Oracle-free Reinforcement Learning in Mean-Field Games along a Single Sample Path

Muhammad Aneeq uz Zaman
Research Assistant
CSL, UIUC

Alec Koppel
Research Scientist
J.P. Morgan, USA

Sujay Bhatt
Research Scientist
J.P. Morgan, USA

Tamer Başar
Research Professor
CSL, UIUC

Abstract

We consider online reinforcement learning in Mean-Field Games (MFGs). Unlike traditional approaches, we alleviate the need for a mean-field oracle by developing an algorithm that approximates the Mean-Field Equilibrium (MFE) using the single sample path of the generic agent. We call this *Sandbox Learning*, as it can be used as a warm-start for any agent learning in a multi-agent non-cooperative setting. We adopt a two time-scale approach in which an online fixed-point recursion for the mean-field operates on a slower time-scale, in tandem with a control policy update on a faster time-scale for the generic agent. Given that the underlying Markov Decision Process (MDP) of the agent is communicating, we provide finite sample convergence guarantees in terms of convergence of the mean-field and control policy to the mean-field equilibrium. The sample complexity of the Sandbox learning algorithm is $\mathcal{O}(\epsilon^{-4})$ where ϵ is the MFE approximation error. This is similar to works which assume access to oracle. Finally, we empirically demonstrate the effectiveness of the sandbox learning algorithm in diverse scenarios, including those where the MDP does not necessarily have a single communicating class.

1 INTRODUCTION

Mean-Field Game (MFG) framework, concurrently introduced by Huang et al. Huang et al. (2006, 2007) and Lasry & Lions Lasry and Lions (2006, 2007), addresses some of the challenges faced by the widely applicable Multi-Agent Reinforcement Learning (MARL) framework Shoham et al. (2007); Ghasemi et al. (2020); Zhang et al. (2021); Mao et al. (2022). In particular, MFG framework captures the limiting case where the number of agents $N \to \infty$ and this deals with the non-stationarity of the environment caused by agents best responding to each other - referred to as the “curse of many agents" Sonu et al. (2017). In the infinite population setting, the effect of individual deviation becomes negligible causing any strategic interaction among the agents to disappear. As a result, it becomes sufficient to consider without loss of generality the interaction between a generic agent and the aggregate behavior of other agents (the mean-field). The solution concept used in MFGs (analog of Nash equilibrium) is called the Mean-Field Equilibrium (MFE). The MFE prescribes a set of control policies which are known to be ϵ-Nash for a large class of N-agent games Saldi et al. (2018), such that $\epsilon \to 0$ as $N \to \infty$. Hence finding the MFE presents a viable method to solving large population games. In this work we propose an RL algorithm to approximate the (stationary) MFE Guo et al. (2019); Xie et al. (2021) without assuming access to a mean-field oracle (henceforth referred to as oracle).

Most literature in RL for MFGs assumes access to such an oracle, which is capable of simulating the aggregate behavior of a large number of agents under a given control policy. But this assumption may be prohibitive and the generic agent may not have access to such an oracle, but only knows its own state, action and reward sequence. Hence the question arises:

Can the generic agent provably learn the stationary MFE without access to a mean-field oracle?

We answer this question in the affirmative by proposing an RL algorithm which computes the MFE without access to an oracle, but instead using the single sample path of the agent (without re-initializations) to approximate the aggregate behavior of large number of agents. We also provide high confidence finite sample bounds for approximation of the MFE to an arbitrary degree. We term this learning approach *Sandbox Learning*, since it allows an agent to approximate

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

equilibrium policies in a multi-agent non-cooperative environment, without interacting with other agents or an oracle. As a result, sandbox learning can be used to provide a *warm-start* to agents before entering an N-agent non-cooperative learning environment.

1.1 Main Results

Our core technical insight is that, instead of assuming access to the oracle, the problem may be cast as a stochastic fixed point problem using the generic agent's single sample path, thus allowing development of *oracle-free* RL algorithm for the MFG. In contrast, prior works require access to mean-field oracle Guo et al. (2019); Xie et al. (2021); Anahtarcı et al. (2019); Fu et al. (2020),which is a strong assumption, as it implicitly assumes the knowledge of the distribution of all other agents, which never holds in practice.The main results of the paper are as follows.

1. To efficiently learn the MFE and avoid degenerate policies, the Sandbox learning algorithm simultaneously updates the mean-field and the policy of the agent. This simultaneous update induces a time-varying Markov Chain (MC) for the generic agent which complicates the analysis of the algorithm. In Section 3, we craft episodic learning rates for the sole purpose of making the MC *slowly* time-varying inside the episode, making the algorithm amenable for analysis.

2. In Section 4, we provide finite sample analysis of Q-learning and dynamics matrix estimation under the slowly time-varying MC setting, using a communicating MDP condition from literature Arslan and Yüksel (2016). This condition generalizes the pre-existing conditions for RL-MFGs in literature. The slowly time-varying MC setting is shown to introduce a small *drift* in the approximation error, which can be reduced by slowing the inter-episodic learning. Lemmas 2 and 3 might be of independent interest to researchers in RL for time-varying MDPs.

3. The estimates of Q-function and dynamics matrix are used to construct approximate optimality and consistency operators, respectively. These operators are used to update the policy and mean-field using two time-scale learning. Finally in Section 4 Corollary 1, we obtain finite sample convergence bounds of this two time-scale algorithm to an ϵ-neighborhood of stationary MFE, under a standard contraction mapping assumption.

4. In Section 5, we numerically illustrate the effectiveness of the Sandbox learning algorithm on a congestion game. We empirically demonstrate that the Sandbox learning algorithm performs well even in the absence of the communicating MDP assumption, if there is a single closed communicating class. This is due to the fact that the MC transitions to the communicating class in finite time.

Proofs of theoretical claims are provided in the Supplementary Materials.

1.2 Relevant Literature

The work most closely related to this paper is Angiuli et al. (2022) which uses a unified-RL algorithm to solve the MFG problem in cooperative and non-cooperative settings, but lacks rigorous analysis of the RL algorithm. The key differences are that (a) the algorithm in Angiuli et al. (2022) relies on re-initializations while our algorithm operates on a single sample path, (b) the algorithm proposed in Angiuli et al. (2022) updates the Q-function at a faster time-scale while ours updates the control policy at a faster time-scale, and (c) we explicitly define the learning rates to have a certain episodic structure. These differences are shown to be pivotal in obtaining the finite sample convergence bounds for the Sandbox learning algorithm. Below we provide a table juxtaposing our work with the contributions of other works in RL for MFGs.

	Oracle-less?	Single sample path	Finite sample bounds
Elie et al. (2019)	✗	✗	✗
Cui and Koeppl (2021)	✗	✗	✗
Fu et al. (2020)	✗	✗	✓
Guo et al. (2019)	✗	✗	✓
Anahtarci et al. (2022)	✗	✗	✓
Xie et al. (2021)	✗	✗	✓
Angiuli et al. (2022)	✓	✗	✗
This work	✓	✓	✓

A complete literature review is provided in the Section 6.

2 FORMULATION & BACKGROUND

Consider an infinite horizon N-agent game over finite state and action spaces $\mathcal{S}$ and $\mathcal{A}$, respectively. The state and action of agent $i \in [N]$ at time t are denoted by $s_t^i \in \mathcal{S}$ and $a_t^i \in \mathcal{A}$, respectively. Agent i's initial state is drawn from a distribution $s_1^i \sim p_1 \in \mathcal{P}(\mathcal{S})$, and the state dynamics of the agent is coupled with the other agents through the empirical distribution $e_t^N := \frac{1}{N}\sum_{j\in[N]} \mathbb{1}\{s_t^j = s\}$, where we also include agent i, without any loss of generality. Agent i generates its actions using policy $\pi_t^i \in \Pi_t^i := \{\pi_t^i \mid \pi_t^i : \mathcal{S} \times \mathcal{P}(\mathcal{S}) \to \mathcal{P}(\mathcal{A})\}$, dependent on its state and the empirical distribution e_t^N. The state of agent i transitions according to

$$s_{t+1}^i \sim P(\cdot \mid s_t^i, a_t^i, e_t^N), s_1^i \sim p_1, a_t^i \sim \pi_t^i(s_t^i, e_t^N). \quad (1)$$

Similarly, the reward accrued to the agent depends on its state, action, and the empirical distribution at time t, $r_t^i = R(s_t^i, a_t^i, e_t^N) \in [0, 1]$. The presence of e_t^N in both (1) and r_t^i is a key point of departure from a standard MDP setting, as it permits other agents' possibly non-cooperative behavior to determine the evolution of the state and the reward of agent i. The over-arching goal of each agent $i = 1, \dots, N$ is to maximize its total reward discounted by a factor $0 < \rho < 1$, defined as

$$V^i(\pi^i, \pi^{-i}) = \mathbb{E}\Big[\sum_{t=1}^{\infty} \rho^t R(s_t^i, a_t^i, e_t^N) \mid s_t^i \sim p_1\Big], \quad (2)$$

where $\pi^i := (\pi_1^i, \pi_2^i, \dots) \in \Pi^i$ is the policy of agent i and $\pi^{-i} := \{\pi^j\}_{j \in [N] \setminus i}$ is the concatenation of policies of all other agents. In an N-agent non-cooperative game, the dominant solution concept is a Nash equilibrium, where none of the agents can increase their total reward by unilaterally deviating from its Nash policy. Based upon this notion, we define an ϵ-Nash equilibrium as follows.

Definition 1 (Başar and Olsder (1998))**.** *A set of policies $\pi^* = \{\pi^{1*}, \dots, \pi^{N*}\}$ is termed an ϵ-Nash equilibrium if $\forall i \in [N]$, $V^i(\pi^{i*}, \pi^{-i*}) + \epsilon > V^i(\pi^i, \pi^{-i*}), \forall \pi^i \in \Pi^i$.*

If $\epsilon \to 0$, ϵ-Nash approaches Nash equilibirum. Due to the exponential dependence on the number of agents N required to compute exact Nash equilibria Başar and Olsder (1998), we restrict focus to computing ϵ-Nash equilibria. In the case that the number of agents $N \to \infty$, known as the mean-field equilibrium (MFE), one obtains an ϵ-Nash equilibrium Saldi et al. (2018); Moon and Başar (2014), specifically, $\epsilon \to 0$ as $N \to \infty$.

Therefore, subsequently, we focus on the MFG, the infinite population analog of the N-agent game.The empirical distribution is replaced in that case by a mean field distribution $\mu = \lim_{N,t \to \infty} e_t^N$, its infinite population stationary counterpart. The stationary MFE of the MFG is guaranteed to exist under certain Lipschitzness assumptions Saldi et al. (2018); Jovanovic and Rosenthal (1988) (Assumption 1). As in the N-agent game, the generic agent in a MFG has state space $\mathcal{S}$, action space $\mathcal{A}$, and the initial distribution of its state is $p_1 \sim \mathcal{P}(\mathcal{S})$. Next, we define the agent's transition dynamics (1) and total reward (2) in the mean-field setting with mean-field $\mu \in \mathcal{P}(\mathcal{S})$:

$$s_{t+1} \sim P(\cdot \mid s_t, a_t, \mu), s_1 \sim p_1, a_t \sim \pi(s_t, \mu). \quad (3)$$

The actions of the generic agent are generated using a stationary stochastic policy $\pi \in \Pi := \{\pi : \mathcal{S} \times \mathcal{P}(\mathcal{S}) \to \mathcal{P}(\mathcal{A})\}$. We restrict ourselves to the set of stationary policies, without loss of generality, since the optimal control policy for an MDP induced by stationary μ is also stationary Puterman (2014). The instantaneous reward r_t accrued to a generic agent at time t is dependent on its state, control action, and the mean-field, that is, $r_t = R(s_t, a_t, \mu)$. The generic agent aims to maximize its total discounted reward given the mean-field μ and with the discount factor $0 < \rho < 1$,

$$V_{\pi,\mu} := \mathbb{E}\Big[\sum_{t=1}^{\infty} \rho^t R(s_t, a_t, \mu) \mid s_1 \sim p_1\Big]. \quad (4)$$

Next we define the Mean-Field Equilibrium (MFE) by introducing two operators. First define the *optimality* operator $\Gamma_1(\mu) := \mathrm{argmax}_\pi \, V_{\pi,\mu}$ as the operator which outputs the optimal policy for the MDP induced by mean-field μ. We consider policies where the probability is split evenly among optimal actions for a given state and mean-field. We also define $\Gamma_2(\pi, \mu)$ as the *consistency* operator which computes mean-field consistent with the policy π and mean-field μ. If $\mu' = \Gamma_2(\pi, \mu)$, then $\forall s' \in \mathcal{S}$

$$\mu'(s') = \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} P(s' \mid s, a, \mu)\pi(a \mid s, \mu)\mu(s). \quad (5)$$

This is also referred to as the Fokker-Planck-Kolmogorov equation in the literature Bensoussan et al. (2015), and versions of it appear in the literature on probability flow equations in MDPs Puterman (2014). Consistency means that if infinitely many agents (with initial distribution μ) follow a control policy π, the resulting distribution will be μ'. Using these two operators, we can define the MFE of the MFG as follows.

Definition 2 (Saldi et al. (2018))**.** *The pair $(\tilde{\pi}, \tilde{\mu})$ is an MFE of the MFG if $\tilde{\pi} = \Gamma_1(\tilde{\mu})$ and $\tilde{\mu} = \Gamma_2(\tilde{\pi}, \tilde{\mu})$.*

Intuitively this two-part coupled definition can be interpreted as (1) $\tilde{\pi}$ is the optimal policy for the MDP induced by mean-field $\tilde{\mu}$, and (2) mean-field $\tilde{\mu}$ is consistent with the control policy $\tilde{\pi}$. A naive way of approximating the MFE could be through repeated use of the composite operator $\Gamma_2(\Gamma_1(\cdot), \cdot)$ but this iteration is known to be non-contractive (Cui and Koeppl (2021)). Instead we replace $\Gamma_1(\cdot$ with the *approximate optimality* operator $\Gamma_1^\lambda(\mu) := \mathrm{softmax}_\lambda(\cdot, Q_\mu^*)$, where Q_μ^* is the Q-function of the MDP induced by mean-field μ and, the $\mathrm{softmax}_\lambda(\cdot)$ function is defined as

$$\mathrm{softmax}_\lambda(s, Q)_a := \frac{\exp(\lambda Q(s, a))}{\sum_{a' \in \mathcal{A}} \exp(\lambda Q(s, a'))}, \quad (6)$$

$\forall s \in \mathcal{S}, \forall a \in \mathcal{A}$. Evidently as $\lambda \to \infty$, $\Gamma_1^\lambda \to \Gamma_1$. Next using the approximate optimality operator Γ_1^λ we define an approximate MFE known as *Boltzman*-MFE (B-MFE).

Definition 3 (Cui and Koeppl (2021))**.** *For a given $\lambda > 0$, the pair (π^*, μ^*) is a Boltzman-MFE (B-MFE) of the MFG if $\pi^* = \hat{\Gamma}_1^\lambda(\mu^*)$ and $\mu^* = \Gamma_2(\pi^*, \mu^*)$.*

The Boltzman-MFE is an approximate MFE and approaches the MFE as $\lambda \to \infty$ (Theorem 4, Cui and Koeppl (2021)). Henceforth, we will devote ourselves to finding the B-MFE for a large enough λ, so as to closely approximate the MFE. Next we introduce the standard contraction mapping assumption in MFGs Guo et al. (2019); Xie et al. (2021).

Assumption 1. *There exists a $\lambda > 0$ and Lipschitz constants d_1, d_2 and d_3 such that*

$$\|\Gamma_1^\lambda(\mu) - \Gamma_1^\lambda(\mu')\|_{TV} \leq d_1\|\mu - \mu'\|_1,$$
$$\|\Gamma_2(\pi, \mu) - \Gamma_2(\pi', \mu)\|_1 \leq d_2\|\pi - \pi'\|_{TV},$$
$$\|\Gamma_2(\pi, \mu) - \Gamma_2(\pi, \mu')\|_1 \leq d_3\|\mu - \mu'\|_1$$

and $d := d_1 d_2 + d_3 < 1$ *for policies* $\pi, \pi' \in \Pi$ *and mean-fields* $\mu, \mu' \in \mathcal{P}(\mathcal{S})$.

Assumption 1 is guaranteed to be true for a small enough $\lambda > 0$ Cui and Koeppl (2021). This results in a trade-off as higher values of λ increase *closeness* between MFE and B-MFE, but may cause Assumption 1 to be violated, and vice-versa. This issue is well-known in MFGs over finite state and action spaces Cui and Koeppl (2021). Contraction mapping conditions (as in Assumption 1) are widely used in RL for standard MFGs Guo et al. (2019); Xie et al. (2021); Fu et al. (2020). Lemma 5 in Guo et al. (2019) provides candidate values for these constants. The $\|\cdot\|_{TV}$ norm used in Assumption 1 is the Total variation bound Cui and Koeppl (2021) and is defined for a function $f : \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ such that $\|f\|_{TV} := \max_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} |f(a \mid s)|$. Under Assumption 1, the existence and uniqueness of the B-MFE of the MFG has been proven in literature Cui and Koeppl (2021); Guo et al. (2019); Xie et al. (2021) using the standard contraction mapping theorem. Hence, the B-MFE approximates the MFE for large values of λ and the MFE is known to be ϵ-Nash for the finite population game (Theorem 2.3 Saldi et al. (2018)). In the next section, we propose an RL algorithm to approximate the B-MFE without access to a mean-field oracle, by utilizing the sample path of a generic agent itself.

3 SANDBOX REINFORCEMENT LEARNING

Consider a setting where a generic agent has no knowledge of the transition probability P, the functional form of the reward R or a mean-field oracle, which is often required in such studies – see Guo et al. (2019); Fu et al. (2020); Xie et al. (2021); Cui and Koeppl (2021). In this section, we propose a Sandbox RL algorithm to compute the B-MFE. Our methodology operates by updating the mean-field and the control policy concurrently using approximations of the optimality and consistency operators, Γ_1^λ and Γ_2, respectively, defined prior to Definition 3. The approximation to Γ_1^λ is defined by softmax$_\lambda(\cdot)$ of estimated Q-function obtained using Q-learning update, whereas approximation of operator Γ_2 relies on estimating the transition probabilities of the Markov Chain (MC) of the generic agent. But the concurrent update of mean-field and control policy causes the MC of the generic agent to be time-varying. This time-varying MC setting may cause instability in the approximation of the operators, resulting in divergence of mean-field and control policy updates.

To ensure good approximation of operators, we adopt an episodic two time-scale learning rate as shown in Figure 1. Inside an episode, the learning rates are summable (or fast-decaying), allowing the degree of non-stationarity in the MC inside the episode to be *slowly time-varying*. Doing so then enables us to ensure that the approximation errors of the optimality and consistency operators are under control.

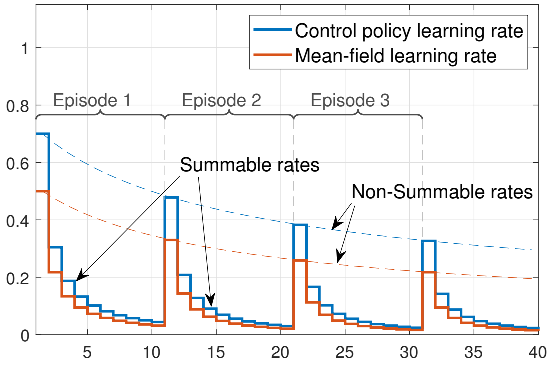

Figure 1: Episodic Two time-scale learning rate

Therefore, given a reasonable estimate for the consistency operator, the control policy is updated on a faster time-scale. Similarly the mean-field is updated at a slower time-scale using the consistency operator. We note that inverting the entity updated on a faster/slower time-scale will result in the solution to the Mean-Field Control problem Angiuli et al. (2022). In the following subsection we describe how we can estimate the two operators.

