_\eta\left(X_{-i}\pi^i\right)\right)$ is $2\sqrt{2}\,\eta\,\sigma_{\max}(X_{-i})^2\,|\mathcal{A}^{-i}|$ smooth with respect to $\|\cdot\|_2$.

Combining with the smoothness of Tsallis entropy, we conclude that $V_X\left(\cdot,\pi^{-i}\right)$ is $L_{\eta,i}$–smooth with respect to $\|\cdot\|_2$ uniformly for all π^{-i}, where we defined $L_{\eta,i} = 2\sqrt{2}\eta\sigma_{\max}^2 X_{-i})|\mathcal{A}^{-i}| + \frac{1}{\eta\delta_i^{3/2}}$.

3) We first compute the gradient $\nabla_1 V_X(\pi^i,\pi^{-i})$ using Danskin's theorem:

$$\nabla_1 V_X(\pi^i,\pi^{-i}) = -(X_i + X_{-i}^\top)\pi^{-i} - \frac{1}{\eta}\nabla\mathcal{H}(\pi^i) + X_{-i}^\top\,\mathrm{Ts}(X_{-i}\pi^i). \tag{13}$$

It follows that

$$\begin{aligned}
&\langle\nabla_1 V_X(\pi^i,\pi^{-i}),\mathrm{Ts}(X_i\pi^{-i}) - \pi^i\rangle \\
&\quad= \langle-(X_i + X_{-i}^\top)\pi^{-i} - \frac{1}{\eta}\nabla\mathcal{H}(\pi^i) + X_{-i}^\top\,\mathrm{Ts}(X_{-i}\pi^i),\mathrm{Ts}(X_i\pi^{-i}) - \pi^i\rangle \\
&\quad= \langle-(X_i + X_{-i}^\top)\pi^{-i} - \frac{1}{\eta}\nabla\mathcal{H}(\pi^i) + X_{-i}^\top\,\mathrm{Ts}(X_{-i}\pi^i),\mathrm{Ts}(X_i\pi^{-i}) - \pi^i\rangle \\
&\qquad+ \langle X_i\pi^{-i} + \frac{1}{\eta}\nabla\mathcal{H}(\mathrm{Ts}(X_i\pi^{-i})),\mathrm{Ts}(X_i\pi^{-i}) - \pi^i\rangle \\
&\quad= \frac{1}{\eta}\langle\nabla\mathcal{H}(\mathrm{Ts}(X_i\pi^{-i})) - \nabla\mathcal{H}(\pi^i),\mathrm{Ts}(X_i\pi^{-i}) - \pi^i\rangle \\
&\qquad+ (\mathrm{Ts}(X_{-i}\pi^i) - \pi^{-i})^\top X_{-i}(\mathrm{Ts}(X_i\pi^{-i}) - \pi^i).
\end{aligned} \tag{14}$$

where Equation 14 from from the optimality condition $X_i\pi^{-i} + \frac{1}{\eta}\nabla\mathcal{H}(\mathrm{Ts}(X_i\pi^{-i})) = 0$.

To proceed, observe that the concavity of $\mathcal{H}(\cdot)$ and the optimality condition imply

$$\begin{aligned}
&\langle \nabla\mathcal{H}(\mathrm{Ts}(X_i\pi^{-i})) - \nabla\mathcal{H}(\pi^i), \mathrm{Ts}(X_i\pi^{-i}) - \pi^i\rangle \\
&\qquad = \langle \nabla\mathcal{H}(\pi^i) - \nabla\mathcal{H}(\mathrm{Ts}(X_i\pi^{-i})), \pi^i - \mathrm{Ts}(X_i\pi^{-i})\rangle \\
&\qquad = \langle \nabla\mathcal{H}(\pi^i), \pi^i - \mathrm{Ts}(X_i\pi^{-i})\rangle - \langle \nabla\mathcal{H}(\mathrm{Ts}(X_i\pi^{-i})), \pi^i - \mathrm{Ts}(X_i\pi^{-i})\rangle \\
&\qquad \le \mathcal{H}(\pi^i) - \mathcal{H}(\mathrm{Ts}(X_i\pi^{-i})) - \langle \nabla\mathcal{H}(\mathrm{Ts}(X_i\pi^{-i})), \pi^i - \mathrm{Ts}(X_i\pi^{-i})\rangle \\
&\qquad = \mathcal{H}(\pi^i) - \mathcal{H}(\mathrm{Ts}(X_i\pi^{-i})) + \eta\langle X_i\pi^{-i}, \pi^i - \mathrm{Ts}(X_i\pi^{-i})\rangle \\
&\qquad = \eta\left[(\pi^i)^\top X_i\pi^{-i} + \frac{1}{\eta}\mathcal{H}(\pi^i) - \max_{\hat\pi^i\in\Delta^{|\mathcal{A}^i|}}\left\{(\hat\pi^i)^\top X_i\pi^{-i} + \frac{1}{\eta}\mathcal{H}(\hat\pi^i)\right\}\right].
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\langle \nabla_1 V_X(\pi^i,\pi^{-i}), \mathrm{Ts}(X_i\pi^{-i}) - \pi^i\rangle \le &\left[(\pi^i)^\top X_i\pi^{-i} + \frac{1}{\eta}\mathcal{H}(\pi^i)\right. \\
&\left. - \max_{\hat\pi^i\in\Delta^{|\mathcal{A}^i|}}\left\{(\hat\pi^i)^\top X_i\pi^{-i} + \frac{1}{\eta}\mathcal{H}(\hat\pi^i)\right\}\right] \\
&+ (\mathrm{Ts}(X_{-i}\pi^i) - \pi^{-i})^\top X_{-i}(\mathrm{Ts}(X_i\pi^{-i}) - \pi^i)
\end{aligned}$$

Similarly, we also have

$$\begin{aligned}
&\langle \nabla_2 V_X(\pi^i,\pi^{-i}), \mathrm{Ts}(X_{-i}\pi^i) - \pi^{-i}\rangle \\
&\quad \le \left[(\pi^{-i})^\top X_{-i}\pi^i + \frac{1}{\eta}\mathcal{H}(\pi^{-i}) - \max_{\hat\pi^{-i}\in\Delta^{|\mathcal{A}^{-i}|}}\left\{(\hat\pi^{-i})^\top X_{-i}\pi^i + \frac{1}{\eta}\mathcal{H}(\hat\pi^{-i})\right\}\right] \\
&\quad\quad + (\mathrm{Ts}(X_i\pi^{-i}) - \pi^i)^\top X_i(\mathrm{Ts}(X_{-i}\pi^i) - \pi^{-i})
\end{aligned}$$

Adding up the previous two inequalities we obtain

$$\begin{aligned}
&\langle \nabla_1 V_X(\pi^i,\pi^{-i}), \mathrm{Ts}(X_i\pi^{-i}) - \pi^i\rangle + \langle \nabla_2 V_X(\pi^i,\pi^{-i}), \mathrm{Ts}(X_{-i}\pi^i) - \pi^{-i}\rangle \\
&\quad \le -V_X(\pi^i,\pi^{-i}) + (\mathrm{Ts}(X_i\pi^{-i}) - \pi^i)^\top(X_i + X_{-i}^\top)(\mathrm{Ts}(X_{-i}\pi^i) - \pi^{-i}).
\end{aligned} \tag{15}$$

To control the second term on the r.h.s of Equation 15, observe that

$$\begin{aligned}
&(\mathrm{Ts}(X_i\pi^{-i}) - \pi^i)^\top(X_i + X_{-i}^\top)(\mathrm{Ts}(X_{-i}\pi^i) - \pi^{-i}) \\
&\qquad \le \|\mathrm{Ts}(X_i\pi^{-i}) - \pi^i\|_2\|X_i + X_{-i}^\top\|_2\|\mathrm{Ts}(X_{-i}\pi^i) - \pi^{-i}\|_2 \\
&\qquad \le (\|\mathrm{Ts}(X_i\pi^{-i})\|_2 + \|\pi^i\|_2)\|X_i + X_{-i}^\top\|_2\|\mathrm{Ts}(X_{-i}\pi^i) - \pi^{-i}\|_2 \\
&\qquad \le 2\|X_i + X_{-i}^\top\|_2\|\mathrm{Ts}(X_{-i}\pi^i) - \pi^{-i}\|_2 \\
&\qquad \le c_1\|X_i + X_{-i}^\top\|_2^2 + \frac{1}{c_1}\|\mathrm{Ts}(X_{-i}\pi^i) - \pi^{-i}\|_2^2 \qquad \text{(This is true for all } c_1 > 0) \\
&\qquad \le c_1\|X_i + X_{-i}^\top\|_2^2 + \frac{1}{c_1}(\|\mathrm{Ts}(X_{-i}\pi^i) - \pi^{-i}\|_2^2 + \|\mathrm{Ts}(X_i\pi^{-i}) - \pi^i\|_2^2).
\end{aligned} \tag{16}$$

Next, note that the function

$$F_{X_i}(\pi^i,\pi^{-i}) := \max_{\hat\pi^i}\left\{(\hat\pi^i - \pi^i)^\top X_i\pi^{-i} + \frac{1}{\eta}\mathcal{H}(\hat\pi^i) - \frac{1}{\eta}\mathcal{H}(\pi^i)\right\}$$

is $\frac{1}{\eta}$-strongly convex as a function of π^i uniformly for all π^{-i}. Therefore, we have

$$\begin{aligned} F_{X_i}(\pi^i, \pi^{-i}) &= F_{X_i}(\pi^i, \pi^{-i}) - F_{X_i}(\mathrm{Ts}(X_i\pi^{-i}), \pi^{-i}) \\ &= F_{X_i}(\pi^i, \pi^{-i}) - \min_{\pi^i} F_{X_i}(\pi^i, \pi^{-i}) \\ &\geq \frac{1}{2\eta}\| \mathrm{Ts}(X_i\pi^{-i}) - \pi^i\|_2^2, \end{aligned}$$

which is called the quadratic growth property in optimization literature. It follows that

$$\| \mathrm{Ts}(X_i\pi^{-i}) - \pi^i\|_2^2 \leq 2\eta \max_{\hat{\pi}^i} \left\{ (\hat{\pi}^i - \pi^i)^\top X_i \pi^{-i} + \frac{1}{\eta}\mathcal{H}(\hat{\pi}^i) - \frac{1}{\eta}\mathcal{H}(\pi^i) \right\}.$$

Similarly, we also have

$$\| \mathrm{Ts}(X_{-i}\pi^{i}) - \pi^{-i}\|_2^2 \leq 2\eta \max_{\hat{\pi}^{-i}} \left\{ (\hat{\pi}^{-i} - \pi^{-i})^\top X_{-i} \pi^{i} + \frac{1}{\eta}\mathcal{H}(\hat{\pi}^{-i}) - \frac{1}{\eta}\mathcal{H}(\pi^{-i}) \right\}.$$

Adding up the previous two inequalities

$$\| \mathrm{Ts}(X_{-i}\pi^{i}) - \pi^{-i}\|_2^2 + \| \mathrm{Ts}(X_i\pi^{-i}) - \pi^i\|_2^2 \leq 2\eta V_X(\pi^i, \pi^{-i}).$$

Plugging the previous inequality in Equation 16

$$\begin{aligned} &(\mathrm{Ts}(X_i\pi^{-i}) - \pi^i)^\top (X_i + X_{-i}^\top)(\mathrm{Ts}(X_{-i}\pi^i) - \pi^{-i}) \\ &\qquad \leq c_1 \|X_i + X_{-i}^\top\|_2^2 + \frac{1}{c_1}(\| \mathrm{Ts}(X_{-i}\pi^{i}) - \pi^{-i}\|_2^2 + \| \mathrm{Ts}(X_i\pi^{-i}) - \pi^i\|_2^2) \\ &\qquad \leq c_1 \|X_i + X_{-i}^\top\|_2^2 + \frac{2\eta}{c_1} V_X(\pi^i, \pi^{-i}) \\ &\qquad = 16\eta \|X_i + X_{-i}^\top\|_2^2 + \frac{1}{8} V_X(\pi^i, \pi^{-i}), \end{aligned}$$

where the last line follows from choosing $c_1 = 16\eta$. Using the previous inequality in Equation 15

$$\begin{aligned} \langle \nabla_1 V_X(\pi^i, \pi^{-i}), \mathrm{Ts}(X_i\pi^{-i}) - \pi^i\rangle + \langle \nabla_2 V_X(\pi^i, \pi^{-i}), \mathrm{Ts}(X_{-i}\pi^{i}) - \pi^{-i}\rangle& \\ \leq\ -\frac{7}{8} V_X(\pi^i, \pi^{-i}) + 16\eta\|X_i + X_{-i}^\top\|_2^2&. \end{aligned}$$

4) For any $u^i \in \mathbb{R}^{|\mathcal{A}^i|}$, using the explicit expression of the gradient of $V_X(\cdot,\cdot)$ from Equation 13

$$\begin{aligned} &\langle \nabla_1 V_X(\pi^i, \pi^{-i}), \mathrm{Ts}(u^i) - \mathrm{Ts}(X_i\pi^{-i})\rangle \\ &\quad = \langle -(X_i + X_{-i}^\top)\pi^{-i} - \frac{1}{\eta}\nabla\mathcal{H}(\pi^i) + X_{-i}^\top \mathrm{Ts}(X_{-i}\pi^i), \mathrm{Ts}(u^i) - \mathrm{Ts}(X_i\pi^{-i})\rangle \\ &\quad = \langle -(X_i + X_{-i}^\top)\pi^{-i} - \frac{1}{\eta}\nabla\mathcal{H}(\pi^i) + X_{-i}^\top \mathrm{Ts}(X_{-i}\pi^i), \mathrm{Ts}(u^i) - \mathrm{Ts}(X_i\pi^{-i})\rangle \\ &\qquad + \langle X_i\pi^{-i} + \frac{1}{\eta}\nabla\mathcal{H}(\mathrm{Ts}(X_i\pi^{-i})), \mathrm{Ts}(u^i) - \mathrm{Ts}(X_i\pi^{-i})\rangle \\ &\quad = \frac{1}{\eta}\langle \nabla\mathcal{H}(\mathrm{Ts}(X_i\pi^{-i})) - \nabla\mathcal{H}(\pi^i), \mathrm{Ts}(u^i) - \mathrm{Ts}(X_i\pi^{-i})\rangle \\ &\qquad + (\mathrm{Ts}(X_{-i}\pi^i) - \pi^{-i})^\top X_{-i}(\mathrm{Ts}(u^i) - \mathrm{Ts}(X_i\pi^{-i})) \\ &\quad \leq \frac{1}{\eta}\|\nabla\mathcal{H}(\mathrm{Ts}(X_i\pi^{-i})) - \nabla\mathcal{H}(\pi^i)\|_2 \| \mathrm{Ts}(u^i) - \mathrm{Ts}(X_i\pi^{-i})\|_2 \\ &\qquad + \| \mathrm{Ts}(X_{-i}\pi^i) - \pi^{-i}\|_2 \|X_{-i}\|_2 \| \mathrm{Ts}(u^i) - \mathrm{Ts}(X_i\pi^{-i})\|_2 \\ &\quad \leq \frac{1}{\eta\delta_i^{3/2}}\| \mathrm{Ts}(X_i\pi^{-i}) - \pi^i\|_2 \| \mathrm{Ts}(u^i) - \mathrm{Ts}(X_i\pi^{-i})\|_2 \\ &\qquad + \|X_{-i}\|_2 \| \mathrm{Ts}(X_{-i}\pi^i) - \pi^{-i}\|_2 \| \mathrm{Ts}(u^i) - \mathrm{Ts}(X_i\pi^{-i})\|_2, \end{aligned}$$

where the last inequality follows from the smoothness of $\mathcal{H}(\cdot)$ in Lemma. C.4 (2). Similarly, we also have for any $u^{-i} \in \mathbb{R}^{|\mathcal{A}^{-i}|}$ that

$$
\begin{aligned}
&\langle \nabla_2 V_X(\pi^i, \pi^{-i}), \operatorname{Ts}(u^{-i}) - \operatorname{Ts}(X_{-i}\pi^i) \rangle \\
&\quad \leq \frac{1}{\eta \delta_{-i}^{3/2}} \| \operatorname{Ts}(X_{-i}\pi^i) - \pi^{-i} \|_2 \| \operatorname{Ts}(u^{-i}) - \operatorname{Ts}(X_{-i}\pi^i) \|_2 \\
&\qquad + \|X_i\|_2 \| \operatorname{Ts}(X_i \pi^{-i}) - \pi^i \|_2 \| \operatorname{Ts}(u^{-i}) - \operatorname{Ts}(X_i \pi^i) \|_2 .
\end{aligned}
$$

Adding up the previous two inequalities

$$
\begin{aligned}
&\langle \nabla_1 V_X(\pi^i, \pi^{-i}), \operatorname{Ts}(u^i) - \operatorname{Ts}(X_i\pi^{-i}) \rangle + \langle \nabla_2 V_X(\pi^i, \pi^{-i}), \operatorname{Ts}(u^{-i}) - \operatorname{Ts}(X_{-i}\pi^i) \rangle \\
&\leq \left(\frac{1}{\eta \delta_i^{3/2}} + \frac{1}{\eta \delta_{-i}^{3/2}} + \|X_i\|_2 + \|X_{-i}\|_2 \right) \left(\sum_i \| \operatorname{Ts}(X_i\pi^{-i}) - \pi^i \|_2 \right) \\
&\quad \times \left(\| \operatorname{Ts}(u^i) - \operatorname{Ts}(X_i\pi^{-i}) \|_2 + \| \operatorname{Ts}(u^{-i}) - \operatorname{Ts}(X_i\pi^i) \|_2 \right) \\
&\leq \frac{1}{2} \left(\frac{1}{\eta \delta_i^{3/2}} + \frac{1}{\eta \delta_{-i}^{3/2}} + \|X_i\|_2 + \|X_{-i}\|_2 \right) \\
&\quad \times \Bigg[\bar{c} \left(\| \operatorname{Ts}(X_i\pi^{-i}) - \pi^i \|_2 + \| \operatorname{Ts}(X_{-i}\pi^i) - \pi^{-i} \|_2 \right)^2 \\
&\qquad + \frac{1}{\bar{c}} \left(\| \operatorname{Ts}(u^i) - \operatorname{Ts}(X_i\pi^{-i}) \|_2 + \| \operatorname{Ts}(u^{-i}) - \operatorname{Ts}(X_i\pi^i) \|_2 \right)^2 \Bigg]
\end{aligned}
$$

the last line is true for all $\bar{c} > 0$, then,

$$
\begin{aligned}
&\langle \nabla_1 V_X(\pi^i, \pi^{-i}), \operatorname{Ts}(u^i) - \operatorname{Ts}(X_i\pi^{-i}) \rangle + \langle \nabla_2 V_X(\pi^i, \pi^{-i}), \operatorname{Ts}(u^{-i}) - \operatorname{Ts}(X_{-i}\pi^i) \rangle \\
&\leq \left(\frac{1}{\eta \delta_i^{3/2}} + \frac{1}{\eta \delta_{-i}^{3/2}} + \|X_i\|_2 + \|X_{-i}\|_2 \right) \bar{c} \left(\sum_i \| \operatorname{Ts}(X_i\pi^{-i}) - \pi^i \|_2^2 \right) \\
&\quad + \frac{1}{\bar{c}} \| \operatorname{Ts}(u^i) - \operatorname{Ts}(X_i\pi^{-i}) \|_2^2 + \frac{1}{\bar{c}} \| \operatorname{Ts}(u^{-i}) - \operatorname{Ts}(X_i\pi^i) \|_2^2 \Bigg] \\
&\leq \left(\frac{1}{\eta \delta_i^{3/2}} + \frac{1}{\eta \delta_{-i}^{3/2}} + \|X_i\|_2 + \|X_{-i}\|_2 \right) \Bigg[2\bar{c}\eta V_X(\pi^i, \pi^{-i}) \\
&\quad + \frac{8\eta^2 |\mathcal{A}^i|^2}{\bar{c}} \|u^i - X_i\pi^{-i}\|_2^2 + \frac{8\eta^2 |\mathcal{A}^{-i}|^2}{\bar{c}} \|u^{-i} - X_i\pi^i\|_2^2 \Bigg],
\end{aligned}
$$

where the penultimate line holds because $(a+b)^2 \leq 2(a^2+b^2)$ for all $a, b \in \mathbb{R}$, and last line follows from the quadratic growth property of strongly convex functions and the Lipschitz continuity of the Tsallis function (*cf.* lemma B.3).

□

With the properties of $V_X(\cdot, \cdot)$ established above, we can now use it to study the policy update π_k^i and π_k^{-i}. Specifically, recall that V_X is just a temporary notation in this section for $V_{v,s}$. Therefore, the newly proved smoothness of $V_{v,s}$, the update equation, and lemma C.4 provide us with the drift inequality for $\mathcal{V}_\pi$ in the lemma below.