3.1 Approximate Mean-Field consistency and optimality operators

We start by describing how the Sandbox learning algorithm uses the MC of the generic agent to approximate the consistency operator Γ_2. Recalling the definition of Γ_2 (5), if $\mu' = \Gamma_2(\pi, \mu)$, then

$$\begin{aligned}\mu'(s') &= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} P(s' \mid s, a, \mu) \pi(a \mid s, \mu) \mu(s) \\ &= \sum_{s \in \mathcal{S}} P_{\pi,\mu}(s, s') \mu(s), \quad \forall s' \in \mathcal{S}\end{aligned}$$

where $P_{\pi,\mu}$ is the transition dynamics matrix of the generic agent under control law π and mean-field μ. Hence if $\mu' = \Gamma_2(\pi, \mu)$, the vector $\mu' \in \mathcal{P}(\mathcal{S})$ can be written as

$$\mu' = P_{\pi,\mu}^\top \mu \tag{7}$$

To come up with an estimator for Γ_2 we will need to estimate the dynamics matrix $P_{\pi,\mu}$. Toward this end, we can take a sample path of the Markov chain induced by π and μ of length T to obtain approximation of μ' through the use of an estimation of the occupancy (visitation) measure, and we can determine to what extent this estimate would be optimal through its ability to solve equation (7). More specifically, for a fixed pair of states $(i, j) \in \mathcal{S} \times \mathcal{S}$, the empirical transition probabilities $\hat{P}$ can be computed by keeping track of the state visitation numbers $N(i)$ and $N(i, j)$ as follows:

$$\hat{P}(i, j) = \frac{N(i, j) + 1/S}{N(i) + 1}, \tag{8}$$

where $N(i,j) = \big|\{l \in [T] : s_l = i, s_{l+1} = j\}\big|, N(i) = \sum_{j\in\mathcal{S}} N_t^k(i,j)$ and s_t is the state visited by the MC at time $t \in [T]$. Notice that we use smoothing (by adding $1/S$ and 1 to the numerator and the denominator, respectively) to avoid degenerate cases during the transition probability estimation. The transition probabilities $\hat{P}$ approximate the true transition probabilities $P_{\pi,\mu}$. Hence the approximate consistency operator is then given by $\hat{P}^\top \mu$, and the associated mean-field is updated by sequentially applying $\hat{P}^\top \mu$ with a specific step-size [cf. (12)] in (10), which we defer to the next subsection in order to underscore its concurrence with policy updates that are derived in terms of the Bellman equations.

Now we describe how the Sandbox learning algorithm approximates the optimality operator Γ_1^λ. As described in Section 2 $\Gamma_1^\lambda := \text{softmax}_\lambda(\cdot, Q^*_\mu)$, where $\text{softmax}_\lambda(\cdot)$ is defined in (6) and $Q^*_\mu(s,a) := \text{argmax}_\pi \mathbb{E}[\sum_{t=1}^\infty R(s_t, a_t, \mu)|s_1 = s, a_1 = a]$ is the optimal Q-function for the MDP induced by the mean-field μ and is the fixed point of the Bellman equation

$$Q^*_\mu(s,a) = R(s,a,\mu) + \rho \mathbb{E}_{s'\sim P(\cdot)}[\max_{a'} Q^*_\mu(s',a')].$$

The algorithm uses Q-learning update to approximate the optimal Q-function. The asynchronous Q-learning update Lewis et al. (2012); Even-Dar et al. (2003) can be written as follows,

$$\begin{aligned} &Q_{t+1}(s_t, a_t) = \\ &(1-\beta_t)Q_t(s_t) + \beta_t\big(r_t + \rho \max_{a\in\mathcal{A}} Q_t(s_{t+1}, a)\big), \end{aligned} \quad (9)$$

where $\beta_t := c_\beta/(t+1)^\nu$ and $0.5 < \nu \leq 1$. Let us denote the approximate optimality operator as $\hat{\Gamma}_1 := \text{softmax}_\lambda(\cdot, \hat{Q})$, where $\hat{Q}$ is the estimate obtained using the Q-learning update (9). The estimation error in $\hat{\Gamma}_1$ is due to the estimation error in the Q-learning, and is monotonically increasing with λ. With this technical machinery introduced for the approximate consistency operator and the Bellman operator, we are now ready to introduce the Sandbox learning algorithm. This is the focus of the following subsection.

3.2 Sandbox Reinforcement Learning algorithm

The Sandbox learning algorithm is presented in Algorithm 1. Throughout the algorithm the superscript $k \in [K]$ refers to the episode, and subscript $t \in [T]$ refers to the timestep inside the episode. Each episode k lasts for T timesteps. The state s_1^1 is initialized using distribution p_1 and a new state s_{t+1}^k is generated at each timestep t (line 5), and hence the algorithm evolves over a single sample path (of the generic agent) without re-initialization. The mean-field μ_t^k and the control policy π_t^k are updated at each timestep according to

$$\begin{aligned} \mu_t^k = \mathbb{P}_{S(\epsilon^{\text{net}})}\big[(1-c_{\mu,t}^k)\mu_{t-1}^k + &\quad (10)\\ c_{\mu,t}^k\big((\hat{P}_t^k)^\top \mu_{t-1}^k\big), \mathbb{1}_{\{t=1\}}\big],&\\ \pi_t^k = (1-c_{\pi,t}^k)\pi_{t-1}^k + &\quad (11)\\ c_{\pi,t}^k\big((1-\psi_t^k)\text{softmax}_\lambda(\cdot, Q_t^k) + \psi_t^k \mathbb{1}_{|\mathcal{A}|}\big).& \end{aligned}$$

The update of mean-field involves the operation $\mathbb{P}_{S(\epsilon^{\text{net}})}[\mu, x]$, which projects μ onto the ϵ-net $S(\epsilon^{\text{net}})$ if and only if $x = 1$. This projection step is performed on the first time-step of each episode k. In the analysis (Section 4) we show that $\epsilon^{\text{net}} = \mathcal{O}(\epsilon^2)$ at worst, where $\epsilon > 0$ is the approximation error in the B-MFE. The update of the mean-field is performed using the approximate consistency operator $(\hat{P}_t^k)^\top \mu_{t-1}^k$, and the control policy is updated using the approximate optimality operator $\text{softmax}_\lambda(\cdot, Q_t^k)$. The control policy updates also involve an exploration noise $\psi_t^k \mathbb{1}_{|\mathcal{A}|}$, which results in sufficient exploration of the state-action space (Lemma 1) without effecting the convergence bounds (Theorem 1 & Corollary 1). The expression for exploration coefficient ψ_t^k is provided in the proof of Lemma 1. The learning rates for the update, $c_{\mu,t}^k$ and $c_{\pi,t}^k$, are episodic two time-scale:

$$c_{\mu,t}^k = \frac{c_\mu}{k^\gamma}\frac{1}{t^\zeta}, \quad c_{\pi,t}^k = \frac{c_\pi}{k^\theta}\frac{1}{t^\zeta}, \quad (12)$$

where $0 < \theta < \gamma < 1 < \zeta < \infty$. The episodic nature of the learning rates is due to the $1/t^\zeta$ factor, $\zeta > 1$ in both rates, which makes it summable, resulting in slowly time varying MC inside the episode. The two time-scale nature of the learning rate is due to $\theta < \gamma$ where the update of the policy π_t^k is faster than that of the mean-field μ_t^k. Furthermore, the learning rates $c_{\mu,t}^k$ and $c_{\pi,t}^k$ are non-summable since $0 < \theta, \gamma < 1$. This episodic two time-scale nature is pivotal in proving that Sandbox RL converges to the B-MFE of the MFG as shown in the next section.

4 FINITE TIME BOUNDS FOR SANDBOX LEARNING

Most results in RL for MFGs break down in our setting as they assume a time invariant MC. In contrast, concurrent update of the mean-field and the control policy in the Sandbox learning algorithm induces a time-varying MC. In this section we analyze how the *slowly* time-varying MC under the episodic learning rates (10)-(11) is more amenable to analysis and leads to good approximations of Γ_1^λ and Γ_2 operators. Toward this end we first prove convergence of the transition probability and Q-learning estimation inside each episode $k \in [K]$ in Lemmas 2 and 3. These results are worthy of interest independent of the Sandbox learning algorithm, due to the slowly time-varying MC setting. In contrast, earlier works Guo et al. (2019); Xie et al. (2021); Anahtarcı et al. (2019) deal with approximating just Γ_1^λ under a time invariant MC. Then in Theorem 1 we show that

Algorithm 1: Sandbox RL for MFG

```
1: Initialize: initial state $s_1^1 \sim p_1$, policy $\pi_0^1$ and
   mean-field $\mu_0^1$
2: for $k \in \{1, 2, \dots, K\}$ do
3:     for $t \in \{1, 2, \dots, T\}$ do
4:         Update $\mu_t^k, \pi_t^k$ using (10), (11) respectively.
5:         Generate single transition
           $s_{t+1}^k \sim P(\cdot \mid s_t^k, a_t^k, \mu_t^k)$ and reward
           $r_t^k = R(s_t^k, a_t^k, \mu_t^k)$ with $a_t^k \sim \pi_t^k(s_t^k, \mu_t^k)$.
6:         Transition probability estimation: For
           $(i, j) \in \mathcal{S} \times \mathcal{S}$

               $\hat{P}_{t+1}^k(i, j) = \frac{N_t^k(i, j) + 1/S}{N_t^k(i) + 1}$,    (13)

           where $N_t^k(i, j) = \big|\{l \in [t] : s_l^k = i, s_{l+1}^k = j\}\big|$, $N_t^k(i) = \sum_{j \in \mathcal{S}} N_t^k(i, j)$.
7:         Q-learning: $Q_{t+1}^k(s_t^k, a_t^k) =$
           $(1 - \beta_t) Q_t^k + \beta_t\big(r_t^k + \rho \max_{a \in \mathcal{A}} Q_t^k(s_{t+1}^k, a)\big)$
8:     end for
9:     $\hat{P}_1^{k+1} = \hat{P}_{T+1}^k, Q_1^{k+1} = Q_{T+1}^k, \mu_0^{k+1} =$
       $\mu_T^k, \pi_0^{k+1} = \pi_T^k, s_1^{k+1} = s_{T+1}^k$
10: end for
11: Output: Approximate B-MFE
    $(\frac{1}{K} \sum_{k=1}^{K-1} \pi_1^k, \frac{1}{K} \sum_{k=1}^{K-1} \mu_1^k)$.
```

good approximation of Γ_1^λ and Γ_2 operators (due to good Q-learning and transition probability estimation, respectively) results in decreasing average error in policy and mean-field. Finally, in Corollary 1, we present finite sample analysis for the two time-scale Sandbox learning algorithm.

Lemma 2 presents error bounds on transition probability estimation (8) for a slowly time-varying MC, under a communicating MDP condition as given below. Assumption 2 *generalizes* the pre-existing conditions for RL-MFGs in literature. The online RL-MFG works of Guo et al. (2019) and Xie et al. (2021) (and references therein Shah and Xie (2018); Farahmand et al. (2016)) assume either a covering time assumption or require i.i.d. samples from stationary distribution. The offline RL-MFG works of Anahtarcı et al. (2019) and Fu et al. (2020) rely on i.i.d. samples from unique stationary distribution of MC which requires ergodicity. Communicating MDP (Assumption 2) is more general than covering time or ergodicity conditions Chandrasekaran and Tewari (2021). Before stating the communicating MDP condition, we introduce the set $S(\epsilon^{\text{net}})$ which is a set of mean-field distributions. This set (also termed ϵ-net Guo et al. (2019) over $\mathcal{P}(\mathcal{S})$) defined as $S(\epsilon^{\text{net}}) = \{\mu^1, \dots, \mu^{N_{\text{net}}}\} \subset \mathcal{P}(\mathcal{S})$ is a finite set of simplexes over $\mathcal{S}$ such that $\|\mu - \mu'\|_1 \leq \epsilon^{\text{net}}$ for any $\mu \in \mathcal{P}(\mathcal{S}), \exists \mu' \in S(\epsilon^{\text{net}})$. The existence of the set is guaranteed due to the compactness of $\mathcal{P}(\mathcal{S})$.

Assumption 2. (Communicating MDPs Arslan and Yüksel (2016)) *For any mean-field $\tilde{\mu} \in S(\epsilon^{net})$ (which is a finite set) and any pair of states $s, s' \in \mathcal{S}$, there exists a finite horizon $H(\tilde{\mu})$ such that for $t \geq H(\tilde{\mu})$ there exists a set of actions $\tilde{a}_1, \dots, \tilde{a}_t$,*

$$P\big(s_t = s' \mid a_1 = \tilde{a}_1, \dots, a_t = \tilde{a}_t, s_1 = s, \mu = \tilde{\mu}\big) > 0.$$

Informally Assumption 2 means that every agent in the game has a path from any state to any other state for mean-fields in $S(\epsilon^{\text{net}})$. Assumption 2 is satisfied in several real-world scenarios. Production of an exhaustible resource by competing producers (e.g. oil) is a typical multi-agent setting where Assumptionm 2 is satisfied Guéant et al. (2011), since the agents can achieve any level of reserve by increasing/decreasing their production. Capital accumulation games Fershtman and Muller (1984) and asset management games Lacker and Zariphopoulou (2019) have a similar structure thus satisfying Assumptionm 2. It is also satisfied in cyber-security applications Kolokoltsov and Bensoussan (2016), as any infection state can be reached by choosing the right policy and a strictly positive MF ϵ-net. In Section 5 we numerically investigate a setting where such an assumption is not satisfied. Next, under the communicating MDP assumption, we prove sufficient exploration of state and action space, under the policy update (11).

Lemma 1. (Sufficient Exploration) *If Assumption 2 is satisfied, stochastic kernel $P(\cdot \mid s, a, \mu)$ is Lipschitz in μ, and ζ is large enough, then under the control policy update (Algorithm 1 line 4), there exists a $\sigma \in (0, 1)$ such that for any $(s, a) \in |\mathcal{S} \times \mathcal{A}|$ and $t \geq H := \max_{\tilde{\mu} \in S(\epsilon^{net})} H(\tilde{\mu})$, $P((s_t, a_t) = (s, a) \mid \mathcal{F}_{t-H}) \geq \sigma$.*

Lemma 1 implies that the communicating MDP assumption coupled with the policy update (11), is more general than the sufficient exploration condition used in Q-learning for MDPs (Qu and Wierman (2020) Assumption 3) as well as for N-player stochastic games (Hu and Wellman (2003) Assumption 1). Furthermore, it is more general than an ergodicity assumption used in the stochastic optimization literature Srikant and Ying (2019). We further note that the sufficient exploration condition has also been used in two time-scale settings in the literature Wu et al. (2020). Next we quantify the error in transition probability estimation in Lemma 2 under Assumption 2 and Lipschitz conditions on transition probability P Angiuli et al. (2022). The estimation error is denoted by ϵ_P^k, and is the norm of the difference between the transition probability estimate $\hat{P}_T^k$ and the true transition probability induced by the control policy and the mean-field at the first timestep (π_1^k, μ_1^k, respectively).

Lemma 2. *Given that Assumption 2 is satisfied and transition probability $P_{\pi,\mu}$ is Lipschitz in policy π and mean-field μ such that $\|P_{\pi,\mu} - P_{\pi',\mu}\|_F \leq L_P^\pi \|\pi - \pi'\|_{TV}$ and $\|P_{\pi,\mu} - P_{\pi,\mu'}\|_F \leq L_P^\mu \|\mu - \mu'\|_1$, the error in transition probability estimation for episode k is*

$$\begin{aligned} \epsilon_P^k &:= \|\hat{P}_T^k - P_{\pi_1^k, \mu_1^k}\|_F \\ &= \tilde{\mathcal{O}}(T^{-1/2}) + \tilde{\mathcal{O}}(T^{-1}) + \mathcal{O}(2^{1-\zeta}) \end{aligned}$$

with probability at least $1-\delta_P$ where $P_{\pi_1^k,\mu_1^k}$ is the transition probability under control law π_1^k and mean-field μ_1^k.

The Lipschitz conditions in Lemma 1 will follow if transition probability is continuous in the mean-field μ and the policy $\pi(\cdot \mid s)$, due to the compactness of mean-field and policy spaces, $\mathcal{P}(\mathcal{S})$ and $\mathcal{P}(\mathcal{A})$, respectively. And in most real-world examples, such as asset Reis and Platonov (2019) and crowd management Priuli (2014), continuity of transition probability w.r.t. mean-field and policy is ensured. Lemma 2 shows that the estimation error ϵ_P^k contains a drift term $\mathcal{O}(2^{1-\zeta})$ due to the slowly time-varying MC setting which can be decreased by increasing the inter-episodic learning parameter ζ. Aside from drift, ϵ_P^k grows at $\tilde{\mathcal{O}}(T^{-1/2})+\tilde{\mathcal{O}}(T^{-1})$, where $\tilde{\mathcal{O}}$ hides logarithmic factors. Hence increasing the duration of episode T will result in decrease in estimation error. The proof of the lemma is given in the Supplementary Notes and relies on Freedman's inequality Freedman (1975).

Next we analyze the error in Q-learning estimation (9) for each episode $k\in[K]$. This update has been shown to converge to the optimal Q function under a sufficient exploration condition (stronger than Assumption 2) for a time invariant MC Even-Dar et al. (2003); Qu and Wierman (2020). In Lemma 3 we show that this update converges (albeit with a drift) under the comunicating MDP condition for the slowly time-varying MC setting and with $0.5<\nu\leq 1$. This estimation error is denoted by ϵ_Q^k, and is the norm of the difference between the estimate of the optimal Q-function Q_T^k and the true Q-function $Q_1^{*,k} := Q^*_{\mu_1^k}$ for the MDP induced by the mean-field μ_1^k. As in Lemma 2, a drift term $\mathcal{O}(2^{1-\zeta})$ creeps in due to the slowly time-varying MC setting.

Lemma 3. *Under Assumption 2, the estimation error in Q-learning for episode k is*

$$\begin{aligned}\epsilon_Q^k &:= \|Q_T^k - Q_1^{*,k}\|_\infty\\ &= \mathcal{O}(T^{1-2\nu})+\mathcal{O}(T^{1-\zeta-\nu})+\tilde{\mathcal{O}}(T^{1/2-\nu})+\mathcal{O}(2^{1-\zeta})\end{aligned}$$

with probability at least $1-\delta_Q$.

The slowly time-varying MC also contributes $\mathcal{O}(T^{1-\zeta-\nu})$ error component but that is dominated by the $\mathcal{O}(T^{1-2\nu})$ term due to $\nu\leq 1<\zeta$. The error terms, which are $\mathcal{O}(T^{1-2\nu})$ and $\tilde{\mathcal{O}}(T^{1/2-\nu})$, are decreasing for increasing T, and hence to get a small ϵ_Q^k we need a large enough T. Combining the bounds from Lemmas 2 and 3 we can surmise that given that the inter-episode learning parameter ζ and the episode length T are large enough, the transition probability estimation and Q-learning will be good enough, leading to good approximations of the consistency and optimality operators, Γ_2 and Γ_1^λ, respectively.

Now we are in a position to present Theorem 1, relying on good approximations of the consistency and optimality operators. Theorem 1 below bounds the average error in policy $e_\pi^k := \|\pi_1^k - \Gamma_1^\lambda(\mu_1^k)\|_{TV}$ and mean-field $e_\mu^k := \|\mu_1^k-\mu^*\|_1$ over episodes $k\in[K]$, given that $\epsilon_P^k\leq\epsilon_P, \epsilon_Q^k\leq\epsilon_Q/\log(K)$ for some $\epsilon_P,\epsilon_Q>0$.

Theorem 1. *Let the approximation errors be denoted by $e_\pi^k := \|\pi_1^k - \Gamma_1^\lambda(\mu_1^k)\|_{TV}$ and $e_\mu^k := \|\mu_1^k-\mu^*\|_1$ and $\epsilon^{net}\leq(c_\mu\bar{d}\epsilon)/K^\gamma$ for $\epsilon>0$. Under Assumptions 1-2, with the estimation errors satisfying $\epsilon_P^k\leq\epsilon_P,\epsilon_Q^k\leq\epsilon_Q/\lambda$, for some $\epsilon_P,\epsilon_Q>0$, the average approximation errors decrease at the following rates:*

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K-1}e_\pi^k =&\mathcal{O}(K^{\theta-1})+\mathcal{O}(\epsilon_Q)+\mathcal{O}(K^{\theta-\gamma})\\ &+\mathcal{O}(K^{-1})+\mathcal{O}(2^{1-\zeta}),\\ \frac{1}{K}\sum_{k=1}^{K-1}e_\mu^k =&\mathcal{O}(K^{\gamma-1})+\mathcal{O}(\epsilon)+\mathcal{O}(\epsilon_P)+\frac{1}{K}\sum_{k=1}^{K-1}e_\pi^k\end{aligned}$$

with probability at least $1-\delta_Q$, where $0<\theta<\gamma<1<\zeta<\infty$.