Lemma C.5 (Policy drift)**.** *The following inequality holds for all $k \geq 0$:*

$$
\begin{aligned}
\sum_s \mathbb{E}[V_{v,s}(\pi_{k+1}^i(s), \pi_{k+1}^{-i}(s))] \leq{}& \left(1 - \frac{3\beta_k}{4}\right) \sum_s \mathbb{E}[V_{v,s}(\pi_k^i(s), \pi_k^{-i}(s))] + \frac{4|\mathcal{S}|\mathcal{A}^2}{\ell_\eta (1-\gamma)^2} \beta_k^2 \\
&+ \frac{2048\mathcal{A}^4 \beta_k \eta^3}{\ell_\eta^3 (1-\gamma)^2} \sum_{i=1,2} \sum_s \mathbb{E}[\|q_k^i(s) - \mathcal{T}^i(v^i)(s)\pi_k^{-i}(s)\|_2^2] \\
&+ 16|\mathcal{S}|\mathcal{A}\beta_k \eta \|v^i + v^{-i}\|_\infty^2 .
\end{aligned}
$$

Proof. Since $\min_{i=1,2}\min_{s,a^i}\pi_k^i(a^i|s)\geq \ell_\eta$ (by lemma 4.1), then lemma C.4 (2) implies that the function $V_{v,s}(\pi^i,\pi^{-i})$ as a function of π^i is $L_{\eta,i}$ – smooth on $\{\pi^i\in\Delta^{|\mathcal{A}^i|}\mid \min_{a^i}\pi^i(a^i)\geq\ell_\eta\}$ uniformly for all π^{-i}, where

$$L_{\eta,i}:=\left(2\sqrt{2}\,\eta\,\sigma^2_{\max}(\mathcal{T}^{-i}(v^{-i})(s))\,|\mathcal{A}^{-i}|+\frac{1}{\eta\ell_e ta^{3/2}}\right).$$

We next bound $L_{\eta,i}$ from above. Since $\|v^i\|_\infty\leq 1/(1-\gamma)$ and $\|v^{-i}\|_\infty\leq 1/(1-\gamma)$, we have for any (s,a^i,a^{-i}) that

$$\begin{aligned}|\mathcal{T}^{-i}(v^{-i})(s,a^{-i},a^i)| &\leq |\mathcal{R}^{-i}(s,a^{-i},a^i)|+\gamma\mathbb{E}[|v^{-i}(S_1)|\mid S_0=s,A_0^i=a^i,A_0^{-i}=a^{-i}]\\ &\leq 1+\frac{\gamma}{1-\gamma}\\ &=\frac{1}{1-\gamma},\end{aligned}$$

which implies

$$\sigma_{\max}(\mathcal{T}^{-i}(v^{-i})(s))=\|\mathcal{T}^{-i}(v^{-i})(s))\|_2\leq\frac{\sqrt{|\mathcal{A}^i||\mathcal{A}^{-i}|}}{1-\gamma}\leq\frac{\mathcal{A}}{1-\gamma}. \tag{17}$$

As a result, we have by $\frac{1}{\eta}\leq 1$ and $\ell_\eta\leq 1$ that

$$L_{\eta,i}=\eta\sigma^2_{\max}(\mathcal{T}^{-i}(v^{-i})(s))+\frac{1}{\eta\ell_\eta}\leq\frac{\eta\mathcal{A}^2}{(1-\gamma)^2}+\frac{1}{\eta\ell_\eta}\leq\frac{2\mathcal{A}^2}{\ell_\eta(1-\gamma)^2}:=L_\eta.$$

Similarly, $V_{v,s}(\pi^i,\pi^{-i})$ is also L_η – smooth on the set $\{\pi^{-i}\in\Delta^{|\mathcal{A}^{-i}|}\mid\min_{a^{-i}}\pi^i(a^{-i})\geq\ell_\eta\}$ uniformly for all π^i.

Using the smoothness of $V_{v,s}(\cdot,\cdot)$ established above, for any $s\in\mathcal{S}$, we have by the policy update equation that

$$\begin{aligned}&V_{v,s}(\pi^i_{k+1}(s),\pi^{-i}_{k+1}(s))\\ &\quad\leq V_{v,s}(\pi_k^i(s),\pi_k^{-i}(s))+\beta_k\langle\nabla_2 V_{v,s}(\pi_k^i(s),\pi_k^{-i}(s)),\mathrm{Ts}(q_k^{-i}(s))-\pi_k^{-i}(s)\rangle\\ &\qquad+\beta_k\langle\nabla_1 V_{v,s}(\pi_k^i(s),\pi_{k+1}^{-i}(s)),\mathrm{Ts}(q_k^i(s))-\pi_k^i(s)\rangle\\ &\qquad+\frac{L_\eta\beta_k^2}{2}\|\,\mathrm{Ts}(q_k^i(s))-\pi_k^i(s)\|_2^2+\frac{L_\eta\beta_k^2}{2}\|\,\mathrm{Ts}(q_k^{-i}(s))-\pi_k^{-i}(s)\|_2^2\\ &\quad\leq V_{v,s}(\pi_k^i(s),\pi_k^{-i}(s))\\ &\qquad+\underbrace{\beta_k\langle\nabla_2 V_{v,s}(\pi_k^i(s),\pi_k^{-i}(s)),\mathrm{Ts}(\mathcal{T}^{-i}(v^{-i})(s)\pi_k^i(s))-\pi_k^{-i}(s)\rangle}_{\hat{N}_1}\\ &\qquad+\underbrace{\beta_k\langle\nabla_1 V_{v,s}(\pi_k^i(s),\pi_{k+1}^{-i}(s)),\mathrm{Ts}(\mathcal{T}^{i}(v^{i})(s)\pi_k^{-i}(s))-\pi_k^{i}(s)\rangle}_{\hat{N}_2}\\ &\qquad+\underbrace{\beta_k\langle\nabla_2 V_{v,s}(\pi_k^i(s),\pi_k^{-i}(s)),\mathrm{Ts}(q_k^{-i}(s))-\mathrm{Ts}(\mathcal{T}^{-i}(v^{-i})(s)\pi_k^i(s))\rangle}_{\hat{N}_3}\\ &\qquad+\underbrace{\beta_k\langle\nabla_1 V_{v,s}(\pi_k^i(s),\pi_{k+1}^{-i}(s)),\mathrm{Ts}(q_k^{i}(s))-\mathrm{Ts}(\mathcal{T}^{i}(v^{i})(s)\pi_k^{-i}(s))\rangle}_{\hat{N}_4}\\ &\qquad+2L_\eta\beta_k^2.\end{aligned} \tag{18}$$

We next bound the terms $\{\hat{N}_j\}_{1\leq j\leq 4}$ on the r.h.s of Equation 18 using Lemma. C.4 (3) and (4).

First consider $\hat{N}_1+\hat{N}_2$. We have by Lemma. C.4 (3) that

$$\hat{N}_1+\hat{N}_2\leq\,-\frac{7\beta_k}{8}V_{v,s}(\pi_k^i(s),\pi_k^{-i}(s))+16\beta_k\eta\|\mathcal{T}^i(v^i)(s)+\mathcal{T}^{-i}(v^{-i}(s)^\top\|_2^2.$$

To proceed, note that for any $\pi^{-i} \in \mathbb{R}^{|\mathcal{A}^{-i}|}$ satisfying $\|\pi^{-i}\|_2 = 1$, we have

$$\begin{aligned}
\|(\mathcal{T}^i(v^i)(s)+&\mathcal{T}^{-i}(v^{-i}(s)^\top)\pi^i\|_2^2 \\
&=\gamma^2 \sum_{a^i} \left[\sum_{a^{-i}} \mathbb{E}[v^i(s_1) + v^{-i}(s_1) \mid s_0 = s, a_0^i = a^i, a_0^{-i} = a^{-i}]\pi^{-i}(a^{-i})\right]^2 \\
&\leq \gamma^2 \|v^i + v^{-i}\|_\infty^2 \sum_{a^i} \left[\sum_{a^{-i}} \pi^{-i}(a^{-i})\right]^2 \\
&\leq \gamma^2 \mathcal{A} \|v^i + v^{-i}\|_\infty^2,
\end{aligned}$$

which implies

$$\|\mathcal{T}^i(v^i)(s) + \mathcal{T}^{-i}(v^{-i}(s)^\top\|_2^2 \leq \gamma^2 \mathcal{A} \|v^i + v^{-i}\|_\infty^2.$$

It follows that

$$\hat{N}_1 + \hat{N}_2 \leq -\frac{7\beta_k}{8} V_{v,s}(\pi_k^i(s), \pi_k^{-i}(s)) + 16\mathcal{A}\beta_k\eta\|v^i + v^{-i}\|_\infty^2.$$

We next consider $\hat{N}_3 + \hat{N}_4$. Since

$$\max(\|\mathcal{T}^i(v^i)(s)\|_2, \|\mathcal{T}^{-i}(v^{-i})(s)\|_2) \leq \frac{\mathcal{A}}{1-\gamma}, \qquad \text{(See Equation 17)}$$

we have by Lemma. C.4 (4) that

$$\begin{aligned}
\hat{N}_3 + \hat{N}_4 \leq 2\beta_k \left(\frac{1}{\eta \ell_\eta^{3/2}} + \frac{\mathcal{A}}{1-\gamma}\right) \Bigg[& 2\bar{c}\eta V_{v,s}(\pi_k^i(s), \pi_k^{-i}(s)) \\
&+ \frac{8\eta^2\mathcal{A}^2}{\bar{c}} \|q_k^i(s) - \mathcal{T}^i(v^i)(s)\pi_k^{-i}(s)\|_2^2 + \frac{8\eta^2\mathcal{A}^2}{\bar{c}} \|q_k^{-i}(s) - \mathcal{T}^{-i}(v^{-i})(s)\pi_k^i(s)\|_2^2 \Bigg]
\end{aligned}$$

for any $\bar{c} > 0$. By choosing $\bar{c} = \frac{1}{32\eta}(\frac{1}{\eta\ell_\eta^{3/2}} + \frac{\mathcal{A}}{1-\gamma})^{-1}$, we have that

$$\begin{aligned}
\hat{N}_3 + \hat{N}_4 \leq & \frac{\beta_k}{8} V_{v,s}(\pi_k^i(s), \pi_k^{-i}(s)) \\
&+ 512\beta_k \left(\frac{1}{\eta\ell_\eta^{3/2}} + \frac{\mathcal{A}}{1-\gamma}\right)^2 \eta^3 \mathcal{A}^2 \|q_k^i(s) - \mathcal{T}^i(v^i)(s)\pi_k^{-i}(s)\|_2^2 \\
&+ 512\beta_k \left(\frac{1}{\eta\ell_\eta^{3/2}} + \frac{\mathcal{A}}{1-\gamma}\right)^2 \eta^3 \mathcal{A}^2 \|q_k^{-i}(s) - \mathcal{T}^{-i}(v^{-i})(s)\pi_k^i(s)\|_2^2 \\
\leq & \frac{\beta_k}{8} V_{v,s}(\pi_k^i(s), \pi_k^{-i}(s)) + \frac{2048\mathcal{A}^4\beta_k\eta^3}{\ell_\eta^3(1-\gamma)^2} \|q_k^i(s) - \mathcal{T}^i(v^i)(s)\pi_k^{-i}(s)\|_2^2 \\
&+ \frac{2048\mathcal{A}^4\beta_k\eta^3}{\ell_\eta^3(1-\gamma)^2} \|q_k^{-i}(s) - \mathcal{T}^{-i}(v^{-i})(s)\pi_k^i(s)\|_2^2.
\end{aligned}$$

Finally, using the upper bounds we obtained for the terms $\hat{N}_1 + \hat{N}_2$ and $\hat{N}_3 + \hat{N}_4$ in Equation 18

$$\begin{aligned}
V_{v,s}(\pi_{k+1}^i(s), \pi_{k+1}^{-i}(s)) \leq & \left(1 - \frac{3\beta_k}{4}\right) V_{v,s}(\pi_k^i(s), \pi_k^{-i}(s)) + 16\mathcal{A}\beta_k\eta\|v^i + v^{-i}\|_\infty^2. \\
&+ \frac{2048\mathcal{A}^4\beta_k\eta^3}{\ell_\eta^3(1-\gamma)^2} \|q_k^i(s) - \mathcal{T}^i(v^i)(s)\pi_k^{-i}(s)\|_2^2 \\
&+ \frac{2048\mathcal{A}^4\beta_k\eta^3}{\ell_\eta^3(1-\gamma)^2} \|q_k^{-i}(s) - \mathcal{T}^{-i}(v^{-i})(s)\pi_k^i(s)\|_2^2 \\
&+ \frac{4\mathcal{A}^2}{\ell_\eta(1-\gamma)^2}\beta_k^2.
\end{aligned}$$

Summing up both sides of the previous inequality for all s and then taking expectation, we deduce the desired result. □

C.2.2. Analyzing the q-function update

Consider q_k^i generated by the algorithm. We begin by reformulating the update of the q-function as a stochastic approximation algorithm for estimating a time-varying target. Let $F^i : \mathbb{R}^{|\mathcal{S}||\mathcal{A}^i|} \times \mathcal{S} \times \mathcal{A}^i \times \mathcal{A}^{-i} \times \mathcal{S} \mapsto \mathbb{R}^{|\mathcal{S}||\mathcal{A}^i|}$ be an operator defined as

$$[F^i(q^i, s_0, a_0^i, a_0^{-i}, s_1)](s, a^i) = \mathbb{1}_{\{(s,a^i)=(s_0,a_0^i)\}} \left(\mathcal{R}^i(s_0, a_0^i, a_0^{-i}) + \gamma v^i(s_1) - q^i(s_0, a_0^i)\right)$$

for all $(q^i, s_0, a_0^i, a_0^{-i}, s_1)$ and (s, a^i). Then the q–function update can be compactly written as

$$q_{k+1}^i = q_k^i + \alpha_k F^i(q_k^i, S_k, A_k^i, A_k^{-i}, S_{k+1}). \tag{19}$$

Denote the stationary distribution of the Markov chain $\{S_k\}$ induced by the joint policy $\pi_k = (\pi_k^i, \pi_k^{-i})$ by $\pi_k \in \Delta^{|\mathcal{S}|}$, the existence and uniqueness of which is guaranteed by Lemma 4.1 and Lemma 4 (1) of (Chen et al., 2023) (same as Lemma 4 in (Zhang et al., 2023)). Let $\bar{F}_k^i : \mathbb{R}^{|\mathcal{S}||\mathcal{A}^i|} \mapsto \mathbb{R}^{|\mathcal{S}||\mathcal{A}^i|}$ be defined as

$$\bar{F}_k^i(q^i) = \mathbb{E}_{s_0\sim\pi_k(\cdot),a_0^i\sim\pi_k^i(\cdot|s_0),a_k^{-i}\sim\pi_k^{-i}(\cdot|s_0),S_1\sim p(\cdot|s_0,a_0^i,a_0^{-i})} \left[F^i(q^i, s_0, a_0^i, a_0^{-i}, s_1)\right]$$

for all $q^i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}^i|}$. Then Equation 19 can be viewed as a stochastic approximation algorithm for solving the (time-varying) equation $\bar{F}_k^i(q^i) = 0$ with time-inhomogeneous Markovian noise $\{(S_k, A_k^i, A_k^{-i}, S_{k+1})\}_{k\geq 0}$. We next establish the properties of the operators $F^i(\cdot)$ and $\bar{F}_k^i(\cdot)$ in the following lemma.

Lemma C.6. *The following inequalities hold:*

1. *for all (q_1^i, q_2^i) and $(s_0, a_0^i, a_0^{-i}, s_1)$:*
$$\|F^i(q_1^i, s_0, a_0^i, a_0^{-i}, s_1) - F^i(q_2^i, s_0, a_0^i, a_0^{-i}, s_1)\|_2 \leq \|q_1^i - q_2^i\|_2.$$
2. *for all $(s_0, a_0^i, a_0^{-i}, s_1)$: $\|F^i(\mathbf{0}, s_0, a_0^i, a_0^{-i}, s_1)\|_2 \leq \frac{1}{1-\gamma}$.*
3. *$\bar{F}_k^i(q^i) = 0$ has a unique solution $\bar{q}_k^i$, which is explicitly given as $\bar{q}_k^i(s) = \mathcal{T}^i(v^i)(s)\pi_k^{-i}(s)$ for all s.*
4. *$\langle \bar{F}_k^i(q_1^i) - \bar{F}_k^i(q_2^i), q_1^i - q_2^i\rangle \leq -c_\eta \|q_1^i - q_2^i\|_2^2$ for all (q_1^i, q_2^i).*

Using $\|\cdot\|_2^2$ as a Lyapunov function and the equivalent update equation (19), we obtain

$$\begin{aligned}
\mathbb{E}[\|q_{k+1}^i - \bar{q}_{k+1}^i\|_2^2] &= \mathbb{E}[\|q_{k+1}^i - q_k^i + q_k^i - \bar{q}_k^i + \bar{q}_k^i - \bar{q}_{k+1}^i\|_2^2] \\
&= \mathbb{E}[\|q_k^i - \bar{q}_k^i\|_2^2] + \mathbb{E}[\|q_{k+1}^i - q_k^i\|_2^2] + \mathbb{E}[\|\bar{q}_k^i - \bar{q}_{k+1}^i\|_2^2] \\
&\quad + \alpha_k \mathbb{E}[\langle F^i(q_k^i, S_k, A_k^i, A_k^{-i}, S_{k+1}), q_k^i - \bar{q}_k^i\rangle] + \mathbb{E}[\langle q_{k+1}^i - q_k^i, \bar{q}_k^i - \bar{q}_{k+1}^i\rangle] \\
&\quad + \mathbb{E}[\langle q_k^i - \bar{q}_k^i, \bar{q}_k^i - \bar{q}_{k+1}^i\rangle] \\
&= \mathbb{E}[\|q_k^i - \bar{q}_k^i\|_2^2] + \alpha_k \underbrace{\mathbb{E}[\langle \bar{F}_k^i(q_k^i), q_k^i - \bar{q}_k^i\rangle]}_{N_1} \\
&\quad + \alpha_k \underbrace{\mathbb{E}[\langle F^i(q_k^i, S_k, A_k^i, A_k^{-i}, S_{k+1}) - \bar{F}_k^i(q_k^i), q_k^i - \bar{q}_k^i\rangle]}_{N_2} \\
&\quad + \mathbb{E}[\|q_{k+1}^i - q_k^i\|_2^2] + \mathbb{E}[\|\bar{q}_k^i - \bar{q}_{k+1}^i\|_2^2] \\
&\quad + \mathbb{E}[\langle q_{k+1}^i - q_k^i, \bar{q}_k^i - \bar{q}_{k+1}^i\rangle] + \mathbb{E}[\langle q_k^i - \bar{q}_k^i, \bar{q}_k^i - \bar{q}_{k+1}^i\rangle].
\end{aligned} \tag{20}$$

What remains to do is to bound the terms on the r.h.s of the previous inequality. Among them, we want to highlight the two terms N_1 and N_2. For the term N_1, using Lemma C.6 (4), we have

$$N_1 = \mathbb{E}[\langle \bar{F}_k^i(q_k^i), q_k^i - \bar{q}_k^i\rangle] = \mathbb{E}[\langle \bar{F}_k^i(q_k^i) - \bar{F}_k^i(\bar{q}_k^i), q_k^i - \bar{q}_k^i\rangle] \leq -c_\eta \mathbb{E}[\|q_k^i - \bar{q}_k^i\|_2^2], \tag{21}$$

which provides us with the desired drift inequality.

The term N_2 is the difference between $F^i(q_k^i, S_k, A_k^i, A_k^{-i}, S_{k+1})$ and its expected value $\bar{F}_k^i(q_k^i)$. To analyze the latter, under the time-inhomogeneous Markov chain $\{(S_k, A_k^i, A_k^{-i}, S_{k+1})\}$, we note that the policy is updated slower than the q-functions; and the stationary distribution is Lipschitz in the underlying policy (Lemma B.4). This reasoning is at the core of the proof of this result.

Lemma C.7 (Noise term (Lemma A.12 (Chen et al., 2023))). *When $\alpha_{k-z_k,k-1} \leq 1/4$ for all $k \geq z_k$, we have for all $k \geq z_k$ that*

$$N_2 \leq \frac{340|\mathcal{S}|^{3/2}\mathcal{A}^{3/2}\hat{L}_\eta}{(1-\gamma)^2} z_k \alpha_{k-z_k,k-1}.$$

For our choice of α_k and β_k, we find $\lim_{k\to\infty} z_k\alpha_{k-z_k,k-1} = 0$. Therefore, Lemma C.7 implies that the term N_2 vanishes as a function of the episode length.

We next bound the rest of the terms on the r.h.s of Equation 20 in the following lemma.