The challenges in establishing Theorem 1 are due to the two time-scale learning rates and non-regularized MFG setting. The proof of Theorem 1 keeps track of errors e_π^k and e_μ^k for each episode k and the average of these errors is shown to approach 0 due to tight approximation of the optimality and consistency operators and the two time-scale update under the contraction Assumption 1. Apart from the familiar drift terms $\mathcal{O}(2^{1-\zeta})$ and estimation error bounds ϵ_P and ϵ_Q, all other terms are decreasing with increasing total number of episodes K at rates governed by θ,γ and ζ. Next we present a corollary to Theorem 1 which gives us the final bound quantifying the approximation error between output of the Sandbox learning algorithm and the B-MFE of the MFG. The proof of Corollary 1 depends on the result of Theorem 1 and is provided in the Supplementary Notes.

Corollary 1. *If all conditions in Theorem 1 are satisfied, we have*

$$\begin{aligned}\left\|\frac{1}{K}\sum_{k=1}^{K-1}\pi_1^k-\pi^*\right\|_{TV}+\left\|\frac{1}{K}\sum_{k=1}^{K-1}\mu_1^k-\mu^*\right\|_1 =\\ \mathcal{O}(K^{\gamma-1})+\mathcal{O}(2^{1-\zeta})+\mathcal{O}(\epsilon_P)+\mathcal{O}(K^{\theta-1})\\ +\mathcal{O}(\epsilon_Q)+\mathcal{O}(K^{\theta-\gamma})+\mathcal{O}(\epsilon).\end{aligned}$$

The terms $\mathcal{O}(K^{\theta-1}),\mathcal{O}(K^{\gamma-1})$ and $\mathcal{O}(K^{\theta-\gamma})$ enter due to the two-time-scale learning setting and are monotonically decreasing with the number of episodes K. The convergence rates of these terms can be tuned by choosing θ and γ as explained next. The terms $\mathcal{O}(\epsilon_Q)$ and $\mathcal{O}(\epsilon_P)$ enter into the analysis due to the estimation errors in the Q-function and the dynamics matrix P, respectively. These quantities can be made arbitrarily small by increasing the number of timesteps in each episode T, due to Lemmas 1 and 2. Lastly, the term $\mathcal{O}(2^{1-\gamma})$ is a drift term which enters due to inter-episodic learning. This can be made arbitrarily small by increasing the value of γ. But the value of $\gamma \neq \infty$ as that stops inter-episodic learning and may cause degenerate policies. The error introduced by the projection step in the mean-field update line 4 of Algorithm 1 is $\mathcal{O}(\epsilon)$.

If the learning rates are chosen such that $\theta = 0.01, \gamma = 0.5, \nu = 1, \zeta = \Omega(\log(1/\epsilon))$, then the output of Algorithm 1 will be ϵ close to the B-MFE with high probability (for small enough $\epsilon > 0$), given that episode length is $T = \Omega(\epsilon^{-2})$ and the number of episodes is $K = \Omega(\epsilon^{-2})$. Notice that under these conditions, $\epsilon^{\text{net}} = \mathcal{O}(\epsilon^2)$. Hence the sample complexity of the algorithm is $\mathcal{O}(\epsilon^{-4})$. Finally, the difference between the MFE and the B-MFE will be determined by the λ parameter with higher values resulting in a close approximation of MFE by the B-MFE Cui and Koeppl (2021). In the next section we apply the algorithm to a congestion game.

5 NUMERICAL RESULTS

We simulate Sandbox learning for a congestion MFG Toumi et al. (2020) on a grid. In particular we investigate the convergence of the algorithm for the cases (1) when the communicating MDP assumption is satisfied, and (2) when it is not satisfied. The state space $\mathcal{S} = \{1, \dots, 5\}^2$, action space $\mathcal{A} = \{-1, 1\}^2$, and discount factor $\rho = 0.7$. The initial distribution of the agent p_1 is uniform over the state space. If the agent takes action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$, the resulting state will be $s' = \mathbb{P}_{\mathcal{S}}[s + a]$ with probability $1 - p$ or the agent may be 'jostled' into one of the neighboring states, with probability p for small $p > 0$. This stochasticity is meant to model the jostling present in crowds. The agent's reward is $r(s, \mu) = (1 - c \cdot \mu(s)) \cdot R(s)$ where $c = 0.5$ is the congestion averse parameter and $R(s)$ is the state dependent component of the reward. The state-dependent rewards $R(s)$ are concentrated around favorable states $\{(3,3), (3,4), (4,3), (4,4)\}$. Hence the agent prefers states with higher $R(s)$ values but might be deterred by the congestion in the state $\mu(s)$.

The initial control policy π_0^1 and mean-field μ_0^1 are uniform over actions and states, respectively. The initial estimate of the Q-function Q_1^1 is zero and transition probability estimate $\hat{P}_1^1$ is uniform. By choosing learning coefficient $c_\beta = 5$, learning rate $\nu = 0.55$ and episodic length $T = 5 \times 10^4$ we observe in Figure 2 that Q-learning and transition probability estimation converge very well inside episode $k = 1$.

Furthermore, by choosing $c_\mu = c_\pi = 0.5$ and the two time-scale learning rates as $\theta = 0.55 < \gamma = 0.6$ we see that the control policy and mean-field estimates converge after $K = 300$ many episodes in Figure 3. For this particular simulation we forego the projection step as it does not have a significant impact on the convergence of the algorithm.

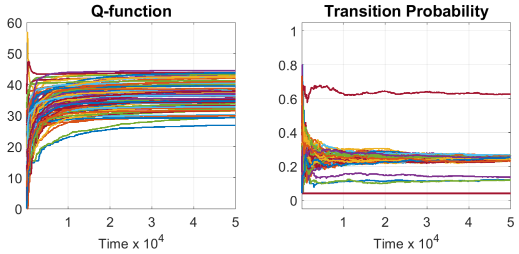

Figure 2: Convergence of transition probability and Q-function estimation for episode $k = 1$

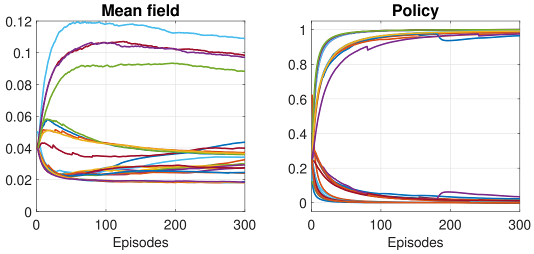

Figure 3: Convergence of mean-field and control policy estimates

Next, we consider the setting where the state space consists of two communicating classes, thereby nullifying the communicating MDP assumption (Assumption 2). In particular, the states $\mathcal{C}^1 := \{(4,5), (5,4), (5,5)\}$ form a communicating class which is not closed and the rest of the state space is a closed communicating class $\mathcal{C}^0$. The reward function $r(s, \mu)$ is exactly the same as the previous simulation but the transition probabilities are modified to prevent transition $\mathcal{C}^0 \to \mathcal{C}^1$, ensuring that $\mathcal{C}^0$ is closed and $\mathcal{C}^1$ is not closed. The learning rates and other estimates are initialized as before. Figure 4 shows the convergence of mean-field and control policy estimates in the Sandbox learning algorithm.

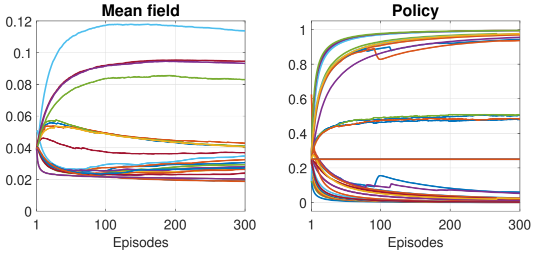

Figure 4: Convergence of mean-field and control policy estimates in the absence of Assumption 2

The simulation shows good convergence properties of the algorithm even in absence of the communicating MDP condition. This is due to the fact that the MC transitions (in finite time) from $\mathcal{C}^1$ into $\mathcal{C}^0$ and stays there, allowing good approximations of Γ_1, Γ_2 operators, resulting in convergence. If there were multiple closed classes with initial distribution spread amongst them, the algorithm's performance would be variable.

6 EXTENDED LITERATURE

Mean-Field Games (MFGs) originated concurrently in the works of Huang et al. (2006, 2007) (termed as Nash Certainty Equivalence) and Lasry and Lions (2006, 2007) (who coined the term MFG). Since its inception, there have been several works extending the classical concept of MFGs in various directions, such as heterogenous agents Moon and Başar (2014), scarce interactions Caines and Huang (2019); Zaman et al. (2021), risk-sensitive criteria Tembine et al. (2013); Moon and Başar (2016); Saldi et al. (2020) and cooperative equilibria Bensoussan et al. (2018); Barreiro-Gomez and Tembine (2021). MFGs have also been applied to a variety of real-world applications such as decentralized charging of EVs Ma et al. (2011), economics Carmona (2020) and congestion dynamics Lachapelle and Wolfram (2011), among others. Although most of these works have been in the continuous time setting, research in discrete-time MFGs which are much more amenable to Reinforcement Learning have also been gaining momentum recently Saldi et al. (2018); Moon and Başar (2014).

RL for MFGs was first dealt with in Guo et al. (2019) for the finite and in Elie et al. (2019) for infinite state and action spaces. The work of Guo et al. (2019) proposes a double-loop RL algorithm for MFGs with finite state and action spaces MFGs, which involves a projection step onto an ϵ-net. This projection step helps in establishing convergence by utilizing a uniform action gap bound over the ϵ-net. A fictitous play algorithm was proposed Elie et al. (2019), involving repeated updates of the mean-field and control policy to approximate the MFE. The first set of works to deal with RL for the benchmark Linear Quadratic (LQ) MFGs were Fu et al. (2020); Zaman et al. (2020, 2022). These works have provided finite sample bounds for the LQ-MFG in the stationary Fu et al. (2020) and the non-stationary Zaman et al. (2020, 2022) settings, by building on policy gradient Fazel et al. (2018) and zero-order stochastic optimization methods Malik et al. (2019) for the Linear Quadratic Regulator problem. The recent work of Yongacoglu et al. (2022) deals with independent learning for a novel setting of N-player mean-field games. The work of Lee et al. (2021) learns the MFE in the special setting of strategic complementarities where a single step of centralized Q-learning followed by a single step of policy execution by many agents is shown to converge to the MFE. The idea of entropy-regularized MFGs was introduced in Xie et al. (2021); Cui and Koeppl (2021) along with existence and uniqueness results and RL algorithms to compute the entropy-regularized MFE. The work of Anahtarcı et al. (2019) also deals with the entropy-regularized MFGs, by utilizing a fitted Q-iteration based approach. There have been several works on Deep-RL techniques for MFGs, such as Perrin et al. (2021); Subramanian and Mahajan (2019), where Perrin et al. (2021) uses Deep RL techniques to learn a flocking model observed in nature and Subramanian and Mahajan (2019) proposes a Neural Network based policy update mechanism. The paper Xie et al. (2021) proposed a single-loop RL algorithm, such that each critic step leads to a mean-field update as well. This is in contrast to the standard RL for MFG algorithms which have a double-loop structure where multiple critic steps can be executed while keeping the mean-field constant. Our work also has a single-loop structure as each critic step of control policy update leads to a concurrent update of the mean-field. Furthermore, we consider learning along a single sample path of the generic agent without re-initializations.

In addition, the majority of the literature in RL for MFGs assume access to an oracle which can provide the mean-field (or a noisy estimate of it) under a given control policy. This work, on the other hand, proposes the Sandbox learning algorithm which uses the sample path of the agent itself to estimate the mean-field. The work closest to our setting is Angiuli et al. (2022) which adopts an oracle-less setting but does not provide a finite sample convergence bound of the RL algorithm. Furthermore, the two time-scale update in Angiuli et al. (2022) updates the Q-function at a faster rate whereas Sandbox learning algorithm updates the control policy at the faster rate using a softmax of the estimated Q-function. Furthermore, we prove that the episodic nature of learning rates in Sandbox learning is crucial to obtaining finite sample convergence guarantees. Sandbox learning can be extended to entropy-regularized setting by employing a fitted Q-iteration (as in Anahtarcı et al. (2019)).

7 CONCLUSION & FUTURE WORK

In this paper we have developed the Sandbox learning algorithm with finite-time guarantees to approximate the stationary MFE of a MFG without access to a mean-field oracle, using the single sample path of the generic agent. The sample complexity of the Sandbox learning algorithm is $\mathcal{O}(\epsilon^{-4})$ where ϵ is the MFE approximation error. The proof of convergence has relied on goodness of transition probability and Q-function estimates (along slowly time-varying MC). The control policy and the mean-field were then updated using two time-scale learning rates and approximate consistency and optimality operators. We also generalize the covering time and ergodicity assumptions in literature by proposing a communicating MDP assumption. This work opens up several interesting research directions. It would be worthwhile to explore finite sample bounds for Sandbox learning under a *weakly* communicating MDP condition. Another important research direction would be to explore how feature embeddings can improve the scalability beyond the tabular MFG setting. Furthermore, extending oracle-less learning to the benchmark setting of Linear Quadratic MFGs may also help in solving several real-world scenarios such as, consensus and formation flying.

Acknowledgements

Research of the first and fourth authors was supported in part by the Air Force Office of Scientific Research (AFOSR) Grant FA9550-19-1-0353. We would like to thank the reviewers for their useful comments.

References

Berkay Anahtarcı, Can Deha Karıksız, and Naci Saldi. Fitted Q-learning in mean-field games. *arXiv preprint arXiv:1912.13309*, 2019.

Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Q-learning in regularized mean-field games. *Dynamic Games and Applications*, pages 1–29, 2022.

Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Laurière. Unified reinforcement Q-learning for mean field game and control problems. *Mathematics of Control, Signals, and Systems*, pages 1–55, 2022.

Gürdal Arslan and Serdar Yüksel. Decentralized Q-learning for stochastic teams and games. *IEEE Transactions on Automatic Control*, 62(4):1545–1558, 2016.

Julian Barreiro-Gomez and Hamidou Tembine. *Mean-field-type Games for Engineers*. CRC Press, 2021.

Tamer Başar and Geert Jan Olsder. *Dynamic Noncooperative Game Theory*. SIAM, 1998.

Alain Bensoussan, Jens Frehse, and Sheung Chi Phillip Yam. The master equation in mean field theory. *Journal de Mathématiques Pures et Appliquées*, 103(6):1441–1474, 2015.

Alain Bensoussan, Tao Huang, and Mathieu Laurière. Mean field control and mean field game models with several populations. *arXiv preprint arXiv:1810.00783*, 2018.

Peter E Caines and Minyi Huang. Graphon mean field games and the GMFG equations: ε-Nash equilibria. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 286–292. IEEE, 2019.

Rene Carmona. Applications of mean field games in financial engineering and economic theory. *arXiv preprint arXiv:2012.05237*, 2020.

Gautam Chandrasekaran and Ambuj Tewari. Learning in online mdps: Is there a price for handling the communicating case? *arXiv preprint arXiv:2111.02024*, 2021.

Kai Cui and Heinz Koeppl. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1909–1917. PMLR, 2021.

Romuald Elie, Julien Pérolat, Mathieu Laurière, Matthieu Geist, and Olivier Pietquin. Approximate fictitious play for mean field games. *arXiv preprint arXiv:1907.02633*, 2019.

Eyal Even-Dar, Yishay Mansour, and Peter Bartlett. Learning rates for Q-learning. *Journal of Machine Learning Research*, 5(1), 2003.

Amir-massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

Maryam Fazel, Rong Ge, Sham M Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476, 2018.

Chaim Fershtman and Eitan Muller. Capital accumulation games of infinite duration. *Journal of Economic Theory*, 33(2):322–339, 1984.

David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

Zuyue Fu, Zhuoran Yang, Yongxin Chen, and Zhaoran Wang. Actor-critic provably finds Nash equilibria of linear-quadratic mean-field games. In *International Conference on Learning Representation*, 2020.

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.

Arman Ghasemi, Amin Shojaeighadikolaei, Kailani Jones, Morteza Hashemi, Alexandru G Bardas, and Reza Ahmadi. A multi-agent deep reinforcement learning approach for a distributed energy marketplace in smart grids. In *2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 1–6. IEEE, 2020.

Olivier Guéant, Jean-Michel Lasry, and Pierre-Louis Lions. *Mean Field Games and Applications*, pages 205–266. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-14660-2. doi: 10.1007/978-3-642-14660-2_3. URL `https://doi.org/10.1007/978-3-642-14660-2_3`.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. In *Advances in Neural Information Processing Systems*, 2019.

Daniel Hsu, Aryeh Kontorovich, David A Levin, Yuval Peres, Csaba Szepesvári, and Geoffrey Wolfer. Mixing time estimation in reversible Markov chains from a single sample path. *The Annals of Applied Probability*, 29(4): 2439–2480, 2019.

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

Minyi Huang, Roland P Malhamé, and Peter E Caines. Large population stochastic dynamic games: Closed-loop Mckean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information & Systems*, 6(3):221–252, 2006.

Minyi Huang, Peter E Caines, and Roland P Malhamé. Large-population cost-coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized ε-Nash equilibria. *IEEE Transactions on Automatic Control*, 52(9):1560–1571, 2007.

Boyan Jovanovic and Robert W Rosenthal. Anonymous sequential games. *Journal of Mathematical Economics*, 17(1):77–87, 1988.

Vassili N Kolokoltsov and Alain Bensoussan. Mean-field-game model for botnet defense in cyber-security. *Applied Mathematics & Optimization*, 74:669–692, 2016.

Aimé Lachapelle and Marie-Therese Wolfram. On a mean field game approach modeling congestion and aversion in pedestrian crowds. *Transportation Research Part B: Methodological*, 45(10):1572–1589, 2011.

Daniel Lacker and Thaleia Zariphopoulou. Mean field and n-agent games for optimal investment under relative performance criteria. *Mathematical Finance*, 29(4):1003–1038, 2019.

Jean-Michel Lasry and Pierre-Louis Lions. Jeux à champ moyen. i–le cas stationnaire. *Comptes Rendus Mathématique*, 343(9):619–625, 2006.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2(1):229–260, 2007.

Kiyeob Lee, Desik Rengarajan, Dileep Kalathil, and Srinivas Shakkottai. Reinforcement learning for mean field games with strategic complementarities. In *International Conference on Artificial Intelligence and Statistics*, pages 2458–2466. PMLR, 2021.

Frank L Lewis, Draguna Vrabie, and Vassilis L Syrmos. *Optimal Control*. John Wiley & Sons, 2012.

Zhongjing Ma, Duncan S Callaway, and Ian A Hiskens. Decentralized charging control of large populations of plug-in electric vehicles. *IEEE Transactions on Control Systems Technology*, 21(1):67–78, 2011.

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter Bartlett, and Martin Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2916–2925. PMLR, 2019.

Weichao Mao, Lin Yang, Kaiqing Zhang, and Tamer Başar. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 15007–15049. PMLR, 2022.

Jun Moon and Tamer Başar. Discrete-time LQG mean field games with unreliable communication. In *53rd IEEE Conference on Decision and Control*, pages 2697–2702. IEEE, 2014.

Jun Moon and Tamer Başar. Linear quadratic risk-sensitive and robust mean field games. *IEEE Transactions on Automatic Control*, 62(3):1062–1077, 2016.

Sarah Perrin, Mathieu Laurière, Julien Pérolat, Matthieu Geist, Romuald Élie, and Olivier Pietquin. Mean field games flock! the reinforcement learning way. *arXiv preprint arXiv:2105.07933*, 2021.

Fabio S Priuli. First order mean field games in crowd dynamics. *arXiv preprint arXiv:1402.7296*, 2014.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and Q-learning. In *Conference on Learning Theory*, pages 3185–3205. PMLR, 2020.

Gonçalo dos Reis and Vadim Platonov. Forward utilities and mean-field games under relative performance concerns. In *From Particle Systems to Partial Differential Equations*, pages 227–251. Springer, 2019.

Naci Saldi, Tamer Başar, and Maxim Raginsky. Markov–Nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization*, 56(6):4256–4287, 2018.

Naci Saldi, Tamer Başar, and Maxim Raginsky. Approximate markov-nash equilibria for discrete-time risk-sensitive mean-field games. *Mathematics of Operations Research*, 45(4):1596–1620, 2020.

Devavrat Shah and Qiaomin Xie. Q-learning with nearest neighbors. *Advances in Neural Information Processing Systems*, 31, 2018.