Lemma C.8 (Other terms). *The following inequalities hold for all $k \geq 0$.*

1. $\mathbb{E}[\|q_{k+1}^i - q_k^i\|_2^2] \leq \frac{4|\mathcal{S}|\mathcal{A}\alpha_k^2}{(1-\gamma)^2}$.
2. $\mathbb{E}[\|\bar{q}_k^i - \bar{q}_{k+1}^i\|_2^2] \leq \frac{4|\mathcal{S}|\mathcal{A}\beta_k^2}{(1-\gamma)^2}$.
3. $\mathbb{E}[\langle q_{k+1}^i - q_k^i, \bar{q}_k^i - \bar{q}_{k+1}^i\rangle] \leq \frac{4|\mathcal{S}|\mathcal{A}\alpha_k\beta_k}{(1-\gamma)^2}$.
4. $\mathbb{E}[\langle q_k^i - \bar{q}_k^i, \bar{q}_k^i - \bar{q}_{k+1}^i\rangle] \leq \frac{17\eta\mathcal{A}^2\beta_k}{(1-\gamma)^2}\mathbb{E}[\|q_k^i - \bar{q}_k^i\|_2^2] + \frac{\beta_k}{16}\sum_s \mathbb{E}[V_{v,s}(\pi_k^i(s), \pi_k^{-i}(s))]$.

Proof. This lemma was proved in (Chen et al., 2023), we only provide the proof of the fourth part as it is slightly different.

4) For any $k \geq 0$, we have

$$\begin{aligned}\langle q_k^i - \bar{q}_k^i, \bar{q}_k^i - \bar{q}_{k+1}^i\rangle &= \beta_k \sum_s \langle q_k^i(s) - \bar{q}_k^i(s), \mathcal{T}^i(v^i)(s)(\mathrm{Ts}(q_k^{-i}(s)) - \pi_k^{-i}(s))\rangle \\ &\leq \beta_k\left(\frac{\hat{c}\|q_k^i - \bar{q}_k^i\|_2^2}{2} + \frac{\sum_s \|\mathcal{T}^i(v^i)(s)(\mathrm{Ts}(q_k^{-i}(s)) - \pi_k^{-i}(s))\|_2^2}{2\hat{c}}\right),\end{aligned} \tag{22}$$

where $\hat{c}$ is an arbitrary positive real number. We next analyze the second term on the r.h.s of the previous inequality. For any $s \in \mathcal{S}$, we have

$$\begin{aligned}&\|\mathcal{T}^i(v^i)(s)(\mathrm{Ts}(q_k^{-i}(s)) - \pi_k^{-i}(s))\|_2 \\ &\quad= \|\mathcal{T}^i(v^i)(s)(\mathrm{Ts}(q_k^{-i}(s)) - \mathrm{Ts}(\bar{q}_k^{-i}(s)) + \mathrm{Ts}(\mathcal{T}^{-i}(v^{-i})(s)\pi_k^i(s)) - \pi_k^{-i}(s))\|_2 \\ &\quad\leq \underbrace{\|\mathcal{T}^i(v^i)(s)(\mathrm{Ts}(q_k^{-i}(s)) - \mathrm{Ts}(\bar{q}_k^{-i}(s)))\|_2}_{B_1} \\ &\qquad + \underbrace{\|\mathcal{T}^i(v^i)(s)(\mathrm{Ts}(\mathcal{T}^{-i}(v^{-i})(s)\pi_k^i(s)) - \pi_k^{-i}(s))\|_2}_{B_2}.\end{aligned}$$

Since Tsallis–smoothing $\mathrm{Ts}(\cdot)$ is $2\sqrt{2}\eta\mathcal{A}$ – Lipschitz continuous with respect to $\|\cdot\|_2$, we have

$$\begin{aligned}B_1 &\leq \|\mathcal{T}^i(v^i)(s)\|_2\|\,\mathrm{Ts}(q_k^{-i}(s)) - \mathrm{Ts}(\bar{q}_k^{-i}(s))\|_2 \\ &\leq \frac{2\sqrt{2}\eta\mathcal{A}^2}{1-\gamma}\|q_k^{-i}(s) - \bar{q}_k^{-i}(s)\|_2.\end{aligned}$$

We next analyze the term B_2. Using the quadratic growth property of strongly convex functions, we obtain

$$\begin{aligned}B_2 &= \|\mathcal{T}^i(v^i)(s)(\mathrm{Ts}(\mathcal{T}^{-i}(v^{-i})(s)\pi_k^i(s)) - \pi_k^{-i}(s))\|_2 \\ &\leq \|\mathcal{T}^i(v^i)(s)\|_2\|\,\mathrm{Ts}(\mathcal{T}^{-i}(v^{-i})(s)\pi_k^i(s)) - \pi_k^{-i}(s)\|_2 \\ &\leq \frac{\sqrt{2\eta}\mathcal{A}}{1-\gamma}V_{v,s}^{1/2}(\pi_k^i(s), \pi_k^{-i}(s)).\end{aligned}$$

Combining the upper bounds we obtained for the terms B_1 and B_2

$$\begin{aligned}
&\sum_s \|\mathcal{T}^i(v^i)(s)(\mathrm{Ts}(q_k^{-i}(s)) - \pi_k^{-i}(s))\|_2^2 \\
&\qquad \leq \sum_s (B_1 + B_2)^2 \\
&\qquad \leq 2\sum_s (B_1^2 + B_2^2) \\
&\qquad \leq 2\sum_s \left(\frac{8\mathcal{A}^4\eta^2}{(1-\gamma)^2}\|q_k^{-i}(s)) - \bar{q}_k^{-i}(s)\|_2^2 + \frac{2\eta\mathcal{A}^2}{(1-\gamma)^2} V_{v,s}(\pi_k^i(s), \pi_k^{-i}(s)) \right) \\
&\qquad = \frac{16\eta^2\mathcal{A}^4}{(1-\gamma)^2}\|q_k^{-i} - \bar{q}_k^{-i}\|_2^2 + \frac{4\eta\mathcal{A}^2}{(1-\gamma)^2}\sum_s V_{v,s}(\pi_k^i(s), \pi_k^{-i}(s)).
\end{aligned}$$

Coming back to Equation 22 and using the previous inequality

$$\begin{aligned}
&\langle q_k^i - \bar{q}_k^i, \bar{q}_k^i - \bar{q}_{k+1}^i \rangle \\
&\leq \beta_k \left(\frac{\hat{c}\|q_k^i - \bar{q}_k^i\|_2^2}{2} + \frac{\sum_s \|\mathcal{T}^i(v^i)(s)(\mathrm{Ts}(q_k^{-i}(s)) - \pi_k^{-i}(s))\|_2^2}{2\hat{c}} \right) \\
&\leq \beta_k \left(\frac{\hat{c}\|q_k^i - \bar{q}_k^i\|_2^2}{2} + \frac{8\eta^2\mathcal{A}^4}{\hat{c}(1-\gamma)^2}\|q_k^{-i} - \bar{q}_k^{-i}\|_2^2 + \frac{2\eta\mathcal{A}^2}{\hat{c}(1-\gamma)^2}\sum_s V_{v,s}(\pi_k^i(s), \pi_k^{-i}(s)) \right).
\end{aligned}$$

Choosing $\hat{c} = \frac{32\eta\mathcal{A}^2}{(1-\gamma)^2}$ in the previous inequality and then taking total expectation

$$\mathbb{E}[\langle q_k^i - \bar{q}_k^i, \bar{q}_k^i - \bar{q}_{k+1}^i \rangle] \leq \frac{17\eta\mathcal{A}^2\beta_k}{(1-\gamma)^2}\mathbb{E}[\|q_k^i - \bar{q}_k^i\|_2^2] + \frac{\beta_k}{16}\sum_s \mathbb{E}[V_{v,s}(\pi_k^i(s), \pi_k^{-i}(s))].$$

□

Since we obtained upper bounds on all the terms on the r.h.s of Equation 20, we deduce the one-step Lyapunov drift inequality for q_k^i. The same analysis also entails a drift inequality for q_k^{-i}. Both results are presented in the following lemma.

Lemma C.9 (Lyapunov drift for q–functions ((Chen et al., 2023), Lemma A.12)). *For all $k \geq z_k$ and $i \in \{1, 2\}$:*

$$\begin{aligned}
\mathbb{E}[\|q_{k+1}^i - \bar{q}_{k+1}^i\|_2^2] \leq &\left(1 - c_\eta\alpha_k + \frac{17\eta\mathcal{A}^2\beta_k}{(1-\gamma)^2}\right)\mathbb{E}[\|q_k^i - \bar{q}_k^i\|_2^2] \\
&+ \frac{352|\mathcal{S}|^{3/2}\mathcal{A}^{3/2}\hat{L}_\eta}{(1-\gamma)^2} z_k\alpha_k\alpha_{k-z_k,k-1} + \frac{\beta_k}{16}\sum_s \mathbb{E}[V_{v,s}(\pi_k^i(s), \pi_k^{-i}(s))].
\end{aligned}$$

where $z_k = t(\ell_\eta, \beta_k)$ is a uniform upper bound on the uniform mixing time with accuracy β_k, see Equation 8).

C.3. Bounding Coupled Drift Inequalities

We first restate the drift inequalities from previous sections. Recall the notations: $\mathcal{V}_q(t,k) = \sum_{i=1,2}\|q_{t,k}^i - \bar{q}_{t,k}^i\|_2^2$, $\mathcal{V}_\pi(t,k) = \sum_s V_{v_t,s}(\pi_{t,k}^i(s), \pi_{t,k}^{-i}(s))$, and $\mathcal{F}_t$ as the history of the algorithm right before the t-th outer-loop iteration. Note that v_t^i and v_t^{-i} are both measurable with respect to $\mathcal{F}_t$. In what follows, we denote $\mathbb{E}_t[\,\cdot\,]$ for $\mathbb{E}[\,\cdot \mid \mathcal{F}_t]$.

- **Lemma C.2:** It holds for all $t \geq 0$ and $i = 1, 2$ that

$$\begin{aligned}
\|v_{t+1}^i - v_*^i\|_\infty \leq \gamma\|v_t^i - v_*^i\|_\infty &+ 2\max_{s\in\mathcal{S}} V_{v_t,s}(\pi_{t,K}^i(s), \pi_{t,K}^{-i}(s)) + \frac{16}{\eta}\sqrt{\mathcal{A}} \\
&+ \max_{s\in\mathcal{S}}\|\bar{q}_{t,K}^i(s) - q_{t,K}^i(s)\|_\infty + 2\gamma\|v_t^i + v_t^{-i}\|_\infty,
\end{aligned} \tag{23}$$

  where $\bar{q}_{t,K}^i(s) := \mathcal{T}^i(v_t^i)(s)\pi_{t,K}^{-i}(s)$ for all $s \in \mathcal{S}$.