Yoav Shoham, Rob Powers, and Trond Grenager. If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 171(7):365–377, 2007.

Ekhlas Sonu, Yingke Chen, and Prashant Doshi. Decision-theoretic planning under anonymity in agent populations. *Journal of Artificial Intelligence Research*, 59:725–770, 2017.

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

Jayakumar Subramanian and Aditya Mahajan. Reinforcement learning in stationary mean-field games. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 251–259, 2019.

Hamidou Tembine, Quanyan Zhu, and Tamer Başar. Risk-sensitive mean-field games. *IEEE Transactions on Automatic Control*, 59(4):835–850, 2013.

Noureddine Toumi, Roland Malhamé, and Jerome Le Ny. A tractable mean field game model for the analysis of crowd evacuation dynamics. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 1020–1025. IEEE, 2020.

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.

Qiaomin Xie, Zhuoran Yang, Zhaoran Wang, and Andreea Minca. Learning while playing in mean-field games: Convergence and optimality. In *International Conference on Machine Learning*, pages 11436–11447. PMLR, 2021.

Bora Yongacoglu, Gürdal Arslan, and Serdar Yüksel. Independent learning and subjectivity in mean-field games. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 2845–2850. IEEE, 2022.

Muhammad Aneeq Uz Zaman, Kaiqing Zhang, Erik Miehling, and Tamer Başar. Reinforcement learning in non-stationary discrete-time linear-quadratic mean-field games. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 2278–2284. IEEE, 2020.

Muhammad Aneeq Uz Zaman, Sujay Bhatt, and Tamer Başar. Adversarial linear-quadratic mean-field games over multigraphs. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 209–214. IEEE, 2021.

Muhammad Aneeq Uz Zaman, Erik Miehling, and Tamer Başar. Reinforcement learning for non-stationary discrete-time linear–quadratic mean-field games in multiple populations. *Dynamic Games and Applications*, pages 1–47, 2022.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021.

Supplementary Materials

A Proofs of Results in Section 4

Throughout this section the cardinalities of the state and action spaces are denoted by $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$, respectively.

A.1 Proof of Lemma 1

Proof. Let us consider exploration noise coefficient of the form $\psi_t^k = \psi$ for some $\psi \in (0, 1 - c_\pi)$, and c_π and θ defined in Section 3.2. The proof consists of two parts; in the first part we show that under the policy update (11) using learning coefficient ψ the probability of any action $a \in \mathcal{A}$ for any state $s \in \mathcal{S}$, will have a uniform lower bound $\pi_t^k(a \mid s) \geq \underline{\pi} > 0$ for all $k \in [K], 1 < t \leq T$. In the second step, we show that using this uniform lower bound and the communicating MDP assumption (Assumption 2), the sufficient exploration condition as shown in the statement of Lemma 1 holds.

To prove the lower bound $\underline{\pi}$, we start by proving a lower bound on $\pi_t^k(a \mid s)$ for a given episode $k \in [K]$ and $1 \leq t \leq T$. Recalling the update (11) for $1 < t \leq T$

$$\pi_t^k = (1 - c_{\pi,t}^k)\pi_{t-1}^k + c_{\pi,t}^k\big((1-\psi)\text{softmax}_\lambda(\cdot, Q_t^k) + \psi\mathbb{1}_{|\mathcal{A}|}\big)$$

where the last part on the right hand side is the uniform exploration noise $\mathbb{1}_{|\mathcal{A}|} = (\frac{1}{|\mathcal{A}|}, \dots, \frac{1}{|\mathcal{A}|})$. This can be written as follows:

$$\pi_t^k = \tilde{c}_{\pi,[1,t]}^k \pi_0^k + \sum_{l=1}^{t} c_{\pi,[l,t]}^k (1-\psi)\text{softmax}_\lambda(\cdot, Q_l^k) + \sum_{l=1}^{t} c_{\pi,[l,t]}^k \psi \mathbb{1}_{|\mathcal{A}|}, \tag{14}$$

where

$$\tilde{c}_{\pi,[l,t]}^k = \prod_{s=l}^{t}(1 - c_{\pi,s}^k), \quad c_{\pi,[l,t]}^k = c_{\pi,l}^k \prod_{s=l+1}^{t}(1 - c_{\pi,s}^k).$$

From the control policy update (14) we analyze the last part of the right-hand-side. In particular, we will prove that $\sum_{l=2}^{t+1} c_{\pi,[l,t+1]}^k \geq \sum_{l=2}^{t} c_{\pi,[l,t]}^k$ for $1 \leq t \leq T$. First we see that,

$$c_{\pi,[l,t]}^k = c_{\pi,l}^k(1 - c_{\pi,t}^k)\prod_{s=l+1}^{t-1}(1 - c_{\pi,s}^k) = (1 - c_{\pi,t}^k)c_{\pi,[l,t-1]}^k \tag{15}$$

Using this equality, we can show

$$\begin{aligned}\sum_{l=1}^{t+1} c_{\pi,[l,t+1]}^k &= c_{\pi,[t+1,t+1]}^k + \sum_{l=1}^{t} c_{\pi,[l,t+1]}^k = c_{\pi,t+1}^k + (1 - c_{\pi,t+1}^k)\sum_{l=1}^{t} c_{\pi,[l,t]}^k \\ &= c_{\pi,t+1}^k\left(1 - \sum_{l=1}^{t} c_{\pi,[l,t]}^k\right) + \sum_{l=1}^{t} c_{\pi,[l,t]}^k \geq \sum_{l=1}^{t} c_{\pi,[l,t]}^k \end{aligned} \tag{16}$$

where the second equality is obtained using (15). In the inequality we have used the fact that $\sum_{l=1}^{t} c_{\pi,[l,t]}^k \leq 1$ for all $1 \leq t \leq T$ which can be shown using inductive reasoning. The base case is true since $c_{\pi,[1,1]}^k = c_{\pi,1}^k = \frac{c_\pi}{k^\theta} < 1$ since $c_\pi \leq 1$ and $\zeta > 1$. Now assume that $\sum_{l=1}^{t} c_{\pi,[l,t]}^k \leq 1$; then,

$$\sum_{l=1}^{t+1} c_{\pi,[l,t+1]}^k = c_{\pi,t+1}^k + (1 - c_{\pi,t+1}^k)\sum_{l=1}^{t} c_{\pi,[l,t]}^k \leq c_{\pi,t+1}^k + 1 - c_{\pi,t+1}^k = 1$$

Hence we have proved that $\sum_{l=2}^{t} c_{\pi,[l,t]}^k \geq c_{\pi,[1,1]}^k = c_{\pi,1}^k$ for all $1 \leq t \leq T$. Let us define $\underline{\pi}_t^k := \min_{a\in\mathcal{A}, s\in\mathcal{S}} \pi_t^k(a \mid s)$ as the lower limit on exploration noise in policy π_t^k for $1 \leq t \leq T$. Using (14) and the definition of $\underline{\pi}_t^k$ and we can deduce

$$\underline{\pi}_t^k \geq \tilde{c}_{\pi,[1,t]}^k \underline{\pi}_0^k + \sum_{l=1}^{t} c_{\pi,[l,t]}^k \psi \geq \tilde{c}_{\pi,[1,t]}^k \underline{\pi}_0^k + c_{\pi,1}^k \psi / |\mathcal{A}|$$

for $1 \leq t \leq t$ where the second inequality uses $\sum_{l=1}^{t} c^k_{\pi,[l,t]} \geq c^k_{\pi,[1,1]} = c^k_{\pi,1}$ for all $1 \leq t \leq T$. This gives us a k-dependent lower bound on $\underline{\pi}^k_t$. Next we convert it into a uniform lower bound independent of k. Let us define a uniform lower bound dependent on k, $\underline{\pi}^k := \min_{1<t\leq T} \underline{\pi}^k_t$. Using (11), we can write

$$\begin{aligned}\pi^k_1 &= (1-c^k_{\pi,1})\pi^k_0 + c^k_{\pi,1}\big((1-\psi)\text{softmax}_\lambda(\cdot, Q^k_1) + \psi\mathbb{1}_{|\mathcal{A}|}\big)\\ &= (1-c^k_{\pi,1})\pi^{k-1}_T + c^k_{\pi,1}\big((1-\psi)\text{softmax}_\lambda(\cdot, Q^k_1) + \psi\mathbb{1}_{|\mathcal{A}|}\big)\end{aligned}$$

Using this relation and the definition of $\underline{\pi}^k$, we deduce that

$$\underline{\pi}^k \geq (1-c^k_{\pi,1})\underline{\pi}^{k-1} + c^k_{\pi,1}\psi/|\mathcal{A}|$$

Using recursion we can write this as

$$\underline{\pi}^k \geq \tilde{\vartheta}_{1,k}\underline{\pi}^1 + \sum_{l=1}^{k} \vartheta_{l,k}\psi/|\mathcal{A}|, \text{ where } \tilde{\vartheta}_{l,k} = \prod_{s=l}^{k}(1-c^s_{\pi,1}) \text{ and } \vartheta_{l,k} = c^l_{\pi,1}\tilde{\vartheta}_{l+1,k} \tag{17}$$

Next we will show that $\sum_{l=1}^{k} \vartheta_{l,k}\psi/|\mathcal{A}| \geq \vartheta_{1,1}\psi/|\mathcal{A}| = c_\pi\psi/|\mathcal{A}|$ for all $k \in [K]$ which coupled with (17) naturally leads to the conclusion that $\underline{\pi}^k \geq c_\pi\psi/|\mathcal{A}|$. By definition, we have

$$\begin{aligned}\sum_{l=1}^{k} \vartheta_{l,k}\psi/|\mathcal{A}| &= \vartheta_{k,k}\psi/|\mathcal{A}| + \sum_{l=1}^{k-1} \vartheta_{l,k}\psi/|\mathcal{A}| = c^k_{\pi,1}\psi/|\mathcal{A}| + (1-c^k_{\pi,1})\sum_{l=1}^{k-1} \vartheta_{l,k-1}\psi/|\mathcal{A}|\\ &= c^k_{\pi,1}\Big(\psi - \sum_{l=1}^{k-1}\vartheta_{l,k-1}\psi\Big)/|\mathcal{A}| + \sum_{l=1}^{k-1}\vartheta_{l,k-1}\psi/|\mathcal{A}| \geq \sum_{l=1}^{k-1}\vartheta_{l,k-1}\psi/|\mathcal{A}|\end{aligned}$$

where the last inequality follows from the fact that $\sum_{l=1}^{k} \vartheta_{l,k} \leq 1$ for all $k \in [K]$ which can be shown using inductive reasoning. The base case is true since $\vartheta_{1,1} = c_\pi < 1$. Now assume that $\sum_{l=1}^{k} \vartheta_{l,k} \leq 1$; then

$$\sum_{l=1}^{k+1} \vartheta_{l,k+1} = \vartheta_{k+1,k+1} + (1-c^{k+1}_{\pi,1})\sum_{l=1}^{k}\vartheta_{l,k} \leq c^{k+1}_{\pi,1} + 1 - c^{k+1}_{\pi,1} = 1$$

We have proved that $\sum_{l=1}^{k} \vartheta_{l,k}\psi \geq \vartheta_{1,1}\psi = c_\pi\psi$ for all $k \in [K]$, and hence using (17) we can deduce that $\underline{\pi}^k \geq c_\pi\psi/|\mathcal{A}|$ which implies $\pi^k_t(a|s) \geq \underline{\pi} := c_\pi\psi/|\mathcal{A}| > 0$ for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

Next we show how using this uniform lower bound $\underline{\pi}$ along with the communicating MDP assumption (Assumption 2) sufficient exploration can be proved. From Assumption 2 for any $i, j \in \mathcal{S}$ and $\tilde{\mu} \in S(\epsilon^{\text{net}})$ we know that there exists a finite horizon $H(\tilde{\mu})$ and a set of actions $\tilde{a}_1, \ldots, \tilde{a}_{H(\tilde{\mu})}$ such that

$$P^{\tilde{\mu}}_{ij} := P\big(s_{H(\tilde{\mu})} = j \mid a_1 = \tilde{a}_1, \ldots, a_{H(\tilde{\mu})} = \tilde{a}_{H(\tilde{\mu})}, s_1 = i, \mu = \tilde{\mu}\big) > 0. \tag{18}$$

Let us define $\underline{P} := \min_{i,j\in\mathcal{S},\tilde{\mu}\in S(\epsilon^{\text{net}})} P^{\tilde{\mu}}_{ij}$ and due to the finiteness of $\mathcal{S}$ and $S(\epsilon^{\text{net}})$ we can guarantee $\underline{P} > 0$. Due to the Lipschitzness of stochastic kernel P with respect to μ for any $s \in \mathcal{S}, a \in \mathcal{A}$, and $\mu, \mu' \in \Delta(\mathcal{S})$,

$$\|P(\cdot \mid s, a, \mu) - P(\cdot \mid s, a, \mu')\|_1 \leq L_\mu\|\mu - \mu'\|_1$$

for Lipschitz constant $L_\mu > 0$. This allows us to lower bound $P(s'|s, a, \mu^k_t)$ where μ^k_t evolves according to the update rule (10). Hence for any $s, s' \in \mathcal{S}, a \in \mathcal{A}$ and $\mu^k_t \in \Delta(\mathcal{S})$,

$$\begin{aligned}P(s'|s, a, \mu^k_t) &\geq P(s'|s, a, \mu^k_1) - |P(s'|s, a, \mu^k_t) - P(s'|s, a, \mu^k_1)|\\ &\geq P(s'|s, a, \mu^k_1) - L_\mu\sum_{k=2}^{\infty} c^k_{\mu,k}\\ &\geq P(s'|s, a, \mu^k_1) - \frac{L_\mu c_\mu}{2\varsigma}\end{aligned} \tag{19}$$

and we know that $\mu_1^k \in S(\epsilon^{\text{net}})$ due to the projection step in (10). As a result, the change in stochastic kernel $P(\cdot \mid s, a, \mu_t^k)$ can be bounded by controlling the inter-episodic learning coefficient ζ. Let us define the probability of reaching state $j \in \mathcal{A}$ from state $i \in \mathcal{S}$ in time $t \geq H := \max_{\tilde{\mu} \in S(\epsilon^{\text{net}})} H(\tilde{\mu})$ under the stochastic kernels induced by mean-field $(\mu_1^k, \dots, \mu_t^k)$ as $P(s_t = j | s_1 = i, (\mu_1^k, \dots, \mu_t^k))$. From Assumption 2 we know that there exists a set of actions $(\tilde{a}_1, \dots, \tilde{a}_t)$ such that $P(s_t = j | s_1 = i, (a_1 = \tilde{a}_1, \dots, a_t = \tilde{a}_t), (\mu_1^k, \dots, \mu_1^k)) \geq \underline{P}$. Using these facts we can deduce,

$$
\begin{aligned}
P(s_t = j | s_1 = i, (\mu_1^k, \dots, \mu_t^k)) &\geq P(s_t = j | s_1 = i, (a_1 = \tilde{a}_1, \dots, a_t = \tilde{a}_t), (\mu_1^k, \dots, \mu_t^k)) \\
&= \prod_{l=1}^{t-1} \sum_{s_1 = i, s_l \in \mathcal{S}, s_k = j} P(s_{l+1} = \tilde{s}_{l+1} | s_l = \tilde{s}_l, a_l = \tilde{a}_l, \mu_l^k) \pi_l^k (a_l = \tilde{a}_l | s_l = \tilde{s}_l) \\
&\geq \prod_{l=1}^{t-1} \sum_{s_1 = i, s_l \in \mathcal{S}, s_k = j} P(s_{l+1} = \tilde{s}_{l+1} | s_l = \tilde{s}_l, a_l = \tilde{a}_l, \mu_l^k) \underline{\pi} \\
&\geq \prod_{l=1}^{t-1} \Big(\sum_{s_1 = i, s_l \in \mathcal{S}, s_k = j} P(s_{l+1} = \tilde{s}_{l+1} | s_l = \tilde{s}_l, a_l = \tilde{a}_l, \mu_1^k) - |\mathcal{S}| \frac{L_\mu c_\mu}{2^\zeta} \Big) \underline{\pi} \\
&\geq \prod_{l=1}^{t-1} \sum_{s_1 = i, s_l \in \mathcal{S}, s_k = j} P(s_{l+1} = \tilde{s}_{l+1} | s_l = \tilde{s}_l, a_l = \tilde{a}_l, \mu_1^k) - C(t) |\mathcal{S}| \frac{L_\mu c_\mu}{2^\zeta} \underline{\pi} \\
&\geq \underline{P} - C(2H) |\mathcal{S}| \frac{L_\mu c_\mu}{2^\zeta} \underline{\pi}^{2H} \geq \underline{P}/2
\end{aligned}
$$

The first inequality is obtained using the fact that under exploration noise $\psi_t^k = \psi \in (0, 1 - c_\pi)$, $\pi_t^k(a|s)$ is lower bounded by $\underline{\pi}$. The second inequality follows from Lipschitzness of stochastic kernel P and inter-episodic learning coefficient $\zeta > 1$ as shown in (19). The third inequality is obtained by using the fact that probability $P(s'|s, a, \mu) \leq 1$ and for ζ high enough $|\mathcal{S}| \frac{L_\mu c_\mu}{2^\zeta} \leq 1$. The fourth inequality uses the lower bound on $P_{ij}^{\tilde{\mu}}$ for $i, j \in \mathcal{S}$ and $\tilde{\mu} \in S(\epsilon^{\text{net}})$ as shown in (18). The final bound uses the fact that for ζ large enough $C(2H)|\mathcal{S}| \frac{L_\mu c_\mu}{2^\zeta} \underline{\pi}^{2H} \leq \underline{P}/2$. □

A.2 Proof of Lemma 2

Proof. In this proof we provide finite sample convergence bounds for the transition probability estimation (8) under the slowly time-varying MC setting. The proof of Lemma 2 relies on Freedman's inequality Freedman (1975) similar to the analysis of Theorem 4 in Hsu et al. (2019). Our lemma generalizes transition probability estimation for the slowly time-varying MC setting. The proof relies on introducing a stochastic process Y_t (dependent on visitation of a fixed pair of states i, j) which is shown to be a Martingale difference sequence. The transition probability estimation error is shown to be a function of the sum of Y_t and a drift term due to the slowly time-varying MC setting. The drift term is shown to be small due to the Lipschitz property of transition dynamics and the slowly time-varying MC. Then, using Freedman's inequality, we show that the estimation error is monotonically decreasing with the visitation number of the pair of states i, j. Finally, we prove a high confidence lower bound on the visitation number of any pair of states i, j under the sufficient exploration condition (Lemma 1), yielding our convergence result.

We recall the definition of $\epsilon_P^k := \|\hat{P}_T^k - P_1^k\|_F$ where we use P_t^k as a shorthand for $P_{\pi_t^k, \mu_t^k}$. We use Freedman's inequality to obtain the estimation error for estimator $\hat{P}_T^k$ as in Hsu et al. (2019). Furthermore, since we are dealing with a single episode k we suppress the use of episode k for clarity. Let $\mathcal{F}_t$ be the σ-field generated by $\{s_1, \mu_1, \pi_1, \dots, s_t, \mu_t, \pi_t\}$. Let us start by fixing a pair of states (i, j) for any $i, j \in \mathcal{S}$. Next let us define a stochastic process Y_t such that $Y_1 := 0$ and for $t \geq 2$

$$
Y_t := \mathbb{1}\{s_{t-1} = i\} \big(\mathbb{1}\{s_t = j\} - P_{t-1}(i, j) \big)
$$

where $P_{t-1}(i, j)$ is the transition probability of going from state i to j from time $t-1$ to t. The stochastic process $(Y_t)_{t \in [T]}$ is a Martingale Difference Sequence since Y_t is $\mathcal{F}_t$-measurable, and for $t \geq 2$

$$
\begin{aligned}
\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] &= \mathbb{E}\big[\mathbb{1}\{s_{t-1} = i\} \big(\mathbb{1}\{s_t = j\} - P_{t-1}(i, j) \big) \mid \mathcal{F}_{t-1} \big], \\
&= \mathbb{1}\{s_{t-1} = i\} \Big(P_{t-1}(i, j) - P_{t-1}(i, j) \Big) = 0.
\end{aligned}
$$

Furthermore, $\forall t \in [T]$, $Y_t \in [-P_{t-1}(i,j), 1 - P_{t-1}(i,j)] \subset [-1,1]$. Summing up Y_t for $t \in [T]$,

$$
\begin{aligned}
S_T := \sum_{t=1}^{T} Y_t &= \sum_{t=2}^{T} \mathbb{1}\{s_{t-1} = i\}\big(\mathbb{1}\{s_t = j\} - P_{t-1}(i,j)\big), \\
&= \sum_{t=2}^{T} \mathbb{1}\{s_{t-1} = i\}\big(\mathbb{1}\{s_t = j\} - P_1(i,j) + P_1(i,j) - P_{t-1}(i,j)\big), \\
&= N_{i,j} - N_i P_1(i,j) + \sum_{t=2}^{T} \mathbb{1}\{s_{t-1} = i\}\tilde{P}_{t-1}(i,j), \qquad (20)
\end{aligned}
$$

where $\tilde{P}_t := P_1 - P_t$ is the drift in the true transition probability. For use in Freedman's inequality, consider the process

$$
\begin{aligned}
&\mathbb{E}[Y_t^2 \mid \mathcal{F}_{t-1}] = \mathbb{E}\big[\mathbb{1}\{s_{t-1} = i\}\big(\mathbb{1}\{s_t = j\}P_{t-1}(i,j) + P_{t-1}^2(i,j)\big) \mid \mathcal{F}_{t-1}\big], \\
&= \mathbb{1}\{s_{t-1} = i\}P_{t-1}(i,j)\big(1 - P_{t-1}(i,j)\big), \\
&= \mathbb{1}\{s_{t-1} = i\}\big(P_1(i,j) - \tilde{P}_{t-1}(i,j) - P_1^2(i,j) + 2P_1(i,j)\tilde{P}_{t-1}(i,j) - \tilde{P}_{t-1}^2(i,j)\big), \\
&= \mathbb{1}\{s_{t-1} = i\}P_1(i,j)\big(1 - P_1(i,j)\big) + \mathbb{1}\{s_{t-1} = i\}\tilde{P}_{t-1}(i,j)\big(2P_1(i,j) - 1 - \tilde{P}_{t-1}(i,j)\big).
\end{aligned}
$$

Since $\mathbb{E}[Y_t^2 \mid \mathcal{F}_{t-1}] \geq 0$, both terms in the above expression are positive. Hence its summation V_T will be

$$
V_T := \sum_{t=2}^{T} \mathbb{E}[Y_t^2 \mid \mathcal{F}_{t-1}] = N_i P_1(i,j)\big(1 - P_1(i,j)\big) + \sum_{t=2}^{T} \mathbb{1}\{s_{t-1} = i\}\tilde{P}_{t-1}(i,j)\big(2P_1(i,j) - 1 - \tilde{P}_{t-1}(i,j)\big). \qquad (21)
$$

Again both parts of the above expression are positive. Recalling (13) we write the estimation error as

$$
\begin{aligned}
\hat{P}_T(i,j) - P_1(i,j) &= \frac{N_{i,j} - N_i P_1(i,j) + 1/S - P_1(i,j)}{N_i + 1}, \\
&= \frac{N_{i,j} - N_i P_1(i,j)}{N_i + 1} + \frac{1/S - P_1(i,j)}{N_i + 1}, \\
&= \frac{S_T - \sum_{t=2}^{T} \mathbb{1}\{s_{t-1} = i\}\tilde{P}_{t-1}(i,j)}{N_i + 1} + \frac{1/S - P_1(i,j)}{N_i + 1},
\end{aligned}
$$

where the last equality is due to (20). Applying Corollary 1 from Hsu et al. (2019), which is based on Freedman's inequality, we get

$$
\begin{aligned}
&|\hat{P}_T(i,j) - P_1(i,j)| \\
&\leq \sqrt{\frac{2cV_T\tau_{T,\delta_P}}{(N_i+1)^2}} + \frac{4\tau_{T,\delta_P} + \sum_{t=2}^{T} \mathbb{1}\{s_{t-1} = i\}|\tilde{P}_{t-1}(i,j)| + |1/S - P_1(i,j)|}{N_i + 1}, \\
&= \Bigg(\frac{2cN_i P_1(i,j)\big(1 - P_1(i,j)\big)\tau_{T,\delta_P}}{(N_i+1)^2} \\
&\qquad\qquad + \frac{2c\sum_{t=2}^{T} \mathbb{1}\{s_{t-1} = i\}\tilde{P}_{t-1}(i,j)\big(2P_1(i,j) - 1 - \tilde{P}_{t-1}(i,j)\big)\tau_{T,\delta_P}}{(N_i+1)^2}\Bigg)^{\frac{1}{2}} \\
&\qquad\qquad\qquad + \frac{4\tau_{T,\delta_P} + \sum_{t=2}^{T} \mathbb{1}\{s_{t-1} = i\}|\tilde{P}_{t-1}(i,j)| + |1/S - P_1(i,j)|}{N_i + 1}, \\
&\leq \sqrt{\frac{2cN_i P_1(i,j)\big(1 - P_1(i,j)\big)\tau_{T,\delta_P}}{(N_i+1)^2}} + \frac{\sqrt{2c\sum_{t=2}^{T}|\tilde{P}_{t-1}(i,j)|\tau_{T,\delta_P}}}{N_i + 1} \\
&\qquad\qquad\qquad + \frac{4\tau_{T,\delta_P} + \sum_{t=2}^{T}|\tilde{P}_{t-1}(i,j)| + |1/S - P_1(i,j)|}{N_i + 1}, \\
&\leq \sqrt{\frac{2c\tau_{T,\delta_P}}{N_i + 1}} + \frac{\sqrt{2c\sum_{t=2}^{T}|\tilde{P}_{t-1}(i,j)|\tau_{T,\delta_P}}}{N_i + 1} + \frac{4\tau_{T,\delta_P} + \sum_{t=2}^{T}|\tilde{P}_{t-1}(i,j)| + |1/S - P_1(i,j)|}{N_i + 1}, \qquad (22)
\end{aligned}
$$

with probability at least $1 - \delta_P/(2S^2)$, where $\tau_{T,\delta_P} = \mathcal{O}(\log(\frac{2S^3 \log(T)}{\delta_P}))$. We used equation (21) to obtain the second inequality. Analyzing $|\tilde{P}_t(i,j)|$,

$$\begin{aligned} |\tilde{P}_t(i,j)| &\leq \|P_1 - P_t\|_F = \|P_{\pi_1,\mu_1} - P_{\pi_t,\mu_t}\|_F, \\ &\leq \sum_{l=1}^{t-1} \big(\|P_{\pi_l,\mu_l} - P_{\pi_l,\mu_{l+1}}\|_F + \|P_{\pi_l,\mu_{l+1}} - P_{\pi_{l+1},\mu_{l+1}}\|_F \big), \\ &\leq \sum_{l=1}^{t-1} \big(L_P^{\mu} \|\mu_{l+1} - \mu_l\|_1 + L_P^{\pi} \|\pi_{l+1} - \pi_l\|_{TV} \big) \\ &\leq \sum_{l=2}^{t} \big(L_P^{\mu} c_{\mu,l} + L_P^{\pi} c_{\pi,l} \big), \\ &\leq (L_P^{\mu} c_{\mu} + L_P^{\pi} c_{\pi}) \sum_{l=2}^{t} l^{-\zeta}, \\ &\leq \frac{L_P^{\mu} c_{\mu} + L_P^{\pi} c_{\pi}}{\zeta - 1} 2^{1-\zeta} = \tilde{L}_P 2^{1-\zeta}. \end{aligned} \tag{23}$$

where $c_{\mu,t} := c_{\mu} t^{-\zeta}$, $c_{\pi,t} := c_{\pi} t^{-\zeta}$ and $\tilde{L}_P := 10(L_P^{\mu} c_{\mu} + L_P^{\pi} c_{\pi})$ for $\zeta \geq 1.1$. The second inequality above is due to the Lipschitz conditions on P in Lemma 2 and the third inequality is due to the fact that $\|\mu\|_1, \|\pi\|_{TV} \leq 1$ for any $\mu \in \mathcal{P}(\mathcal{S})$ and $\pi \in \mathcal{P}$. Now the estimation error can be bounded using (22) and (23):

$$\begin{aligned} &|\hat{P}_T(i,j) - P_1(i,j)| \\ &\leq \sqrt{\frac{2c\tau_{T,\delta_P}}{N_i + 1}} + \frac{\sqrt{2c\tau_{T,\delta_P} \tilde{L}_P T}}{N_i + 1} 2^{\frac{1-\zeta}{2}} + \frac{4\tau_{T,\delta_P} + \tilde{L}_P T 2^{1-\zeta} + |1/S - P_1(i,j)|}{N_i + 1}, \end{aligned} \tag{24}$$

with probability at least $1 - \delta_P/(2S^2)$. Next we need to lower bound N_i. In the following lemma we show that due to the sufficient exploration condition, N_i grows at least linearly with T with high probability.

Lemma 4. *Using Lemma 1, $N_i \geq T/T_e$ with probability at least $1 - \delta_P/(2S^2)$, where*

$$T_e := \mathcal{O}\Big(\frac{1}{\sigma} \log\Big(\frac{2S^3}{\delta_P}\Big)\Big). \tag{25}$$

Proof. For a fixed state $i \in \mathcal{S}$, define event $\mathcal{E}^k$ such that $\sum_{t=1}^{kT_e} \mathbb{1}\{i_t = i\} \geq k$, for a given integer k such that $1 \leq k \leq K_e := \lceil T/T_e \rceil$. We show that $\mathcal{E}^K$ is a high probability event given the sufficient exploration condition (Lemma 1). For a given $i \in \mathcal{S}$, define a random variable,

$$X_t^i = \mathbb{I}\{i_t = i\} - \mathbb{E}[\mathbb{I}\{i_t = i\} \mid \mathcal{F}_{t-\tau}]$$

This random variable is an $\mathcal{F}_t$ adapted process with $\mathbb{E}[X_t^i \mid \mathcal{F}_{t-\tau}] = 0$ and $|X_t^i| \leq 1$. Let l be an integer $0 \leq l \leq \tau$. For a fixed l, define the process $Y_{l,k}^i = X_{k\tau+l}^i$ and define filtration $\tilde{\mathcal{F}}_{l,k} := \mathcal{F}_{k\tau+l}$. We can deduce that

$$\mathbb{E}[Y_{l,k}^i \mid \tilde{\mathcal{F}}_{l,k-1}] = \mathbb{E}[X_{k\tau+l}^i \mid \mathcal{F}_{k\tau+l-\tau}] = 0, |Y_{l,k}^i| \leq 1,$$

and $Y_{l,k}^i$ is $\tilde{\mathcal{F}}_{l,k}$ measurable. Combining these facts, $Y_{l,k}^i$ is a Martingale Difference Sequence. Using Azuma-Hoeffding inequality and Lemma 1 we can deduce that for a given $i \in \mathcal{S}$ and $k = K_e$ where $T_e := \mathcal{O}(\ln(2S^3/\delta_P)/\sigma)$, $\sum_{t=1}^{T} \mathbb{I}\{i_t = i\} \geq K_e$ with probability at least $1 - \delta_P/(2S^3)$. Taking a union bound over all $i \in \mathcal{S}$, we get $\sum_{t=0}^{T} \mathbb{I}\{i_t = i\} \geq K_e, \forall i \in \mathcal{S}$ with probability at least $1 - \delta_P/(2S^2)$. □

Using (24) and Lemma 4 the estimation error can be written as

$$\begin{aligned} &|\hat{P}_T(i,j) - P_1(i,j)| \\ &\leq \sqrt{\frac{2c\tau_{T,\delta_P} T_e}{T}} + \sqrt{\frac{2c\tau_{T,\delta_P} \tilde{L}_P}{T}} T_e 2^{\frac{1-\zeta}{2}} + \frac{(4\tau_{T,\delta_P} + |1/S - P_1(i,j)|) T_e}{T} + \tilde{L}_P T_e 2^{1-\zeta}, \end{aligned}$$

with probability at least $1-\delta_P/S^2$. Using a union bound over all pairs $(i,j)\in\mathcal{S}\times\mathcal{S}$, the definition of Frobenius norm and the equivalence between 1 and 2 vector norms,

$$\|\hat{P}-P_1\|_F=\tilde{\mathcal{O}}(T^{-1/2})+\tilde{\mathcal{O}}(T^{-1})+\mathcal{O}(2^{1-\zeta}).$$

with probability at least $1-\delta_P$. Hence we have completed the proof. □

A.3 Proof of Lemma 3

Proof. In this proof we provide finite sample convergence bounds for the Q-learning update (9) within the slowly time-varying MC setting. The proof of Lemma 3 follows an approach similar to proof of Theorem 4 in Qu and Wierman (2020) and extends the results to a slowly time-varying MC and learning exponent $0.5<\nu\le 1$, which is empirically observed to have better convergence properties. We also find that convergence cannot be guaranteed for $0<\nu\le 0.5$. The proof starts with breaking down the error ϵ_Q^k into several components. Then we obtain bounds on those components by proving certain properties like boundedness and the Martingale Difference Sequence property. Following that we prove that the error accumulated due to the slowly time-varying MC setting is small due to the Lipschitz properties of the transition probability and reward function. Finally combining all these results, the total error itself is shown to be converging using the contraction mapping property of the discounted Bellman update.

This proof uses the fact that the coefficient of the learning rate c_β in the Q-learning update (9) is lower bounded by $\frac{1}{\sigma}\max\left\{\nu+\zeta-1,\frac{1}{(1-\sqrt{\rho})}\right\}$. In this proof we suppress the use of superscript k since we are dealing with a single episode. Recall the definition of $\epsilon_Q:=\|Q_T-Q_1^*\|_\infty$ where $Q_1^*:=Q_{\mu_1}^*$ the optimal Q-function for the MDP induced by mean-field μ_1. The Q-learning update can be written down as:

$$Q_{t+1}=Q_t+\beta_t[e_{i_t}^T[F(\mu_t,Q_t)-Q_t]+w(t,\mu_t)]e_{i_t} \tag{26}$$

where $\beta_t=\frac{c_\beta}{(t+1)^\nu}$ and

$$\begin{aligned} w(t,\mu_t)&=\rho[\max_{a\in\mathcal{A}}Q_t(s_{t+1},a)-\mathbb{E}_{s'\sim P(\cdot|s_t,a_t,\mu_t)}[\max_{a'\in\mathcal{A}}Q_t(s',a')], \\ F(\mu_t,Q_t)(s,a)&=r(s,a,\mu_t)+\rho\mathbb{E}_{s'\sim P(\cdot|s_t,a_t,\mu_t)}[\max_{a'\in\mathcal{A}}Q_t(s',a')]\end{aligned} \tag{27}$$

The noise $w(\cdot,\cdot)$ is bounded by $\bar{w}$, is measurable with respect to $\mathcal{F}_{t+1}$ and $\mathbb{E}[w(t,\mu_t)\mid\mathcal{F}_t]=0$. We further decompose the update rule using $D_t:=\mathbb{E}[e_{i_t}^Te_{i_t}\mid\mathcal{F}_{t-\tau}]$. The matrix D_t is a diagonal matrix with elements $(d_{t,i})_{i\in\mathcal{S}\times\mathcal{A}}$, where $d_{t,i}=\mathbb{P}(i_t=i\mid\mathcal{F}_{t-\tau})$, and from the sufficient exploration condition (Lemma 1) we know that $d_{t,i}\ge\sigma>0$.

$$\begin{aligned} Q_{t+1}&=Q_t+\beta_tD_t(F(\mu_t,Q_t)-Q_t)+\beta_t(e_{i_t}^Te_{i_t}-D_t)(F(\mu_t,Q_t)-Q_t)+\beta_tw(t,\mu_t)e_{i_t}, \\ &=Q_t+\beta_tD_t(F(\mu_t,Q_t)-Q_t)+\beta_t(e_{i_t}^Te_{i_t}-D_t)(F(\mu_{t-\tau},Q_{t-\tau})-Q_{t-\tau})+\beta_tw(t,\mu_t)e_{i_t} \\ &\qquad+\beta_t(e_{i_t}^Te_{i_t}-D_t)(F(\mu_t,Q_t)-F(\mu_{t-\tau},Q_{t-\tau})-Q_t+Q_{t-\tau})\end{aligned}$$

Let us define

$$\begin{aligned} \epsilon_t&:=(e_{i_t}^Te_{i_t}-D_t)(F(\mu_{t-\tau},Q_{t-\tau})-Q_{t-\tau})+\beta_tw(t,\mu_t)e_{i_t}, \\ \phi_t&:=(e_{i_t}^Te_{i_t}-D_t)(F(\mu_t,Q_t)-F(\mu_{t-\tau},Q_{t-\tau})-Q_t+Q_{t-\tau})\end{aligned}$$

The process ϵ_t is $\mathcal{F}_{t+1}$ measurable and

$$\begin{aligned} &\mathbb{E}[\epsilon_t\mid\mathcal{F}_{t-\tau}] \\ &=\mathbb{E}[e_{i_t}^Te_{i_t}-D_t\mid\mathcal{F}_{t-\tau}](F(\mu_{t-\tau},Q_{t-\tau})-Q_{t-\tau})+\mathbb{E}[\mathbb{E}[w(t,\mu_t)\mid\mathcal{F}_t]e_{i_t}\mid\mathcal{F}_{t-\tau}]=0.\end{aligned}$$

Hence, ϵ_t is a shifted Martingale Difference Sequence. Writing down the Q-function as a sum from τ (Lemma 1) to t, we get

$$Q_{t+1}=\tilde{B}_{\tau-1,t}Q_\tau+\sum_{k=\tau}^{t}B_{k,t}F(\mu_k,Q_k)+\sum_{k=\tau}^{t}\beta_t\tilde{B}_{k,t}(\epsilon_k+\phi_k), \tag{28}$$

where $B_{k,t} = \beta_k D_k \prod_{l=k+1}^{t}(I - \beta_l D_l)$, $\tilde{B}_{k,t} = \prod_{l=k+1}^{t}(I - \beta_l D_l)$, and $B_{k,t}$ and $\tilde{B}_{k,t}$ are diagonal matrices composed of elements $b_{k,t,i}$ and $\tilde{b}_{k,t,i}$ respectively. We also define $\beta_{k,t}$ and $\tilde{\beta}_{k,t}$ such that

$$\beta_{k,t} := \beta_k \prod_{l=k+1}^{t} (1 - \beta_l \sigma) \geq b_{k,t,i}, \qquad \tilde{\beta}_{k,t} := \prod_{l=k+1}^{t} (1 - \beta_l \sigma) \geq \tilde{b}_{k,t,i}$$

Next we compute the optimality gap $e_t^Q = \|Q_t - Q_t^*\|_\infty$, where Q_t^* is the fixed point of the operator $F(\mu_t, \cdot)$.