- **Lemma C.3:** It holds for all $t \geq 0$ that

$$\|v_{t+1}^i + v_{t+1}^{-i}\|_\infty \leq \gamma \|v_t^i + v_t^{-i}\|_\infty + \sum_{i=1,2} \|q_{t,K}^i - \bar{q}_{t,K}^i\|_2. \tag{24}$$

- **Lemma C.5:** It holds for all $t, k \geq 0$ that

$$\begin{aligned}\mathbb{E}_t[\mathcal{V}_\pi(t, k+1)] \leq& \left(1 - \frac{3\beta_k}{4}\right) \mathbb{E}_t[\mathcal{V}_\pi(t,k)] + \frac{2048\mathcal{A}^4 \beta_k \eta^3}{\ell_\eta^3 (1-\gamma)^2} \mathbb{E}_t[\mathcal{V}_q(t,k)] \\ &+ 16|\mathcal{S}|\mathcal{A}\beta_k \eta \|v_t^i + v_t^{-i}\|_\infty^2 + \frac{4|\mathcal{S}|\mathcal{A}^2 \beta_k^2}{\ell_\eta (1-\gamma)^2}.\end{aligned} \tag{25}$$

- **Lemma C.9:** It holds for all $t \geq 0$ and $k \geq z_k$ that

$$\begin{aligned}\mathbb{E}_t[\mathcal{V}_q(t, k+1)] \leq& \left(1 - c_\eta \alpha_k + \frac{17\eta \mathcal{A}^2 \beta_k}{(1-\gamma)^2}\right) \mathbb{E}_t[\mathcal{V}_q(t,k)] \\ &+ \frac{\beta_k}{16} \mathbb{E}_t[\mathcal{V}_\pi(t,k)] + \frac{352|\mathcal{S}|^{3/2} \mathcal{A}^{3/2} \hat{L}_\eta}{(1-\gamma)^2} z_k \alpha_k \alpha_{k-z_k,k-1}.\end{aligned} \tag{26}$$

Adding up equations (25) and (26) entails

$$\begin{aligned}\mathbb{E}_t[\mathcal{V}_\pi(t,k+1) + \mathcal{V}_q(t,k+1)] \leq& \left(1 - \frac{\beta_k}{2}\right) \mathbb{E}_t[\mathcal{V}_\pi(t,k)] + \frac{4|\mathcal{S}|\mathcal{A}^2 \beta_k^2}{\ell_\eta (1-\gamma)^2} \\ &+ \left(1 - c_\eta \alpha_k + \frac{3136\mathcal{A}^4 \beta_k \eta^3}{\ell_\eta^3 (1-\gamma)^2}\right) \mathbb{E}_t[\mathcal{V}_q(t,k)] + 16|\mathcal{S}|\mathcal{A}\beta_k \eta \|v_t^i + v_t^{-i}\|_\infty^2 \\ &+ \frac{352|\mathcal{S}|^{3/2} \mathcal{A}^{3/2} \hat{L}_\eta}{(1-\gamma)^2} z_k \alpha_k \alpha_{k-z_k,k-1} \\ =& \left(1 - \frac{c_{\alpha,\beta}\alpha_k}{2}\right) \mathbb{E}_t[\mathcal{V}_\pi(t,k)] + \left(1 - c_\eta \alpha_k + \frac{3136\mathcal{A}^4 c_{\alpha,\beta} \alpha_k \eta^3}{\ell_\eta^3 (1-\gamma)^2}\right) \mathbb{E}_t[\mathcal{V}_q(t,k)] \\ &+ 16|\mathcal{S}|\mathcal{A}\beta_k \eta \|v_t^i + v_t^{-i}\|_\infty^2 + \frac{4|\mathcal{S}|\mathcal{A}^2 \beta_k^2}{\ell_\eta (1-\gamma)^2} \\ &+ \frac{352|\mathcal{S}|^{3/2} \mathcal{A}^{3/2} \hat{L}_\eta}{(1-\gamma)^2} z_k \alpha_k \alpha_{k-z_k,k-1}.\end{aligned}$$

Note that Condition 1 implies that

$$\frac{3136\mathcal{A}^4 c_{\alpha,\beta} \eta^3}{\ell_\eta^3 (1-\gamma)^2} \leq \frac{c_\eta}{2}.$$

Therefore, we have

$$\begin{aligned}\mathbb{E}_t[\mathcal{V}_\pi(t,k+1) + \mathcal{V}_q(t,k+1)] \leq& \frac{4|\mathcal{S}|\mathcal{A}^2 c_{\alpha,\beta}^2 \alpha_k^2}{\ell_\eta (1-\gamma)^2} + \frac{352|\mathcal{S}|^{3/2} \mathcal{A}^{3/2} \hat{L}_\eta}{(1-\gamma)^2} z_k \alpha_k \alpha_{k-z_k,k-1} \\ &+ \left(1 - \frac{c_{\alpha,\beta}\alpha_k}{2}\right) \mathbb{E}_t[\mathcal{V}_\pi(t,k) + \mathcal{V}_q(t,k)] \qquad (27) \\ &+ 16|\mathcal{S}|\mathcal{A} c_{\alpha,\beta} \alpha_k \eta \|v_t^i + v_t^{-i}\|_\infty^2. \qquad (28)\end{aligned}$$

Linearly decaying rates We consider the following parameter choice: $\alpha_k = \frac{\alpha}{k+h}$, $\beta_k = \frac{\beta}{k+h}$, and $\beta = c_{\alpha,\beta}\alpha$. Iteratively applying Equation 27, we obtain for all $k \geq k_0$

$$\begin{aligned}\mathbb{E}_t[\mathcal{V}_\pi(t,k) + \mathcal{V}_q(t,k)] \lesssim & \frac{4|\mathcal{S}|\mathcal{A}}{(1-\gamma)^2} \underbrace{\prod_{m=k_0}^{k-1}\left(1 - \frac{c_{\alpha,\beta}\alpha_m}{2}\right)}_{\hat{\mathcal{E}}_1} \\ & + \frac{|\mathcal{S}|^{3/2}\mathcal{A}^2\hat{L}_\eta}{(1-\gamma)^2} \underbrace{\sum_{n=k_0}^{k-1} z_n^2\alpha_n^2 \prod_{m=n+1}^{k-1}\left(1 - \frac{c_{\alpha,\beta}\alpha_m}{2}\right)}_{\hat{\mathcal{E}}_2} \\ & + |\mathcal{S}|\mathcal{A}c_{\alpha,\beta}\eta\|v_t^i + v_t^{-i}\|_\infty^2 \underbrace{\sum_{n=k_0}^{k-1} \alpha_n \prod_{m=n+1}^{k-1}\left(1 - \frac{c_{\alpha,\beta}\alpha_m}{2}\right)}_{\hat{\mathcal{E}}_3}.\end{aligned}$$

We next provide estimates for the terms $\{\hat{\mathcal{E}}_j\}_{1\leq j\leq 3}$. Bounds of terms like $\{\hat{\mathcal{E}}_j\}_{1\leq j\leq 3}$ are well-established in existing work studying the convergence rate of iterative algorithms. Specifically, we have that

$$\hat{\mathcal{E}}_1 \leq \left(\frac{k_0+h}{k+h}\right)^{c_{\alpha,\beta}\alpha/2}, \quad \hat{\mathcal{E}}_2 \leq \frac{4ez_k^2\alpha^2}{c_{\alpha,\beta}\alpha/2-1}\frac{1}{k+h}, \text{ and } \hat{\mathcal{E}}_3 \leq \frac{2}{c_{\alpha,\beta}}.$$

It follows that

$$\begin{aligned}\mathbb{E}_t[\mathcal{V}_\pi(t,k) + \mathcal{V}_q(t,k)] \lesssim & \frac{|\mathcal{S}|\mathcal{A}}{(1-\gamma)^2}\left(\frac{k_0+h}{k+h}\right)^{c_{\alpha,\beta}\alpha/2} + \frac{|\mathcal{S}|^{3/2}\mathcal{A}^2\hat{L}_\eta}{(1-\gamma)^2}\frac{z_k^2\alpha^2}{c_{\alpha,\beta}\alpha/2-1}\frac{1}{k+h} \\ & + |\mathcal{S}|\mathcal{A}\eta\|v_t^i + v_t^{-i}\|_\infty^2 \\ \lesssim & \frac{|\mathcal{S}|\mathcal{A}}{(1-\gamma)^2}\left(\frac{\alpha_k}{\alpha_{k_0}}\right)^{c_{\alpha,\beta}\alpha/2} + \frac{|\mathcal{S}|^{3/2}\mathcal{A}^2\hat{L}_\eta}{(1-\gamma)^2}\frac{z_k^2\alpha^2}{c_{\alpha,\beta}\alpha/2-1}\frac{1}{k+h} \\ & + |\mathcal{S}|\mathcal{A}\eta\|v_t^i + v_t^{-i}\|_\infty^2,\end{aligned} \tag{29}$$

which implies

$$\begin{aligned}\mathbb{E}_t[\mathcal{V}_\pi(t,k)] \lesssim & \frac{|\mathcal{S}|\mathcal{A}}{(1-\gamma)^2}\left(\frac{\alpha_k}{\alpha_{k_0}}\right)^{c_{\alpha,\beta}\alpha/2} + \frac{|\mathcal{S}|^{3/2}\mathcal{A}^2\hat{L}_\eta}{(1-\gamma)^2}\frac{z_k^2\alpha^2}{c_{\alpha,\beta}\alpha/2-1}\frac{1}{k+h} \\ & + |\mathcal{S}|\mathcal{A}\eta\|v_t^i + v_t^{-i}\|_\infty^2.\end{aligned}$$

Using the previous bound on $\mathbb{E}_t[\mathcal{V}_\pi(t,k)]$ in Equation 26

$$\begin{aligned}\mathbb{E}_t[\mathcal{V}_q(t,k+1)] \leq & \left(1 - c_\eta\alpha_k + \frac{17\eta\mathcal{A}^2\beta_k}{(1-\gamma)^2}\right)\mathbb{E}_t[\mathcal{V}_q(t,k)] \\ & + \frac{\beta_k}{16}\mathbb{E}_t[\mathcal{V}_\pi(t,k)] + \frac{352|\mathcal{S}|^{3/2}\mathcal{A}^{3/2}\hat{L}_\eta}{(1-\gamma)^2}z_k\alpha_k\alpha_{k-z_k,k-1} \\ \lesssim & \left(1 - \frac{c_\eta\alpha_k}{2}\right)\mathbb{E}_t[\mathcal{V}_q(t,k)] + \frac{|\mathcal{S}|^{3/2}\mathcal{A}^2}{\alpha_{k_0}(1-\gamma)^2}z_k^2\alpha_k^2 \\ & + |\mathcal{S}|\mathcal{A}c_{\alpha,\beta}\alpha_k\eta\|v_t^i + v_t^{-i}\|_\infty^2.\end{aligned}$$