Lemma 5.

$$\begin{aligned} e_{t+1}^Q \leq \tilde{B}_{\tau-1,t} e_\tau^Q + \rho \max_i \sum_{k=\tau}^{t} b_{k,t,i} e_k^Q + \Big\| \sum_{k=\tau}^{t} \beta_k \tilde{B}_{k,t}(\epsilon_k + \phi_k) \Big\|_\infty & \\ + L_Q^\mu \Big[\tilde{\beta}_{\tau-1,t} \sum_{l=\tau}^{t} c_{\mu,l} + \sum_{k=\tau}^{t} \beta_{k,t} \sum_{l=k}^{t} c_{\mu,l} \Big] & \end{aligned} \tag{29}$$

Proof. Using (28) and subtracting Q_{t+1}^* from both sides,

$$Q_{t+1} - Q_{t+1}^* = \tilde{B}_{\tau-1,t}(Q_\tau - Q_{t+1}^*) + \sum_{k=\tau}^{t} B_{k,t}(F(\mu_k, Q_k) - Q_{t+1}^*) + \sum_{k=\tau}^{t} \beta_t \tilde{B}_{k,t}(\epsilon_k + \phi_k)$$

Hence we get,

$$\begin{aligned} e_{t+1}^Q = \tilde{B}_{\tau-1,t} e_\tau^Q + \rho \sup_i \sum_{k=\tau}^{t} b_{k,t,i} e_k^Q + \Big\| \sum_{k=\tau}^{t} \beta_k \tilde{B}_{k,t}(\epsilon_k + \phi_k) \Big\|_\infty & \\ + \Big\| \tilde{B}_{\tau-1,t}(Q_\tau^* - Q_{t+1}^*) + \sum_{k=\tau}^{t} B_{k,t}(Q_k^* - Q_{t+1}^*) \Big\|_\infty & \end{aligned}$$

We can use the Simulation lemma and Lipschitzness of transition probability $P_{\pi,\mu}$ and reward function R_μ with respect to the mean-field μ (with corresponding constants L_P^μ and L_R^μ respectively), to prove Lipschitzness of Q^* with μ. Due to Lipschizness, we know that for $\mu, \mu' \in \mathcal{P}(\mathcal{S})$

$$\|P_{\pi,\mu} - P_{\pi,\mu'}\|_F \leq L_P^\mu \|\mu - \mu'\|_1, \qquad \|R_\mu - R_{\mu'}\|_\infty \leq L_R^\mu \|\mu - \mu'\|_1$$

and using the Simulation Lemma Lewis et al. (2012) we know that

$$\|V_\mu^* - V_{\mu'}^*\|_\infty \leq \Big(L_R^\mu + \frac{L_P^\mu}{2(1-\rho)} \Big) \|\mu - \mu'\|_1$$

where V_μ^* is the value function of the MDP induced by mean-field μ and $(1-\rho)^{-1}$ is an upper bound on the value functions due to bounded rewards. Hence

$$\begin{aligned} \|Q_\mu^*(s,a) - Q_{\mu'}^*(s,a)\|_\infty &= \rho\big(\langle P(\cdot \mid s,a,\mu), V_\mu^*(\cdot) \rangle - \langle P(\cdot \mid s,a,\mu'), V_{\mu'}^*(\cdot) \rangle \big), \\ &= \rho\big(\langle P(\cdot \mid s,a,\mu), V_\mu^*(\cdot) \rangle - \langle P(\cdot \mid s,a,\mu), V_{\mu'}^*(\cdot) \rangle \\ &\qquad + \langle P(\cdot \mid s,a,\mu), V_{\mu'}^*(\cdot) \rangle - \langle P(\cdot \mid s,a,\mu'), V_{\mu'}^*(\cdot) \rangle \big), \\ &\leq \rho \Big(L_R^\mu + \frac{L_P^\mu}{2(1-\rho)} \Big) \|\mu - \mu'\|_1 + \rho \frac{L_P^\mu}{2(1-\rho)} \|\mu - \mu'\|_1 = L_Q^\mu \|\mu - \mu'\|_1 \end{aligned}$$

where $L_Q^\mu := \rho(L_R^\mu + L_P^\mu/(1-\rho))$. And thus

$$\|Q_t^* - Q_{t+1}^*\|_\infty \leq L_Q^\mu \|\mu_t - \mu_{t+1}\|_1$$

now that we have shown the Lipschitzness of Q^* with respect to μ. Furthermore, as $\|\mu_t - \mu_{t+1}\|_1 \leq c_{\mu,t}$, where $c_{\mu,t} := \frac{c_\mu}{(t+1)^\varsigma}$, we get

$$\big\| \tilde{B}_{\tau-1,t}(Q_\tau^* - Q_{t+1}^*) + \sum_{k=\tau}^{t} B_{k,t}(Q_k^* - Q_{t+1}^*) \big\|_\infty \leq L_Q^\mu \Big[\tilde{\beta}_{\tau-1,t} \sum_{l=\tau}^{t} c_{\mu,l} + \sum_{k=\tau}^{t} \beta_{k,t} \sum_{l=k}^{t} c_{\mu,l} \Big]$$

This concludes the proof. □

We next start by bounding the terms ϵ_t and ϕ_t in the error decomposition (29).

Lemma 6.

$$\|\epsilon_t\|_\infty \leq \frac{2}{1-\rho} + C + \bar{w} =: \bar{\epsilon},$$
$$\|\phi_t\|_\infty \leq (L_R^\mu + \frac{L_P^\mu}{1-\rho})\sum_{k=1}^{\tau}\|\mu_{t-k+1} - \mu_{t-k}\|_1 + 2\bar{\epsilon}\sum_{k=1}^{\tau}\beta_{t-k}$$

Proof. Recalling the definition of ϵ_t,

$$\begin{aligned}\|\epsilon_t\|_\infty &= \|(e_{i_t}^T e_{i_t} - D_t)(F(\mu_{t-\tau}, Q_{t-\tau}) - Q_{t-\tau}) + \beta_t w(t, \mu_t) e_{i_t}\|_\infty,\\ &\leq \|e_{i_t}^T e_{i_t} - D_t\|_\infty \|F(\mu_{t-\tau}, Q_{t-\tau}) - Q_{t-\tau}\|_\infty + |w(t, \mu_t)|_\infty \|e_{i_t}\|_\infty,\\ &\leq \|F(\mu_{t-\tau}, Q_{t-\tau})\|_\infty + \|Q_{t-\tau}\|_\infty + \bar{w} \leq \frac{2}{1-\rho} + C + \bar{w} =: \bar{\epsilon}\end{aligned}$$

where $C = 1$ and $\bar{w} = \frac{2}{1-\rho}$ due to $\|Q_t\|_\infty \leq \frac{1}{1-\rho}$, contractive property of F and the definitions of noise w (27) and Q-update (26). Similarly for ϕ we get

$$\begin{aligned}\|\phi_t\|_\infty &= \|(e_{i_t}^T e_{i_t} - D_t)(F(\mu_t, Q_t) - F(\mu_{t-\tau}, Q_{t-\tau}) - Q_t + Q_{t-\tau})\|_\infty,\\ &\leq \|F(\mu_t, Q_t) - F(\mu_{t-\tau}, Q_{t-\tau})\|_\infty + \|Q_{t-\tau} - Q_t\|_\infty,\\ &\leq \sum_{k=1}^{\tau}\|F(\mu_{t-k+1}, Q_{t-k+1}) - F(\mu_{t-k}, Q_{t-k})\|_\infty + \sum_{k=1}^{\tau}\|Q_{t-k+1} - Q_{t-k}\|_\infty. \end{aligned} \tag{30}$$

We first analyze the first summand in equation (30)

$$\begin{aligned}&\|F(\mu_{t-k+1}, Q_{t-k+1}) - F(\mu_{t-k}, Q_{t-k})\|_\infty\\ &\leq \|F(\mu_{t-k+1}, Q_{t-k+1}) - F(\mu_{t-k+1}, Q_{t-k})\|_\infty + \|F(\mu_{t-k+1}, Q_{t-k}) - F(\mu_{t-k}, Q_{t-k})\|_\infty,\\ &\leq \rho\|Q_{t-k+1} - Q_{t-k}\|_\infty + \max_{s,a}|R(s, a, \mu_{t-k+1}) - R(s, a, \mu_{t-k})|\\ &\qquad\qquad + \max_{s,a}|P(\cdot \mid s, a, \mu_{t-k+1}) - P(\cdot \mid s, a, \mu_{t-k})|\frac{1}{1-\rho},\\ &\leq \rho\|Q_{t-k+1} - Q_{t-k}\|_\infty + \Big(L_R^\mu + \frac{L_P^\mu}{1-\rho}\Big)\|\mu_{t-k+1} - \mu_{t-k}\|_1\end{aligned} \tag{31}$$

Similarly the second summand in (30) is

$$\begin{aligned}\|Q_{t-k+1} - Q_{t-k}\|_\infty &= \beta_{t-k}\|e_{i_{t-k}}^T\big(F(\mu_{t-k}, Q_{t-k}) - Q_{t-k} + w(t-k, \mu_{t-k})\big)e_{i_{t-k}}\|_\infty,\\ &\leq \beta_{t-k}\|F(\mu_{t-k}, Q_{t-k})\|_\infty + \beta_{t-k}\|Q_{t-k}\|_\infty + \bar{w} \leq \beta_{t-k}\bar{\epsilon}.\end{aligned} \tag{32}$$

Substituting (31), (32) into (30)

$$\|\phi_t\|_\infty \leq (L_R^\mu + \frac{L_P^\mu}{1-\rho})\sum_{k=1}^{\tau}\|\mu_{t-k+1} - \mu_{t-k}\|_1 + 2\bar{\epsilon}\sum_{k=1}^{\tau}\beta_{t-k}.$$

□

Having proved bounds on ϵ_t and ϕ_t, we now prove some properties of the learning rates $c_{\mu,t} := \frac{c_\mu}{(t+1)^\zeta}$ and $\beta_t = \frac{c_\beta}{(t+1)^\nu}$ where $0.5 < \nu \leq 1$, $\zeta > 1$ and $c_\beta \geq \frac{\nu}{\sigma}$.

Lemma 7. *Below we present some results regarding the learning rate β_t and the associated variables.*

1. $\tilde{\beta}_{k,t} \leq \left(\frac{k+2}{t+2}\right)^{c_\beta\sigma} \leq \left(\frac{k+2}{t+2}\right)^{\nu}$,

2. $\beta_{k,t} \leq \frac{c_\beta}{(k+1)^\nu}\left(\frac{k+2}{t+2}\right)^{c_\beta\sigma} \leq 2\frac{c_\beta}{(t+2)^\nu}$,

3. $\sum_{k=1}^{t} \beta_{k,t}^2 \leq \frac{2c_\beta^2}{2c_\beta\sigma - 2\nu + 1} \frac{1}{(t+2)^{2\nu-1}}$,

4. $\sum_{k=\tau}^{t} \beta_{k,t} \sum_{l=k-\tau}^{k-1} \beta_l \leq \frac{2c_\beta^2(\tau+1)^\nu \tau}{1+c_\beta\sigma-2\nu} \frac{1}{(t+2)^{2\nu-1}}$

Proof. For part (1) we start by recalling the definition of $\tilde{\beta}_{k,t}$ for $k \in [t]$

$$\begin{aligned}
\tilde{\beta}_{k,t} &= \prod_{l=k+1}^{t} (1-\beta_l\sigma) \leq \prod_{l=k+1}^{t} 1 - \frac{c_\beta}{l+1} = \prod_{l=k+1}^{t} e^{\log(1-\frac{c_\beta}{l+1})}, \\
&\leq \prod_{l=k+1}^{t} e^{-\frac{c_\beta}{l+1}} = e^{-\sum_{l=k+1}^{t}\frac{c_\beta}{l+1}} = \exp\left(- \sum_{l=k+1}^{t} \frac{c_\beta}{l+1} \right), \\
&\leq \exp\left(- \int_{k+1}^{t+1} \frac{c_\beta\sigma}{y+1} dy \right) = \exp\left(- c_\beta\sigma \log\left(\frac{t+2}{k+2}\right)\right), \\
&= \left(\frac{k+2}{t+2}\right)^{c_\beta\sigma} \leq \left(\frac{k+2}{t+2}\right)^{\nu}.
\end{aligned}$$

The first inequality is due to the fact that $\beta_t := \frac{c_\beta}{(t+1)^\nu} > \frac{c_\beta}{t+1}$ since $\nu < 1$. The last inequality is due to the fact that $c_\beta\sigma \geq \nu$ and $k \leq t$.

For part (2), recalling the definition of $\beta_{k,t}$ and using the bound on $\tilde{\beta}_{k,t}$, we get

$$\beta_{k,t} = \beta_k\tilde{\beta}_{k,t} \leq \frac{c_\beta}{(k+1)^\nu}\left(\frac{k+2}{t+2}\right)^{c_\beta\sigma} \leq 2\frac{c_\beta}{(t+2)^\nu}.$$

For part (3), analyzing each summand

$$\begin{aligned}
\beta_{k,t}^2 &\leq \frac{c_\beta^2}{(k+1)^{2\nu}}\left(\frac{k+2}{t+2}\right)^{2c_\beta\sigma} = \frac{c_\beta^2}{(t+2)^{2c_\beta\sigma}}\frac{(k+2)^{2c_\beta\sigma}}{(k+1)^{2\nu}}, \\
&\leq \frac{2c_\beta^2}{(t+2)^{2c_\beta\sigma}}(k+1)^{2c_\beta\sigma-2\nu}
\end{aligned}$$

Substituting into the sum

$$\begin{aligned}
\sum_{k=1}^{t}\beta_{k,t}^2 &\leq \sum_{k=1}^{t}\frac{2c_\beta^2}{(t+2)^{2c_\beta\sigma}}(k+1)^{2c_\beta\sigma-2\nu} \leq \frac{2c_\beta^2}{(t+2)^{2c_\beta\sigma}}\int_{1}^{t+1}(y+1)^{2c_\beta\sigma-2\nu}dy, \\
&\leq \frac{2c_\beta^2}{(t+2)^{2c_\beta\sigma}}\frac{(t+2)^{2c_\beta\sigma-2\nu+1}}{2c_\beta\sigma-2\nu+1} = \frac{2c_\beta^2}{2c_\beta\sigma-2\nu+1}\frac{1}{(t+2)^{2\nu-1}}
\end{aligned}$$

For part (4), as $k-\tau \leq l \leq k-1$, in the expression $\sum_{k=\tau}^{t}\beta_{k,t}\sum_{l=k-\tau}^{k-1}\beta_l$, we get

$$\beta_l = \frac{c_\beta}{(l+1)^\nu} \leq \frac{c_\beta}{(k-\tau+1)^\nu} \leq \frac{c_\beta(\tau+1)^\nu}{(k+1)^\nu}$$

The summation can be written down as

$$\begin{aligned}
\sum_{k=\tau}^{t}\beta_{k,t}\sum_{l=k-\tau}^{k-1}\beta_l &\leq \sum_{k=\tau}^{t}\beta_{k,t}\frac{c_\beta\tau(\tau+1)^\nu}{(k+1)^\nu} \leq \sum_{k=\tau}^{t}\frac{c_\beta\tau}{(k+1)^\nu}\left(\frac{k+2}{t+2}\right)^{c_\beta\sigma}\frac{c_\beta(\tau+1)^\nu}{(k+1)^\nu}, \\
&\leq \sum_{k=\tau}^{t}\frac{2c_\beta^2(\tau+1)^\nu\tau}{(t+2)^{c_\beta\sigma}}(k+1)^{c_\beta\sigma-2\nu} \leq \frac{2c_\beta^2(\tau+1)^\nu\tau}{(t+2)^{c_\beta\sigma}}\int_{\tau}^{t+1}(y+1)^{2\nu-c_\beta\sigma}dy, \\
&\leq \frac{2c_\beta^2(\tau+1)^\nu\tau}{(t+2)^{c_\beta\sigma}}\frac{(t+2)^{1+c_\beta\sigma-2\nu}}{1+c_\beta\sigma-2\nu} \leq \frac{2c_\beta^2(\tau+1)^\nu\tau}{1+c_\beta\sigma-2\nu}\frac{1}{(t+2)^{2\nu-1}}.
\end{aligned}$$

Hence the inequalities have been proved. □

Having proved some properties of the learning rates in Lemma 7 we are now able to bound the two parts of the quantity $\Big\|\sum_{k=\tau}^{t}\beta_k\tilde{B}_{k,t}(\epsilon_k+\phi_k)\Big\|_\infty$ as follows. The bound on the first quantity relies on the properties of the learning rates and the second bound relies on the fact that ϵ_t is a Martingale Difference sequence.

Lemma 8.

$$\Big\|\sum_{k=\tau}^{t}\beta_k\tilde{B}_{k,t}\phi_k\Big\|_\infty \leq \frac{C_\phi^1}{(t+2)^{2\nu-1}} + \frac{C_\phi^2}{(t+2)^{\zeta+\nu-1}},$$
$$\Big\|\sum_{k=\tau}^{t}\beta_k\tilde{B}_{k,t}\epsilon_k\Big\|_\infty \leq \frac{C_\epsilon}{(t+2)^{\nu-1/2}}$$

with probability at least $1-\delta_Q$*, where*

$$C_\phi^1 = \frac{4c_\beta^2(1+\tau)^\nu\tau}{1+c_\beta\sigma-2\nu}\bar{\epsilon}, \tag{33}$$
$$C_\phi^2 = \Big(L_R^\mu + \frac{L_P^\mu}{1-\rho}\Big)\frac{2c_\mu c_\beta\tau(1+\tau)^\zeta}{c_\beta\sigma-\nu-\zeta+1}, \tag{34}$$
$$C_\epsilon = \frac{10\bar{\epsilon}}{\sqrt{2c_\beta\sigma-2\nu+1}}\sqrt{(\tau+1)c_\beta^2\log\Big(\frac{2(\tau+1)T^2SA}{\delta_Q}\Big)}. \tag{35}$$

Proof. We start with the first inequality

$$\begin{aligned}\Big\|\sum_{k=\tau}^{t}\beta_k\tilde{B}_{k,t}\phi_k\Big\|_\infty &\leq \sum_{k=\tau}^{t}\beta_{k,t}\|\phi_k\|_\infty\\ &\leq \sum_{k=\tau}^{t}\beta_{k,t}\Big(\Big(L_R^\mu+\frac{L_P^\mu}{1-\rho}\Big)\sum_{l=1}^{\tau}\|\mu_{k-l+1}-\mu_{k-l}\|_\infty + 2\bar{\epsilon}\sum_{l=1}^{\tau}\beta_{k-l}\Big),\\ &\leq 2\bar{\epsilon}\sum_{k=\tau}^{t}\beta_{k,t}\sum_{l=k-\tau}^{k-1}\beta_l + \Big(L_R^\mu+\frac{L_P^\mu}{1-\rho}\Big)\sum_{k=tau}^{t}\beta_{k,t}\sum_{l=k-\tau}^{k-1}\|\mu_{l+1}-\mu_l\|_\infty,\\ &\leq C_\phi^1\frac{1}{(t+2)^{2\nu-1}} + \Big(L_R^\mu+\frac{L_P^\mu}{1-\rho}\Big)\sum_{k=tau}^{t}\beta_{k,t}\sum_{l=k-\tau}^{k-1}c_{\mu,l},\end{aligned} \tag{36}$$

Now we obtain an upper bound for the expression $\sum_{k=\tau}^{t}\beta_{k,t}\sum_{l=k-\tau}^{k-1}c_{\mu,l}$. Due to the fact that $k-\tau\leq l\leq k-1$ in the expression $\sum_{k=tau}^{t}\beta_{k,t}\sum_{l=k-\tau}^{k-1}c_{\mu,t}$ and thus $c_{\mu,l}=\frac{c_\mu}{(l+1)^\zeta}\leq\frac{c_\mu(\tau+1)^\zeta}{(k+1)^\zeta}$

$$\begin{aligned}\sum_{k=tau}^{t}\beta_{k,t}\sum_{l=k-\tau}^{k-1}c_{\mu,l} &\leq \sum_{k=\tau}^{t}\beta_{k,t}\frac{c_\mu(\tau+1)^\zeta\tau}{(k+1)^\zeta} \leq \sum_{k=\tau}^{t}\frac{c_\mu\tau}{(k+1)^\zeta}\left(\frac{k+2}{t+2}\right)^{c_\beta\sigma}\frac{c_\beta(\tau+1)^\zeta}{(k+1)^\nu},\\ &\leq 2\sum_{k-\tau}^{t}\frac{c_\mu c_\beta\tau(\tau+1)^\zeta}{(t+2)^{c_\beta\sigma}}\frac{1}{(k+1)^{\nu+\zeta-c_\beta\sigma}},\\ &\leq \frac{c_\mu c_\beta\tau(\tau+1)^\zeta}{(t+2)^{c_\beta\sigma}}\frac{(t+2)^{c_\beta\sigma-\nu-\zeta+1}}{c_\beta\sigma-\nu-\zeta+1} \leq \frac{c_\mu c_\beta\tau(\tau+1)^\zeta}{c_\beta\sigma-\nu-\zeta+1}\frac{1}{(t+2)^{\nu+\zeta-1}},\end{aligned} \tag{37}$$

since $c_\beta\sigma\geq\nu+\zeta-1$. Substituting (37) into (36) we get

$$\Big\|\sum_{k=\tau}^{t}\beta_k\tilde{B}_{k,t}\phi_k\Big\|_\infty \leq \frac{C_\phi^1}{(t+2)^{2\nu-1}} + \frac{C_\phi^2}{(t+2)^{\zeta+\nu-1}}.$$

Next we move to the second inequality. Recalling the definition of ϵ_t

$$\epsilon_t = (e_{i_t}^T e_{i_t} - D_t)(F(\mu_{t-\tau},Q_{t-\tau}) - Q_{t-\tau}) + \beta_t w(t,\mu_t)e_{i_t}$$

which is $\mathcal{F}_{t+1}$ shifted martingale difference sequence, $\mathbb{E}[\epsilon_t \mid \mathcal{F}_t - \tau] = 0$. We will use a variant of the Azuma-Hoeffding bound which can handle *shifted* Martingale Difference Sequences Qu and Wierman (2020). Each element in the vector $\sum_{k=\tau}^{t} \beta_k \tilde{B}_{k,t} \epsilon_k$ can be upper bounded by $|\sum_{k=\tau}^{t} \beta_k \epsilon_{i,k} \tilde{b}_{k,t,i}|$ where $\epsilon_{i,k}$ is the ith element in the vector ϵ_k. Using Lemmas 13 & 14 from Qu and Wierman (2020) we get

$$\begin{aligned}
\Big|\sum_{k=\tau}^{t} \beta_k \epsilon_{i,k} \tilde{b}_{k,t,i}\Big| = \Big|\sum_{k=\tau}^{t} \beta_k \epsilon_{i,k} \prod_{l=k+1}^{t} (1-\beta_l d_{l,i})\Big| &\leq \sup_{\tau \leq k_0 < t} \left(\Big|\sum_{k=k_0+1}^{t} \beta_{k,t} \epsilon_{i,k}\Big| + 2\bar{\epsilon}\beta_{k_0,t}\right), \\
&\leq \bar{\epsilon}\sqrt{2(\tau+1)\sum_{k=\tau+1}^{t} \beta_{k,t}^2 \log\left(\frac{2(\tau+1)tSA}{\delta_Q}\right)} + \sup_{\tau \leq k_0 \leq t} 2\bar{\epsilon}\beta_{k_0,t}, \\
&\leq \frac{2\bar{\epsilon}}{\sqrt{2c_\beta\sigma - 2\nu + 1}}\sqrt{\frac{(\tau+1)c_\beta^2}{(t+2)^{2\nu-1}} \log\left(\frac{2(\tau+1)tSA}{\delta_Q}\right)} \\
&\qquad + \sup_{\tau \leq k_0 \leq t} 2\bar{\epsilon}\frac{c_\beta}{(k_0+1)^\nu}\left(\frac{k_0+2}{t+2}\right)^{c_\beta\sigma}, \\
&\leq \frac{2\bar{\epsilon}}{\sqrt{2c_\beta\sigma - 2\nu + 1}}\sqrt{\frac{(\tau+1)c_\beta^2}{(t+2)^{2\nu-1}} \log\left(\frac{2(\tau+1)tSA}{\delta_Q}\right)} + 4\bar{\epsilon}\frac{c_\beta}{(t+2)^\nu}, \\
&\leq \frac{10\bar{\epsilon}}{\sqrt{2c_\beta\sigma - 2\nu + 1}}\sqrt{\frac{(\tau+1)c_\beta^2}{(t+2)^{2\nu-1}} \log\left(\frac{2(\tau+1)tSA}{\delta_Q}\right)}.
\end{aligned}$$

with probability at least $1 - \delta_Q/SA$. Applying the union bound over $\forall i \in \mathcal{S} \times \mathcal{A}$, we get

$$\Big\|\sum_{k=\tau}^{t} \beta_k \tilde{B}_{k,t} \epsilon_k\Big\|_\infty \leq \frac{C_\epsilon}{(t+2)^{\nu-1/2}}, \quad C_\epsilon = \frac{10\bar{\epsilon}}{\sqrt{2c_\beta\sigma - 2\nu + 1}}\sqrt{(\tau+1)c_\beta^2 \log\left(\frac{2(\tau+1)tSA}{\delta}\right)}$$

with probability at least $1 - \delta_Q$. □

Now we aim to bound the last term in (29)

Lemma 9. *If $\zeta > 1$, then*

$$\tilde{\beta}_{\tau-1,t}\sum_{l=\tau}^{t} c_{\mu,l} + \sum_{k=\tau}^{t} \beta_{k,t} \sum_{l=k}^{t} c_{\mu,l} \leq \frac{c_\mu}{\zeta - 1}\frac{1}{\tau^{\zeta-1}}$$

Proof.

$$\begin{aligned}
\tilde{\beta}_{\tau-1,t}\sum_{l=\tau}^{t} c_{\mu,l} + \sum_{k=\tau}^{t} \beta_{k,t} \sum_{l=k}^{t} c_{\mu,l} &\leq \tilde{\beta}_{\tau-1,t}\sum_{l=\tau}^{t} c_{\mu,l} + \sum_{k=\tau}^{t} \beta_{k,t} \sum_{l=\tau}^{t} c_{\mu,l} \\
&\leq \sum_{l=\tau}^{t} c_{\mu,l} \\
&\leq \frac{c_\mu}{\zeta - 1}\frac{1}{\tau^{\zeta-1}}.
\end{aligned}$$

□

Now that we have bounded all the terms in (29) we will show that the error term e_t^Q can be bounded by a decreasing function of time t. Toward this end we introduce a lemma that will help us with the proof of the main result.

Lemma 10. *For any $0 < w < 1$ and $t \geq \tau$,*

$$e_t := \sum_{k=\tau}^{t} b_{k,t,i}\frac{1}{(k+1)^w} \leq \frac{1}{\sqrt{\rho}(t+2)^w}, \qquad g_t := \sum_{k=\tau}^{t} b_{k,t,i}\frac{1}{\tau^{\zeta-1}} \leq \frac{1}{\sqrt{\rho}\tau^{\zeta-1}}$$

Proof. Recall that $b_{k,t,i} = \beta_k d_{k,i} \prod_{l=k+1}^{t}(1 - \beta_l d_{l,i})$. We first prove the inequality for e_t by recursion. We start with the base case.

$$\begin{aligned}
e_\tau &= b_{\tau,\tau,i}\frac{1}{(\tau+1)^w} = \beta_\tau d_{\tau,i}\frac{1}{(\tau+1)^w},\\
&= \beta_\tau d_{\tau,i}\left(\frac{\tau+2}{\tau+1}\right)^w \frac{1}{(\tau+2)^w} = \beta_\tau d_{\tau,i}\left(1+\frac{1}{\tau+1}\right)^w \frac{1}{(\tau+2)^w}
\end{aligned}$$

and since τ is chosen such that $\left(1+\frac{1}{\tau+1}\right)^w \leq \frac{1}{\sqrt{\rho}}$ for $w \leq 1$, we have

$$e_\tau \leq \frac{1}{\sqrt{\rho}(\tau+2)^w}.$$

Now assume that for some $t > \tau$, $e_{t-1} \leq \frac{1}{\sqrt{\rho}(t+1)^w}$. Then

$$\begin{aligned}
e_t &= \sum_{k=\tau}^{t-1} b_{k,t,i}\frac{1}{(k+1)^w} + b_{t,t,i}\frac{1}{(t+1)^w},\\
&= (1-\beta_t d_{t,i})\sum_{k=\tau}^{t-1} b_{k,t-1,i}\frac{1}{(k+1)^w} + \beta_t d_{t,i}\frac{1}{(t+1)^w},\\
&= (1-\beta_t d_{t,i})e_{t-1} + \beta_t d_{t,i}\frac{1}{(t+1)^w},\\
&\leq (1-\beta_t d_{t,i})\frac{1}{\sqrt{\rho}(t+1)^w} + \beta_t d_{t,i}\frac{1}{(t+1)^w},\\
&= \frac{1-\beta_t d_{t,i}(1-\sqrt{\rho})}{\sqrt{\rho}(t+1)^w},\\
&\leq \left(1-\frac{c_\beta \sigma}{(t+1)^\nu}(1-\sqrt{\rho})\right)\frac{1}{\sqrt{\rho}(t+1)^w},\\
&= \left(1-\frac{c_\beta \sigma}{(t+1)^\nu}(1-\sqrt{\rho})\right)\frac{(t+2)^w}{(t+1)^w}\frac{1}{\sqrt{\rho}(t+2)^w},\\
&= \left(1-\frac{c_\beta \sigma}{(t+1)^\nu}(1-\sqrt{\rho})\right)\left(1+\frac{1}{t+1}\right)^w\frac{1}{\sqrt{\rho}(t+2)^w}.
\end{aligned}$$

For any $x > -1$, $(1+x) \leq e^x$ and thus

$$\left(1-\frac{c_\beta \sigma}{(t+1)^\nu}(1-\sqrt{\rho})\right)\left(1+\frac{1}{t+1}\right)^w \leq e^{-\frac{c_\beta\sigma}{(t+1)^\nu}(1-\sqrt{\rho})+\frac{w}{t+1}} \leq 1,$$

where the last inequality is due to $c_\beta \geq \frac{1}{(1-\sqrt{\rho})\sigma}$. Hence we have proved that

$$e_t \leq \frac{1}{\sqrt{\rho}(t+2)^w}.$$

Now we prove the inequality for g_t using recursion again. For the base case it is easy to see that

$$g_\tau = b_{\tau,\tau,i}\frac{1}{\tau^{\zeta-1}} = \beta_\tau d_{\tau,i}\frac{1}{\tau^{\zeta-1}} \leq \frac{1}{\sqrt{\rho}\tau^{\zeta-1}}.$$

Now assume that $g_{t-1} \leq \frac{1}{\sqrt{\rho}\tau^{\zeta-1}}$. Then

$$\begin{aligned}
g_t &= (1-\beta_t d_{t,i})g_{t-1} + \beta_t d_{t,i}\frac{1}{\tau^{\zeta-1}} \leq (1-\beta_t d_{t,i})\frac{1}{\sqrt{\rho}\tau^{\zeta-1}} + \beta_t d_{t,i}\frac{1}{\tau^{\zeta-1}},\\
&\leq \frac{1-\beta_t d_{t,i}(1-\sqrt{\rho})}{\sqrt{\rho}\tau^{\zeta-1}} \leq \left(1-\frac{c_\beta\sigma}{(t+1)^\nu}(1-\sqrt{\rho})\right)\frac{1}{\sqrt{\rho}\tau^{\zeta-1}} \leq \frac{1}{\sqrt{\rho}\tau^{\zeta-1}},
\end{aligned}$$

which proves the recursion step and completes the proof. □

Now we prove the main result using Lemma 10. Recalling (29),

$$e^Q_{t+1} \leq \tilde{B}_{\tau-1,t} e^Q_\tau + \rho \sup_i \sum_{k=\tau}^{t} b_{k,t,i} e^Q_k + \Big\| \sum_{k=\tau}^{t} \beta_k \tilde{B}_{k,t} (\epsilon_k + \phi_k) \Big\|_\infty \\ + L^\mu_Q \Big[\tilde{\beta}_{\tau-1,t} \sum_{l=\tau}^{t} c_{\mu,l} + \sum_{k=\tau}^{t} \beta_{k,t} \sum_{l=k}^{t} c_{\mu,l} \Big] \tag{38}$$

Using Lemmas 8 and 9 and using $C_\mu := 10 L^\mu_Q c_\mu$ for $\zeta \geq 1.1$,

$$e^Q_{t+1} \leq \tilde{B}_{\tau-1,t} e^Q_\tau + \rho \sup_i \sum_{k=\tau}^{t} b_{k,t,i} e^Q_k + \frac{C^1_\phi}{(t+2)^{2\nu-1}} + \frac{C^2_\phi}{(t+2)^{\zeta+\nu-1}} + \frac{C_\epsilon}{(t+2)^{\nu-1/2}} + \frac{C_\mu}{\tau^{\zeta-1}} \tag{39}$$

We will prove that $e^Q_t \leq \frac{\bar{C}_1}{(t+1)^{2\nu-1}} + \frac{\bar{C}_2}{(t+1)^{\zeta+\nu-1}} + \frac{\bar{C}_3}{(t+1)^{\nu-1/2}} + \frac{\bar{C}_4}{\tau^{\zeta-1}}$ using induction. The base case is trivially true; now assume this to be true for t:

$$\begin{aligned} e^Q_{t+1} &\leq \tilde{B}_{\tau-1,t} e^Q_\tau + \rho \sup_i \sum_{k=\tau}^{t} b_{k,t,i} \Big(\frac{\bar{C}_1}{(t+1)^{2\nu-1}} + \frac{\bar{C}_2}{(t+1)^{\zeta+\nu-1}} + \frac{\bar{C}_3}{(t+1)^{\nu-1/2}} + \frac{\bar{C}_4}{\tau^{\zeta-1}} \Big) \\ &\qquad + \frac{C^1_\phi}{(t+2)^{2\nu-1}} + \frac{C^2_\phi}{(t+2)^{\zeta+\nu-1}} + \frac{C_\epsilon}{(t+2)^{\nu-1/2}} + \frac{C_\mu}{\tau^{\zeta-1}}, \\ &\leq \frac{\sqrt{\rho}\bar{C}_1 + C^1_\phi}{(t+2)^{2\nu-1}} + \frac{\sqrt{\rho}\bar{C}_2 + C^2_\phi}{(t+2)^{\zeta+\nu-1}} + \frac{\sqrt{\rho}\bar{C}_3 + C_\epsilon + 2(\tau+1)^\nu/(1-\rho)}{(t+2)^{\nu-1/2}} + \frac{\sqrt{\rho}\bar{C}_4 + C_\mu}{\tau^{\zeta-1}}, \\ &\leq \frac{\bar{C}_1}{(t+2)^{2\nu-1}} + \frac{\bar{C}_2}{(t+2)^{\zeta+\nu-1}} + \frac{\bar{C}_3}{(t+2)^{\nu-1/2}} + \frac{\bar{C}_4}{\tau^{\zeta-1}} \end{aligned}$$

with probability $1 - \delta_Q$ (using a union bound type argument) where

$$\bar{C}_1 = \frac{C^1_\phi}{1-\sqrt{\rho}}, \bar{C}_2 = \frac{C^2_\phi}{1-\sqrt{\rho}}, \bar{C}_3 = \frac{C_\epsilon + 2(\tau+1)^\nu/(1-\rho)}{1-\sqrt{\rho}}, \bar{C}_4 = \frac{C_\mu}{1-\sqrt{\rho}},$$

Finally

$$\begin{aligned} \epsilon_Q &= \|Q_T - Q^*_1\|_\infty, \\ &\leq \|Q_T - Q^*_T\|_\infty + \|Q^*_T - Q^*_1\|_\infty, \\ &\leq e^Q_T + \sum_{t=1}^{T-1} \|Q^*_{t+1} - Q^*_t\|_\infty, \\ &\leq \frac{\bar{C}_1}{(t+2)^{2\nu-1}} + \frac{\bar{C}_2}{(t+2)^{\zeta+\nu-1}} + \frac{\bar{C}_3}{(t+2)^{\nu-1/2}} + \frac{\bar{C}_4}{\tau^{\zeta-1}} + L^\mu_Q \sum_{k=1}^{T-1} \|\mu_{t+1} - \mu_t\|_1, \\ &\leq \frac{\bar{C}_1}{(t+2)^{2\nu-1}} + \frac{\bar{C}_2}{(t+2)^{\zeta+\nu-1}} + \frac{\bar{C}_3}{(t+2)^{\nu-1/2}} + \frac{\bar{C}_4}{\tau^{\zeta-1}} + L^\mu_Q \sum_{k=1}^{T-1} c_{\mu,t}, \\ &= \mathcal{O}(T^{1-2\nu}) + \mathcal{O}(T^{1-\zeta-\nu}) + \tilde{\mathcal{O}}(T^{1/2-\nu}) + \mathcal{O}(2^{1-\zeta}) \end{aligned}$$

given that $\zeta \geq 1.1$ with probability at least $1 - \delta_Q$. □

A.4 Proof of Theorem 1

Proof. In this proof we provide finite sample bounds for the convergence of approximation errors in control policy and mean-field, e^k_π and e^k_μ, respectively. We start by characterizing the approximation errors in control policy and mean field e^k_π and e^k_μ on the first timestep in each episode k. Then the evolutions of these approximation errors are studied under two timescale learning rates. First we analyze the approximation error in control policy e^k_π which is evolving at a faster learning rate compared to the approximation error in the mean-field e^k_μ. This error is shown to converge due to the good approximation of the Q-function (Lemma 3), increase of Lipschitz coefficient λ^k at a logarithmic rate and fast learning rate c_π^k. Next the approximation error in mean-field e_μ^k (which is evolving under the slower timescale) is also shown to converge due to the good transition dynamics estimation (Lemma 2), the contraction mapping property (Assumption 1) and the convergence of e_π^k.

First we recall the update rules in Algorithm 1

$$\begin{aligned}
\mu_t^k &= \mathbb{P}_{S(\epsilon^{\text{net}})}\big[(1-c_{\mu,t}^k)\mu_{t-1}^k + c_{\mu,t}^k\hat{\Gamma}_{1,t}^k, \mathbb{1}_{t=1}\big], \text{ where } \hat{\Gamma}_{1,t}^k = (\hat{P}_t^k)^\top \mu_{t-1}^k \\
\pi_t^k &= (1-c_{\pi,t}^k)\pi_{t-1}^k + c_{\pi,t}^k\big((1-\psi)\hat{\Gamma}_{2,t}^k + \psi\mathbb{1}_{|\mathcal{A}|}\big), \text{ where } \hat{\Gamma}_{2,t}^k = \text{softmax}_\lambda(\cdot, Q_t^k)
\end{aligned}$$

where $\hat{\Gamma}_{1,t}^k$ and $\hat{\Gamma}_{2,t}^k$ are the approximate consistency and optimality operators. The RL update can now be written down for the first timestep of episode $k+1$,

$$\begin{aligned}
\mu_1^{k+1} &= \mathbb{P}_{S(\epsilon^{\text{net}})}\big[(1-c_{\mu,1}^{k+1})\mu_0^{k+1} + c_{\mu,1}^{k+1}(\hat{P}_1^{k+1})^\top\mu_0^{k+1}, 1\big], \\
&= \mathbb{P}_{S(\epsilon^{\text{net}})}\big[(1-c_{\mu,1}^{k+1})\mu_T^{k} + c_{\mu,1}^{k+1}(\hat{P}_T^{k})^\top\mu_T^{k}, 1\big], \\
&= \mathbb{P}_{S(\epsilon^{\text{net}})}\big[(1-c_{\mu,1}^{k+1})(\mu_1^{k}+\Delta_\mu^k) + c_{\mu,1}^{k+1}(\hat{P}_T^{k})^\top(\mu_1^{k}+\Delta_\mu^k), 1\big], \\
\pi_1^{k+1} &= (1-c_{\pi,1}^{k+1})\pi_0^{k+1} + c_{\pi,1}^{k+1}\big((1-\psi)\text{softmax}_\lambda(\cdot, Q_1^{k+1}) + \psi\mathbb{1}_{|\mathcal{A}|}\big), \\
&= (1-c_{\pi,1}^{k+1})\pi_T^{k} + c_{\pi,1}^{k+1}\big((1-\psi)\text{softmax}_\lambda(\cdot, Q_1^{k+1}) + \psi\mathbb{1}_{|\mathcal{A}|}\big), \\
&= (1-c_{\pi,1}^{k+1})(\pi_1^{k}+\Delta_\pi^k) + c_{\pi,1}^{k+1}\big((1-\psi)\text{softmax}_\lambda(\cdot, Q_1^{k+1}) + \psi\mathbb{1}_{|\mathcal{A}|}\big),
\end{aligned}$$

where $\Delta_\mu^k := \mu_T^k - \mu_1^k$ and $\Delta_\pi^k := \pi_T^k - \pi_1^k$ are the drifts in mean-field and policy, respectively, in the episode k. Since all the time indices in the above inequalities are 1, we suppress all time indices from here on. Coupled with the fact that $c_{\mu,1}^{k+1} = c_\mu^{k+1}$ and $c_{\pi,1}^{k+1} = c_\pi^{k+1}$, the update rules can be written as

$$\begin{aligned}
\mu^{k+1} &= \mathbb{P}_{S(\epsilon^{\text{net}})}\big[(1-c_{\mu}^{k+1})(\mu^{k}+\Delta_\mu^k) + c_{\mu}^{k+1}(\hat{P}^{k})^\top(\mu^{k}+\Delta_\mu^k), 1\big], \\
\pi^{k+1} &= (1-c_{\pi}^{k+1})(\pi^{k}+\Delta_\pi^k) + c_{\pi}^{k+1}\big((1-\psi)\text{softmax}_\lambda(\cdot, Q^{k+1}) + \psi\mathbb{1}_{|\mathcal{A}|}\big).
\end{aligned} \tag{40}$$

Here we use $\hat{P}^k := \hat{P}_T^k$ and $Q^k := Q_T^k$ for conciseness. The estimation errors for transition matrix and Q-function are denoted as

$$\epsilon_P^k := \|\hat{P}^k - P_{\pi^k,\mu^k}\|_F, \qquad \epsilon_Q^k := \|Q^k - Q_{\mu^k}^*\|_\infty.$$

Now we compute the evolution of the approximation errors. We start with $e_\pi^k := \|\pi^k - \hat{\Gamma}_1^\lambda(\mu^k)\|_{TV}$:

$$\begin{aligned}
e_\pi^{k+1} &= \|\pi^{k+1} - \hat{\Gamma}_1^\lambda(\mu^{k+1})\|_{TV}, \\
&\le \|\pi^{k+1} - \hat{\Gamma}_1^\lambda(\mu^{k})\|_{TV} + \|\hat{\Gamma}_1^\lambda(\mu^{k}) - \hat{\Gamma}_1^\lambda(\mu^{k+1})\|_{TV}, \\
&\le \|(1-c_\pi^{k+1})(\pi^k + \Delta_\pi^k) + c_\pi^{k+1}\text{softmax}_\lambda(\cdot, Q^k) - \text{softmax}_\lambda(\cdot, Q_{\mu^k}^*)\|_{TV} + d_1\|\mu^{k+1}-\mu^k\|_1 + 2c_\pi^{k+1}\psi, \\
&\le (1-c_\pi^{k+1})\|\pi^k - \hat{\Gamma}_1^\lambda(\mu^k)\|_{TV} + (1-c_\pi^{k+1})\|\Delta_\pi^k\|_{TV} \\
&\qquad\qquad + c_\pi^{k+1}\|\text{softmax}_\lambda(\cdot, Q^k) - \text{softmax}_\lambda(\cdot, Q_{\mu^k}^*)\|_{TV} + d_1\|\mu^{k+1}-\mu^k\|_1 + c_\pi^{k+1}\epsilon/2, \\
&\le (1-c_\pi^{k+1})e_\pi^k + \|\Delta_\pi^k\|_{TV} + c_\pi^{k+1}\|\text{softmax}_\lambda(\cdot, Q^k) - \text{softmax}_\lambda(\cdot, Q_{\mu^k}^*)\|_{TV} \\
&\qquad\qquad + d_1\|\mu^{k+1}-\mu^k\|_1 + c_\pi^{k+1}\epsilon/2.
\end{aligned} \tag{41}$$

where the third inequality is due to $\psi \le \epsilon/4$. To simplify the above expression we prove the Lipschitz property of the softmax$_\lambda(\cdot, Q)$ operator.