Repeatedly using the previous inequality starting from k_0

$$\begin{aligned}\mathbb{E}_t[\mathcal{V}_q(t,k)] \lesssim & \frac{|\mathcal{S}|\mathcal{A}}{(1-\gamma)^2}\left(\frac{\alpha_k}{\alpha_{k_0}}\right)^{c_\eta\alpha/2} + \frac{|\mathcal{S}|^{3/2}\mathcal{A}^2}{\alpha_{k_0}(1-\gamma)^2}z_k^2\alpha_k + \frac{|\mathcal{S}|\mathcal{A}c_{\alpha,\beta}\eta}{c_\eta}\|v_t^i + v_t^{-i}\|_\infty^2 \\ \lesssim & \frac{|\mathcal{S}|^{3/2}\mathcal{A}^2}{\alpha_{k_0}(1-\gamma)^2}z_k^2\alpha_k + \frac{|\mathcal{S}|\mathcal{A}c_{\alpha,\beta}\eta}{c_\eta}\|v_t^i + v_t^{-i}\|_\infty^2\end{aligned}$$

Since $\sum_{i=1,2} \mathbb{E}_t \left[\|q^i_{t,K} - \bar{q}^i_{t,K}\|_2\right] \lesssim \mathbb{E}_t[\mathcal{V}_q(t,K)]^{1/2}$, we have

$$\sum_{i=1,2} \mathbb{E}_t \left[\|q^i_{t,K} - \bar{q}^i_{t,K}\|_2\right] \le \frac{c_1' |\mathcal{S}|^{3/4}\mathcal{A}}{\alpha_{k_0}^{1/2}(1-\gamma)} z_k \alpha_k^{1/2} + \frac{c_2' \sqrt{|\mathcal{S}|\mathcal{A}} c_{\alpha,\beta}^{1/2}\eta^{1/2}}{c_\eta^{1/2}} \|v^i_t + v^{-i}_t\|_\infty, \tag{30}$$

where c_1' and c_2' are numerical constants. Applying the total expectation for both sides of the previous inequality and using Equation 24, we obtain

$$\begin{aligned}\mathbb{E}[\|v^i_{t+1} + v^{-i}_{t+1}\|_\infty] &\le \left(\gamma + \frac{c_2' \sqrt{|\mathcal{S}|\mathcal{A}} c_{\alpha,\beta}^{1/2}\eta^{1/2}}{c_\eta^{1/2}}\right) \mathbb{E}[\|v^i_t + v^{-i}_t\|_\infty] + \frac{c_1' |\mathcal{S}|^{3/4}\mathcal{A}}{\alpha_{k_0}^{1/2}(1-\gamma)} z_k \alpha_k^{1/2} \\ &\le \left(\frac{\gamma+1}{2}\right) \mathbb{E}[\|v^i_t + v^{-i}_t\|_\infty] + \frac{c_1' |\mathcal{S}|^{3/4}\mathcal{A}}{\alpha_{k_0}^{1/2}(1-\gamma)} z_k \alpha_k^{1/2},\end{aligned}$$

where the last line follows from Condition 1. Repeatedly using the previous inequality starting from 0

$$\mathbb{E}[\|v^i_t + v^{-i}_t\|_\infty] \lesssim \frac{2}{1-\gamma}\left(\frac{\gamma+1}{2}\right)^t + \frac{|\mathcal{S}|^{3/4}\mathcal{A}}{\alpha_{k_0}^{1/2}(1-\gamma)^2} z_k \alpha_k^{1/2}. \tag{31}$$

The next step is to bound $\|v^i_t - v^i_*\|_\infty$. Recall from Equation 23 that

$$\begin{aligned}\mathbb{E}[\|v^i_{t+1} - v^i_*\|_\infty] \le \gamma \mathbb{E}[\|v^i_t - v^i_*\|_\infty] + 2\gamma \mathbb{E}[\|v^i_t + v^{-i}_t\|_\infty] + \frac{16}{\eta}\sqrt{\mathcal{A}} \\ + 2\mathbb{E}[\mathcal{V}_\pi(t,K)] + 2\mathbb{E}[\mathcal{V}_q(t,K)]^{1/2}.\end{aligned}$$

Since equations 29 and 30 imply that

$$\begin{aligned}&\mathbb{E}[\|v^i_t + v^{-i}_t\|_\infty] + \mathbb{E}[\mathcal{V}_\pi(t,K)] + \mathbb{E}[\mathcal{V}_q(t,K)]^{1/2} \\ &\qquad \lesssim \frac{|\mathcal{S}|\mathcal{A}\eta}{(1-\gamma)^2}\left(\frac{\gamma+1}{2}\right)^t + \frac{|\mathcal{S}|^2\mathcal{A}^2\hat{L}_\eta}{\alpha_{k_0} c_{\alpha,\beta}(1-\gamma)^3} z_K^2 \alpha_K^{1/2},\end{aligned}$$

we have

$$\begin{aligned}\mathbb{E}[\|v^i_{t+1} - v^i_*\|_\infty] \le \gamma \mathbb{E}[\|v^i_t - v^i_*\|_\infty] + c'' \Bigg[\frac{|\mathcal{S}|\eta\mathcal{A}}{(1-\gamma)^2}\left(\frac{\gamma+1}{2}\right)^t + \frac{\sqrt{\mathcal{A}}}{\eta} \\ + \frac{|\mathcal{S}|^2\mathcal{A}^2\hat{L}_\eta}{\alpha_{k_0} c_{\alpha,\beta}(1-\gamma)^3} z_K^2 \alpha_K^{1/2} \Bigg]\end{aligned}$$

for some numerical constant c''. We use the previous inequality iteratively starting from 0 to time $T-1$ to find

$$\mathbb{E}[\|v^i_T - v^i_*\|_\infty] \lesssim \frac{|\mathcal{S}|\mathcal{A}T\eta}{(1-\gamma)^2}\left(\frac{\gamma+1}{2}\right)^{T-1} + \frac{\sqrt{\mathcal{A}}}{\eta(1-\gamma)} + \frac{|\mathcal{S}|^2\mathcal{A}^2\hat{L}_\eta}{\alpha_{k_0} c_{\alpha,\beta}(1-\gamma)^4} z_K^2 \alpha_K^{1/2}$$

Plugging the previous inequality with equations 29 and 31 in the Nash gap decomposition (lemma C.1), we obtain

$$\begin{aligned}\mathbb{E}[\|v^i_{*,\pi^{-i}_{T,K}} - v^i_{\pi^i_{T,K},\pi^{-i}_{T,K}}\|_\infty] \lesssim \frac{|\mathcal{S}|\mathcal{A}T\eta}{(1-\gamma)^3}\left(\frac{\gamma+1}{2}\right)^{T-1} + \frac{\sqrt{\mathcal{A}}}{\eta(1-\gamma)^2} \\ + \frac{|\mathcal{S}|^2\mathcal{A}^2\hat{L}_\eta}{\alpha_{k_0} c_{\alpha,\beta}(1-\gamma)^5} z_K^2 \alpha_K^{1/2}\end{aligned}$$

Finally,

$$\begin{aligned}\mathbb{E}[NG(\pi^i_{T,K}, \pi^{-i}_{T,K})] \lesssim \frac{|\mathcal{S}|\mathcal{A}T\eta}{(1-\gamma)^3}\left(\frac{\gamma+1}{2}\right)^{T-1} + \frac{\sqrt{\mathcal{A}}}{\eta(1-\gamma)^2} \\ + \frac{|\mathcal{S}|^2\mathcal{A}^2\hat{L}_\eta}{\alpha_{k_0} c_{\alpha,\beta}(1-\gamma)^5} z_K^2 \alpha_K^{1/2}.\end{aligned}$$

This concludes the proof of theorem 3.2.

D. Proof of corollaries

D.1. Proof of corollary 3.3

We have from theorem 3.2 that under self-play, if $c_{\alpha,\beta} \leq \frac{c_\eta \ell_\eta^3 (1-\gamma)^2}{6272\eta^3 |\mathcal{S}|\mathcal{A}^4}$, then algorithm 1 under Assumption 3.1 achieves for all $K \geq k_0$:

$$\mathbb{E}[NG(\pi^i_{T,K}, \pi^{-i}_{T,K})] \leq \frac{\hat{c}_1' |\mathcal{S}|\mathcal{A}T\eta}{(1-\gamma)^3}\left(\frac{\gamma+1}{2}\right)^{T-1} + \frac{\hat{c}_2' |\mathcal{S}|^2 \mathcal{A}^2 \hat{L}_\eta}{\alpha_{k_0} c_{\alpha,\beta}(1-\gamma)^5} \frac{z_K^2 \alpha^{1/2}}{(K+h)^{1/2}} + \frac{\hat{c}_3' \sqrt{\mathcal{A}}}{\eta(1-\gamma)^2}.$$

In the following, we will provide bounds on the terms that appear in the bound above and that depend on η. Namely, we show that:

$$\begin{aligned} z_K &= \mathcal{O}(\log(K)/(\ell_\eta^{2r_b}\mu_{b,min})), \\ \hat{L}_\eta &= \mathcal{O}(\ell_\eta^{-2r_b}) \\ \inf_{\pi\in\Pi_\eta} \inf_{s\in\mathcal{S}} \mu_\pi(s) &= \Omega(\mu_{b,min}\ell_\eta), \end{aligned}$$

First, we prove these inequalities, then we use them to deduce the sample complexity.

Proof of inequalities

1) We know that $z_K = t(\ell_\eta, \beta_k) \leq \frac{t_{\pi_b,\beta_k}}{\ell_\eta^{2r_b}\mu_{b,min}}$ and $t_{\pi_b,\beta_K} = \mathcal{O}(\log(1/\beta_K)$ thanks to the fast mixing of π_b. Therefore, with the choice $\beta_K \propto 1/K$ we obtain that

$$z_K = \mathcal{O}(\log(K)/(\ell_\eta^{2r_b}\mu_{b,min})).$$

2) We have that $\hat{L}_\eta := \frac{2\log(8|\mathcal{S}|/\rho_\eta)}{\log(1/\rho_\eta)}$ with $\rho_\eta = \rho_b^{(\ell_\eta^2)^{r_b}\mu_{b,\min}}$, therefore as $\ell_\eta \to 0$ we obtain $\hat{L}_\eta = \mathcal{O}(\ell_\eta^{-2r_b})$.

3) We have $c_\eta = \mu_\eta \ell_\eta$, where $\mu_\eta := \inf_{\pi\in\Pi_{\ell_\eta}} \min_{s\in\mathcal{S}} \mu_\pi(s)$. To bound μ_η, notice that for any $\pi \in \Pi_{\ell_\eta}, k \in \mathbb{N}$, we have:

$$\begin{aligned} &\forall s' \in \mathcal{S}: \quad \sum_s \mathbb{P}_\pi^k(s'|s)\mu_\pi(s) = \mu_\pi(s') \\ \implies &\forall s' \in \mathcal{S}: \quad \sum_s \left(\sum_a (p^k(s'|s,a)\pi(a)\right)\mu_\pi(s) = \mu_\pi(s') \\ \implies &\forall s' \in \mathcal{S}: \quad \mu_\pi(s') \geq \ell_\eta \sum_s \left(\sum_a (p^k(s'|s,a)\pi_b(a)\right)\mu_\pi(s) \\ \implies &\forall s' \in \mathcal{S}: \quad \mu_\pi(s') \geq \ell_\eta \sum_s \mathbb{P}_{\pi_b}^k(s'|s)\mu_\pi(s), \\ \implies &\forall s' \in \mathcal{S}: \quad \mu_\pi(s') \geq \ell_\eta \sum_s \mu_{\pi_b}(s')\mu_\pi(s) \end{aligned}$$

where the third line follows because $\pi_b(s) \leq 1 \leq \frac{1}{\ell_\eta}\pi(s)$. The fifth one holds because π_b mixes at a geometric rate, and by definition of the mixing time $\mathbb{P}_{\pi_b}^k(s'|s) \to \mu_{\pi_b}(s')$. Therefore, since $\mu_\pi(s') \geq \mu_{b,min}$, we deduce that $\forall s' \in \mathcal{S}: \mu_\pi(s') \geq \mu_{b,min}\ell_\eta$.

Combination: Given lemma 4.1, we know that $\ell_\eta = \mathcal{O}(\eta^{-2})$, therefore, the condition on $c_{\alpha,\beta}$ becomes

$$c_{\alpha,\beta} \leq \frac{c_\eta \ell_\eta^3 (1-\gamma)^2}{6272\eta^3|\mathcal{S}|\mathcal{A}^4} \leq \frac{\mu_{b,min}\ell_\eta^2 \ell_\eta^3 (1-\gamma^2)}{6272\eta^3|\mathcal{S}|\mathcal{A}^4} = \frac{\mu_{b,min}(1-\gamma^2)}{6272\eta^{13}|\mathcal{S}|\mathcal{A}^4}$$

By injecting the obtained bounds in theorem 3.2 we find for all $K \geq k_0$:

$$\mathbb{E}[NG(\pi^i_{T,K}, \pi^{-i}_{T,K})] = \mathcal{O}\left(\eta\left(\frac{\gamma+1}{2}\right)^{T-1} + \eta^{12r_b+13}\frac{\log(K)^2}{\sqrt{K}} + \frac{1}{\eta}\right).$$

The first term in the r.h.s above is exponentially decaying in T. We optimize the choice of η for the two other terms and obtain the desired result for $\eta = K^{1/(24r_b+28)}$.

D.2. Proof of Corollary 3.4

In this section, we show that if the opponent were to fix their policy, then the player converges to the best response of the opponent. Our argument here is inspired by (Sayin et al., 2021), who used it to prove the rationality of decentralized Q-learning.

First, we argue that the Nash gap bound in theorem 3.2 remains true for noisy rewards as long as the zero-sum structure is preserved. Indeed, if we generalize the rewards for actions (a^i, a^{-i}) from $\mathcal{R}^i(a^i, a^{-i})$ to $r^i(a^i, a^{-i}, \xi)$, where $\xi \in \Xi$ is a random variable with distribution μ_ξ over Ξ (a finite set). If the noise is independent of the actions, $r^i + r^{-i} = 0$, and the reward is still uniformly bounded, then the proof still holds.

Second, observe that if the opponent follows a stationary policy π^{-i}, then it can be seen as additional randomness in the rewards. Specifically, consider a fictitious opponent with one available action a^*, and define $\hat{r}^i(a^i, a^*, a^{-i}) = \mathcal{R}^i(a^i, a^{-i})$, $\hat{p}(s' \mid s, a^i, a^*) = \sum_{\pi^{-i}(a^{-i}|s)} p(s' \mid a^i, a^{-i}, s)$ for all (a^i, a^{-i}). In this new zero-sum game, applying corollary 3.3 entails the same sample complexity for player i to find an approximate best response of its opponent.

E. Limitations of common assumptions: an illustrating example

Here we take a deeper look at the MDP of Figure 1, which we re-draw below for ease of readability.

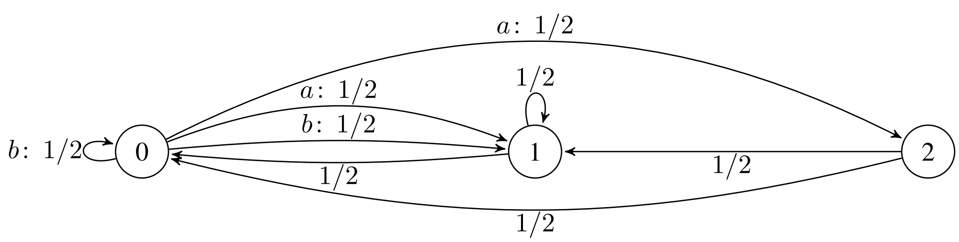

Figure 2. MDP with three states: state 0 has actions a and b, while states 2 and 3 don't have any action. The arrows indicate the possible transitions, which all have a probability of $1/2$.

We also recall the policy π parameterized by $\xi \in [0, 1]$ defined in Equation 2 as:

$$\pi(1, a) = \xi \text{ and } \quad \pi(1, b) = 1 - \xi. \tag{32}$$

Using this policy on the MDP above yields the following transition matrix:

$$P_\xi := \begin{bmatrix} \frac{1-\xi}{2} & \frac{1}{2} & \frac{\xi}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

Therefore, to obtain the corresponding stationary distribution we solve the following for $\mu_\pi = (x, y, z) \in [0, 1]^3$ and $x + y + z = 1$:

$$(x, y, z)P_\xi = (x, y, z)$$

which is equivalent to

$$\begin{cases} \left(\frac{1-\xi}{2}\right)x + \frac{1}{2}y + \frac{1}{2}z = x \\ \frac{1}{2}(x + y + z) = y \\ \frac{\xi}{2}x = z \\ x + y + z = 1 \end{cases}$$

then,

$$\begin{cases} z = \frac{\xi}{2}x \\ y = \frac{2+\xi}{2}x \\ x(2+\xi) = 1 \end{cases}$$

from which we deduce that the stationary distribution is given by:

$$\mu_\pi = \left(\frac{1}{2+\xi}, \frac{1}{2}, \frac{\xi}{4+2\xi}\right).$$

Strong reachability From the explicit formula of the stationary distribution μ_π we can deduce that as $\xi \to 0$ we have $\mu_\pi(3) \to 0$. Then, using Equation 1 we deduce that $T^\pi_{s\to s} \to \infty$. This entails that the strong reachability assumption does not hold in this simple setting. More generally, it is intuitive that stationary distributions of nearly deterministic policies would not have a full support in any non-trivial MDP.

Mixing time Now we shift our focus to the mixing time assumption and we show that the mixing time for this simple MDP and policy can grow arbitrarily if $\xi \to 1$. Consider an initial distribution $\mu_0 = (1/2, 1/4, 1/4)$, and denote by x_k the probability that the state at time k is the zero state. Our goal in the following is to show that the time it takes for x_k to converge to its stationary value can grow arbitrarily if $\xi \to 1$.

First, using a one step transition we have that:

$$x_{k+1} = \frac{1-\xi}{2}x_k + \frac{1}{4} + \frac{1}{2} * \frac{\xi}{2}x_{k-1},$$

which can be equivalently written as:

$$x_{k+1} - \frac{1-\xi}{2}x_k - \frac{\xi}{4}x_{k-1} - \frac{1}{4} = 0$$

and

$$\left(x_{k+1} - \frac{1}{2+\xi}\right) - \frac{1-\xi}{2}\left(x_k - \frac{1}{2+\xi}\right) - \frac{\xi}{4}\left(x_{k-1} - \frac{1}{2+\xi}\right) = 0.$$

The above is a homogeneous linear recurrence relation with constant coefficients and can be solved by solving the second degree equation in r: $r^2 - \frac{1-\xi}{2}r - \frac{\xi}{4} = 0$. The two solutions of this equation are $r_1 = \frac{1-\xi}{4} - \sqrt{\left(\frac{1-\xi}{4}\right)^2 + \xi}$ and $r_2 = \frac{1-\xi}{4} + \sqrt{\left(\frac{1-\xi}{4}\right)^2 + \xi}$ which entails the existence of $\alpha_\xi, \beta_\xi \in \mathbb{R}$ such that:

$$x_k = \alpha_\xi r_1^k + \beta_\xi r_2^k + \frac{1}{2+\xi}.$$

Going back to the mixing time, we have that:

$$\begin{aligned} t_{\pi,\epsilon} &= \min\left\{k \geq 0 \ : \max_{s\in\mathcal{S}} \left\|P_\xi^k(s) - \mu_\pi(s)\right\|_{\mathrm{TV}} \leq \epsilon\right\} \\ &\geq \min_{k\geq 0}\left\{\left\|P_\xi^{2k}(s)(0) - \frac{1}{2+\xi}\right\|_{\mathrm{TV}} \leq \epsilon\right\} \\ &= \min_{k\geq 0}\left\{(\alpha_\xi r_1^{2k} + \beta_\xi r_2^{2k}) \leq \epsilon\right\} \\ &\geq \min_{k\geq 0}\left\{\beta_\xi r_2^{2k} \leq \epsilon - \alpha_\xi r_1^{2k}\right\} \\ &\geq \frac{\log(\beta_\xi/(\epsilon - \alpha_\xi r_1^{2k}))}{\log\left(\frac{1}{(1-\xi)^2/8+\xi}\right)} - 1, \end{aligned}$$

Using the initial distribution $\mu_0 = (1/2, 1/4, 1/4)$ and P_ξ we can easily deduce that for $\xi \to 1$ we obtain $r_1 \to -1$ and $r_2 \to 1$ and $\alpha\xi \to 1/8$ $\beta_\xi \to 1/24$.

Consequently, we have for any $\epsilon \geq 1/3$, when $\xi \to 1$:

$$t_{\pi,\epsilon} \approx \frac{\log \frac{1}{24\epsilon - 3}}{\log \frac{1}{\xi}} - 1$$

which implies that:

$$t_{\pi,\epsilon} \to \infty \quad \text{when } \xi \to 1.$$

Since $\lim_{\xi \to 0} \mu_\pi(0) = +\infty$, then the mixing time can grow arbitrarily if $\xi \to 1$, thus invalidating Assumption 2.3 which completes our proof.

Remark E.1. Note that this computations could also be done with other initial distributions μ_0 as long as they entail non-zero limits for β_ξ when $\xi \to 1$. Indeed, the second root $r_2 \to 1$ when $\xi \to 1$ independently from the initial conditions, and this is the core insight behind this counterexample.