Lemma 11. *The softmax$_\lambda(\cdot, Q)$ satisfies the Lipschitz property for $\lambda > 0$ and $Q : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^+$,*

$$\|\textit{softmax}_\lambda(\cdot, Q) - \textit{softmax}_\lambda(\cdot, Q')\|_{TV} \le \lambda S\sqrt{A}\|Q - Q'\|_\infty.$$

Proof. The Lipschitzness of softmax can be obtained using Proposition 4 in Gao and Pavel (2017). Let us denote the policy under softmax$_\lambda(\cdot, Q)$ as π_Q^λ such that $\pi_Q^\lambda(a|s) = \frac{\exp(\lambda Q(s,a))}{\sum_{a'\in\mathcal{A}}\exp(\lambda Q(s,a'))}$. Now

$$\begin{aligned}
\|\text{softmax}_\lambda(\cdot, Q) - \text{softmax}_\lambda(\cdot, Q')\|_{TV} &= \|\pi_Q^\lambda - \pi_{Q'}^\lambda\|_{TV}, \\
&= \max_{a\in\mathcal{A}}\sum_{s\in\mathcal{S}}\big|\pi_Q^\lambda(a|s) - \pi_{Q'}^\lambda(a|s)\big|
\end{aligned} \tag{42}$$

From Proposition 4 in Gao and Pavel (2017), we know that for any $s \in \mathcal{S}$

$$\begin{aligned}
\|\pi^\lambda_Q(\cdot|s) - \pi^\lambda_{Q'}(\cdot|s)\|_2 = \sqrt{\sum_{a\in\mathcal{A}} \left(\pi^\lambda_Q(a|s) - \pi^\lambda_{Q'}(a|s)\right)^2} &\leq \lambda\|Q(s,\cdot) - Q(s,\cdot)\|_2, \\
&= \lambda\sqrt{\sum_{a\in\mathcal{A}} \left(Q(s,a) - Q'(s,a)\right)^2}, \\
&\leq \lambda\sqrt{A}\|Q(s,\cdot) - Q'(s,\cdot)\|_\infty, \\
&\leq \lambda\sqrt{A}\|Q - Q'\|_\infty. \qquad (43)
\end{aligned}$$

The second inequality is due to the equivalence between 2 and ∞ vector norms. This equivalence also gives us

$$\|\pi^\lambda_Q(\cdot|s) - \pi^\lambda_{Q'}(\cdot|s)\|_2 = \sqrt{\sum_{a\in\mathcal{A}} \left(\pi^\lambda_Q(a|s) - \pi^\lambda_{Q'}(a|s)\right)^2} \geq \max_{a\in\mathcal{A}} \left|\pi^\lambda_Q(a|s) - \pi^\lambda_{Q'}(a|s)\right| \qquad (44)$$

Recalling (42),

$$\begin{aligned}
\|\text{softmax}_\lambda(\cdot, Q) - \text{softmax}_\lambda(\cdot, Q')\|_{TV} &= \max_{a\in\mathcal{A}} \sum_{s\in\mathcal{S}} \left|\pi^\lambda_Q(a|s) - \pi^\lambda_{Q'}(a|s)\right| \\
&\leq \sum_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} \left|\pi^\lambda_Q(a|s) - \pi^\lambda_{Q'}(a|s)\right|, \\
&\leq \lambda S\sqrt{A}\|Q - Q'\|_\infty
\end{aligned}$$

where the last inequality is obtained using (43) and (44). □

Now we can further simplify (41) as:

$$\begin{aligned}
e^{k+1}_\pi &\leq (1 - c^{k+1}_\pi)e^k_\pi + \|\Delta^k_\pi\|_{TV} + c^{k+1}_\pi \lambda S\sqrt{A}\|Q^k - Q^*_{\mu^k}\|_\infty + d_1\|\mu^{k+1} - \mu^k\|_1 + c^{k+1}_\pi \epsilon/2, \\
&\leq (1 - c^{k+1}_\pi)e^k_\pi + \|\Delta^k_\pi\|_{TV} + c^{k+1}_\pi \lambda S\sqrt{A}\epsilon^k_Q + c^{k+1}_\mu d_1(2 + \|\Delta^k_\mu\|_1) + \|\Delta^k_\mu\|_1 + c^{k+1}_\pi \epsilon/2. \qquad (45)
\end{aligned}$$

The first inequality is due to Lemma 11 and the second inequality is due to (40) and the fact that $\|\mu\|_1 \leq 1$ for any $\mu \in \mathcal{P}(\mathcal{S})$. The norms of the drift terms are bounded by

$$\|\Delta^k_\pi\|_{TV} \leq c^k_\pi \sum_{t=2}^{T-1} t^{-\zeta} \leq c^k_\pi \frac{2^{1-\zeta}}{\zeta - 1}, \qquad \|\Delta^k_\mu\|_1 \leq c^k_\mu \sum_{t=2}^{T-1} t^{-\zeta} \leq c^k_\mu \frac{2^{1-\zeta}}{\zeta - 1} \qquad (46)$$

Rearranging the inequality (45),

$$\begin{aligned}
e^k_\pi &\leq \frac{1}{c^{k+1}_\pi}(e^k_\pi - e^{k+1}_\pi) + \frac{2^{1-\zeta}}{\zeta - 1} + \lambda S\sqrt{A}\epsilon^k_Q + \frac{c^{k+1}_\mu}{c^{k+1}_\pi} d_1 \left(2 + \frac{2^{1-\zeta}}{\zeta - 1}\right) + \frac{c^k_\mu}{c^{k+1}_\pi}\frac{2^{1-\zeta}}{\zeta - 1} + \frac{\epsilon}{2}, \\
&\leq \frac{e^k_\pi - e^{k+1}_\pi}{c^{k+1}_\pi} + 10 \cdot 2^{1-\zeta} + \lambda S\sqrt{A}\epsilon^k_Q + 12\frac{c^{k+1}_\mu}{c^{k+1}_\pi} d_1 + 10\frac{c^k_\mu}{c^{k+1}_\pi} + \frac{\epsilon}{2},
\end{aligned}$$

for $\zeta \geq 1.1$. Now taking the average over $k = 1, \ldots, K-1$, we get

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K-1} e_\pi^k \\
&\leq \frac{1}{K}\sum_{k=1}^{K-1}\Big(\frac{e_\pi^k - e_\pi^{k+1}}{c_\pi^{k+1}} + \lambda S\sqrt{A}\epsilon_Q^k + 12\frac{c_\mu^{k+1}}{c_\pi^{k+1}}d_1 + 10\frac{c_\mu^k}{c_\pi^{k+1}}\Big) + 10\cdot 2^{1-\zeta} + \frac{\epsilon}{2}, \\
&\leq \frac{1}{K}\sum_{k=2}^{K-1}\Big(\frac{1}{c_\pi^{k+1}} - \frac{1}{c_\pi^k}\Big)e_\pi^{k+1} + \frac{1}{K}\sum_{k=1}^{K-1}\Big(\lambda S\sqrt{A}\epsilon_Q^k \\
&\qquad\qquad + (12d_1 + 20)\frac{c_\mu^{k+1}}{c_\pi^{k+1}}\Big) + \frac{1}{c_\pi^2 K}e_\pi^1 - \frac{1}{c_\pi^{K+1}K}e_\pi^K + 10\cdot 2^{1-\zeta} + \frac{\epsilon}{2}, \\
&\leq \frac{2}{K}\sum_{k=2}^{K-1}\Big(\frac{1}{c_\pi^{k+1}} - \frac{1}{c_\pi^k}\Big) + \frac{1}{K}\sum_{k=1}^{K-1}\Big(\lambda S\sqrt{A}\epsilon_Q^k \\
&\qquad\qquad + (24d_1 + 40)\frac{c_\mu}{c_\pi}k^{\theta-\gamma}\Big) + \frac{2}{c_\pi^2 K} + 10\cdot 2^{1-\zeta} + \frac{\epsilon}{2}, \\
&\leq \frac{2}{Kc_\pi^K} + S\sqrt{A}\lambda\epsilon_Q^k + \frac{(24d_1+40)c_\mu}{(1+\theta-\gamma)c_\pi}K^{\theta-\gamma} + \frac{2}{c_\pi^2 K} + 10\cdot 2^{1-\zeta} + \frac{\epsilon}{2},
\end{aligned} \tag{47}
$$

where the second to last inequality is due to the fact that $e_\pi^k \leq 2$. Since $\epsilon_Q^k \leq \epsilon_Q/\lambda$, where $\epsilon_Q > 0$, then

$$
\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K-1} e_\pi^k &\leq \frac{2}{Kc_\pi^K} + S\sqrt{A}\epsilon_Q + \frac{(24d_1+40)\bar{\mu}c_\mu}{(1+\theta-\gamma)c_\pi}K^{\theta-\gamma} + \frac{2}{c_\pi^2 K} + 10\cdot 2^{1-\zeta} + \frac{\epsilon}{2}, \\
&\leq \mathcal{O}(K^{\theta-1}) + \mathcal{O}(\epsilon_Q) + \mathcal{O}(K^{\theta-\gamma}) + \mathcal{O}(K^{-1}) + \mathcal{O}(2^{1-\zeta}) + \mathcal{O}(\epsilon)
\end{aligned} \tag{48}
$$

where K is the total number of episodes.

Now we analyze the mean-field approximation error evolution $e_\mu^k := \|\mu^k - \mu^*\|_1$. Let us define $\hat{\mu}^{k+1} := (1 - c_\mu^{k+1})(\mu^k + \Delta_\mu^k) + c_\mu^{k+1}(\hat{P}^k)^\top(\mu^k + \Delta_\mu^k)$. Then,

$$
\begin{aligned}
e_\mu^{k+1} &= \|\mu^{k+1} - \mu^*\|_1 = \|\mathbb{P}_{S(\epsilon^{\text{net}})}[\hat{\mu}^{k+1}, 1] - \Gamma_2(\Gamma_1^\lambda(\mu^*), \mu^*)\|_1 \\
&\leq \|\mathbb{P}_{S(\epsilon^{\text{net}})}[\hat{\mu}^{k+1}, 1] - \hat{\mu}^{k+1}\|_1 + \|\hat{\mu}^{k+1} - \Gamma_2(\Gamma_1^\lambda(\mu^*), \mu^*)\|_1, \\
&\leq (1 - c_\mu^{k+1})\|\mu^k - \mu^*\|_1 + (1 - c_\mu^{k+1})\|\Delta_\mu^k\|_1 + c_\mu^{k+1}\big[\|(\hat{P}^k)^\top(\mu^k + \Delta_\mu^k) - \Gamma_2(\pi^k, \mu^k)\|_1 \\
&\qquad\qquad + \|\Gamma_2(\pi^k, \mu^k) - \Gamma_2(\Gamma_1^\lambda(\mu^*), \mu^*)\|_1\big] + \epsilon^{\text{net}}, \\
&\leq (1 - c_\mu^{k+1})e_\mu^k + \|\Delta_\mu^k\|_1 + c_\mu^{k+1}\big[\|(\hat{P}^k)^\top(\mu^k + \Delta_\mu^k) - \Gamma_2(\pi^k, \mu^k)\|_1 \\
&\qquad + \|\Gamma_2(\pi^k, \mu^k) - \Gamma_2(\Gamma_1^\lambda(\mu^k), \mu^k)\|_1 + \|\Gamma_2(\Gamma_1^\lambda(\mu^k), \mu^k) - \Gamma_2(\Gamma_1^\lambda(\mu^*), \mu^*)\|_1\big] + \epsilon^{\text{net}}, \\
&\leq (1 - c_\mu^{k+1})e_\mu^k + \|\Delta_\mu^k\|_1 + c_\mu^{k+1}\big[\|(\hat{P}^k)^\top(\mu^k + \Delta_\mu^k) - P_{\pi^k,\mu^k}^\top\mu^k\|_1 + d_2\|\pi^k - \Gamma_1^\lambda(\mu^k)\|_1 \\
&\qquad\qquad + (d_1d_2 + d_3)\|\mu^k - \mu^*\|_1\big] + \epsilon^{\text{net}},
\end{aligned}
$$

$$
\begin{aligned}
e_\mu^{k+1} &\leq (1 - c_\mu^{k+1}\bar{d})e_\mu^k + (1 + c_\mu^k)\|\Delta_\mu^k\|_1 + c_\mu^{k+1}\|(\hat{P}^k)^\top - P_{\pi^k,\mu^k}^\top\|_1 + c_\mu^{k+1}d_2e_\pi^k + \epsilon^{\text{net}}, \\
&\leq (1 - c_\mu^{k+1}\bar{d})e_\mu^k + (1 + c_\mu^k)\|\Delta_\mu^k\|_1 + c_\mu^{k+1}\sqrt{S}\|\hat{P}^k - P_{\pi^k,\mu^k}\|_F + c_\mu^{k+1}d_2e_\pi^k + \epsilon^{\text{net}}, \\
&\leq (1 - c_\mu^{k+1}\bar{d})e_\mu^k + 11c_\mu^k 2^{1-\zeta} + c_\mu^{k+1}\sqrt{S}\epsilon_P^k + c_\mu^{k+1}d_2e_\pi^k + \epsilon^{\text{net}}
\end{aligned}
$$

where the second to last inequality is due to the equivalence between induced 1 norm and the Frobenius norm. Rearranging the above inequality,

$$
\begin{aligned}
e_\mu^k &\leq \frac{1}{c_\mu^{k+1}\bar{d}}(e_\mu^k - e_\mu^{k+1}) + 11\frac{c_\mu^k}{c_\mu^{k+1}\bar{d}}2^{1-\zeta} + \frac{\sqrt{S}\epsilon_P^k}{\bar{d}} + \frac{d_2e_\pi^k}{\bar{d}} + \frac{\epsilon^{\text{net}}}{c_\mu^{k+1}\bar{d}}, \\
&\leq \frac{1}{c_\mu^{k+1}\bar{d}}(e_\mu^k - e_\mu^{k+1}) + 22\frac{2^{1-\zeta}}{\bar{d}} + \frac{\sqrt{S}\epsilon_P^k}{\bar{d}} + \frac{d_2e_\pi^k}{\bar{d}} + \frac{\epsilon^{\text{net}}}{c_\mu^{k+1}\bar{d}}
\end{aligned}
$$

Taking average over $k = 1, \dots, K-1$, we get

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K-1} e_\mu^k \le \frac{1}{K}\sum_{k=1}^{K-1}\left[\frac{1}{c_\mu^{k+1}\bar{d}}(e_\mu^k - e_\mu^{k+1}) + \frac{\sqrt{S}\epsilon_P^k}{\bar{d}} + \frac{d_2 e_\pi^k}{\bar{d}} + \frac{\epsilon^{\text{net}}}{c_\mu^{k+1}\bar{d}}\right] + 11\frac{2^{1-\zeta}}{\bar{d}},\\
&\le \frac{\bar{e}_\mu}{c_\mu^K \bar{d} K} + 11\frac{2^{1-\zeta}}{\bar{d}} + \frac{\sqrt{S}\epsilon_P}{\bar{d}} + \frac{1}{K}\sum_{k=1}^{K-1}\left[\frac{d_2 e_\pi^k}{\bar{d}} + \frac{\epsilon^{\text{net}}}{c_\mu^{k+1}\bar{d}}\right],\\
&\le \frac{\bar{e}_\mu}{c_\mu^K \bar{d} K} + 11\frac{2^{1-\zeta}}{\bar{d}} + \frac{\sqrt{S}\epsilon_P}{\bar{d}} + \frac{1}{K}\sum_{k=2}^{K}\frac{\epsilon^{\text{net}} k^\gamma}{c_\mu \bar{d}} + \frac{1}{K}\sum_{k=1}^{K-1}\frac{d_2 e_\pi^k}{\bar{d}},\\
&\le \mathcal{O}(K^{\gamma-1}) + \mathcal{O}(2^{1-\zeta}) + \mathcal{O}(\epsilon_P) + \mathcal{O}(K^{\theta-1}) + \mathcal{O}(\epsilon_Q) + \mathcal{O}(K^{\theta-\gamma}) + \mathcal{O}(K^{-1}) + \mathcal{O}(\epsilon)
\end{aligned}
$$

where the second inequality is obtained using steps similar to (47) and the fact that $\epsilon_P^k \le \epsilon_P$. The last inequality is obtained using (48) and the fact that $\epsilon^{\text{net}} \le c_\mu \bar{d}\epsilon / K^\gamma$. The proof is thus concluded. □

A.5 Proof of Corollary 1

Proof. This is a corollary to Theorem 1:

$$
\begin{aligned}
&\left\|\frac{1}{K}\sum_{k=1}^{K-1}\pi^k - \pi^*\right\| + \left\|\frac{1}{K}\sum_{k=1}^{K-1}\mu^k - \mu^*\right\| \le \frac{1}{K}\sum_{k=1}^{K-1}\|\pi^k - \pi^*\| + \frac{1}{K}\sum_{k=1}^{K-1}\|\mu^k - \mu^*\|,\\
&\le \frac{1}{K}\sum_{k=1}^{K-1}\|\pi^k - \Gamma_1^\lambda(\mu^k)\| + \frac{1}{K}\sum_{k=1}^{K-1}\|\Gamma_1^\lambda(\mu^k) - \pi^*\| + \frac{1}{K}\sum_{k=1}^{K-1}\|\mu^k - \mu^*\|,\\
&\le \frac{1}{K}\sum_{k=1}^{K-1}\|\pi^k - \Gamma_1^\lambda(\mu^k)\| + \frac{d_1+1}{K}\sum_{k=1}^{K-1}\|\mu^k - \mu^*\|,\\
&= \mathcal{O}(K^{\gamma-1}) + \mathcal{O}(2^{1-\zeta}) + \mathcal{O}(\epsilon_P) + \mathcal{O}(K^{\theta-1}) + \mathcal{O}(\epsilon_Q) + \mathcal{O}(\epsilon) + \mathcal{O}(K^{\theta-\gamma}) + \mathcal{O}(K^{-1}).
\end{aligned}
$$

where the last inequality follows from Theorem 1. □