ORACLE-FREE REINFORCEMENT LEARNING IN MEAN-FIELD GAMES ALONG A SINGLE SAMPLE PATH

Muhammad Aneeq uz Zaman
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Urbana IL 61801-2307
mazaman2@illinois.edu

Alec Koppel
J.P. Morgan AI Research
alec.koppel@jpmchase.com

Sujay Bhatt
sujaybhatt.hr@gmail.com

Tamer Başar
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Urbana IL 61801-2307
basar1@illinois.edu

ABSTRACT

We consider online reinforcement learning in Mean-Field Games. In contrast to the existing works, we alleviate the need for a mean-field oracle by developing an algorithm that estimates the mean-field and the optimal policy using a single sample path of the generic agent. We call this *Sandbox Learning*, as it can be used as a warm-start for any agent operating in a multi-agent non-cooperative setting. We adopt a two timescale approach in which an online fixed-point recursion for the mean-field operates on a slower timescale and in tandem with a control policy update on a faster timescale for the generic agent. Under a sufficient exploration condition, 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 $\mathcal{O}(\epsilon^{-4})$. Finally, we empirically demonstrate effectiveness of the sandbox learning algorithm in a congestion game.

1 Introduction

The last decade has seen tremendous progress in single-agent Reinforcement Learning (RL), with methods like DQN, TRPO, PPO and PGQ [1, 2, 3, 4] being developed and successfully applied to applications such as healthcare, transportation, and robotics, to name a few [5]. In the single-agent sequential decision making setting, the agent accumulates rewards by interacting with the environment. Single-agent RL aims to find the policy which maximizes the agent's accumulated total reward without having knowledge of the state dynamics and the reward function of the agent. Since many real-world scenarios like cyber-physical systems [6], financial trading [7], and power markets [8] involve multiple agents, there has been a significant increase in interest in theory and application of Multi-Agent Reinforcement Learning (MARL) [9, 10]. MARL considers a non-cooperative multi-agent game where each agent aims to maximize its private accumulated total reward. Each agent's reward function as well as its state dynamics may be influenced by the other agents. Challenges arise when translating single-agent RL methods to instances with multiple non-cooperative agents due to the resultant non-stationarity of the environment caused by the presence and the actions of the other agents. Moreover, the complexity of the game is further compounded by the fact that an agent best responding to other agents might cause the other agents to change their policies in best responses to the first agent, which is referred to as the "curse of many agents" in game theory [11].

The Mean-Field Game (MFG) framework proposed concurrently by Huang et al. [12, 13] and Lasry & Lions [14, 15], overcomes this difficulty by considering the limiting case where the number of agents $N \to \infty$. In this infinite population setting the effect of individual deviation becomes negligible, causing any strategic interaction among the agents to disappear. Due to this simplification, it becomes sufficient to consider, without any loss of generality the

interaction between a generic agent and the aggregate behavior of other agents (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 [16], 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 steady-state (stationary) MFE [17, 18] 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. Hence the question arises

Can the generic agent provably learn its stationary MFE without access to a mean-field oracle, but only having access to its own state, action and reward sequence?

We answer this question in the affirmative by proposing an RL algorithm and providing 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 equilibrium policies for a multi-agent non-cooperative environment, without interacting with other agents or oracles. 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

The technical novelty of this paper is the finite sample analysis for RL for MFGs in an *oracle-free setting*. We note that previous works that consider similar settings either require stronger conditions such as access to mean-field oracle [17, 18, 19, 20], or establish only asymptotic convergence [21]. The main results of the paper can be summarized as follows.

1. We introduce an episodic two time-scale learning rate for RL in MFG in the oracle-free setting. Simultaneously updating the mean-field and the policy of the agent using its single sample path brings in a technical problem, as it induces a time-varying Markov Chain (MC). We craft the episodic learning rates for the sole purpose of making the MC *slowly* time-varying inside the episode.

2. We provide finite sample analysis of Q-learning and dynamics matrix estimation under the (slowly) time-varying MC setting, using sufficient exploration conditions from the literature [22] (in Lemmas 1 & 2). The slowly time-varying MC setting is shown to introduce a small *drift* in the approximation error, which can be reduced by slowing inter-episodic learning. Lemmas 1 and 2 might be of independent interest to researchers working in RL in time-varying setting.

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, we obtain finite sample convergence bounds of this two time-scale algorithm to an ϵ-neighborhood of stationary MFE.

4. We illustrate the effectiveness of our algorithm using numerical analysis on a simple congestion game.

1.2 Relevant Literature

The work most closely related to this paper is [21] 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 [21] relies on re-initializations while our algorithm operates on a single sample path, (b) the algorithm proposed in [21] 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 pivotal in obtaining the finite sample convergence bounds for the Sandbox learning algorithm as shown in following sections. Below we have provided a table juxtaposing our work with the contributions of several other works in RL for MFGs. A complete literature review is provided in Section 2.

	Needs Oracle/Oracle-less	Single sample path	Finite sample bounds
Guo et al. [17]	Needs Oracle	No	Yes
Elie et al. [23]	Needs Oracle	No	No
Fu et al. [20]	Needs Oracle	No	Yes
Cui & Koeppl [24]	Needs Oracle	No	No
Anahtarci et al. [25]	Needs Oracle	No	Yes
Xie et al. [18]	Needs Oracle	No	Yes
Angiuli et al. [21]	Oracle-less	No	No
This work	Oracle-less	Yes	Yes

The Sandbox learning algorithm involves concurrent update of the mean-field and the control policy of the generic agent. This update relies on estimating the transition dynamics and Q-function of the generic agent. But the Markov Chain (MC) (and by extension the transition probability and Q-function) of the agent depend on the mean-field and control policy which are concurrently being updated. Obtaining convergence bounds in this time-varying MC setting proves to be a hard problem. To ameliorate the effect of time-varying MC, we craft an episodic learning rate such that the learning rate is summable (or fast-decaying) *inside* the episode but non-summable *outside* the episode. This episodic learning rate as depicted in Figure 1 is key to obtaining finite sample convergence bounds. The MC induced by the episodic learning is called *slowly time-varying* Markov chain [26]. Within this slowly time-varying setting, we can show convergence of the transition probability estimation and Q-learning update. These convergence results for time-varying MDPs are of interest independently of the Sandbox learning algorithm. Using these transition probability and Q-function estimators, we use a two time-scale update on the mean-field and the control policy, such that the control policy is updated at a faster rate than the mean-field. Our results can be extended to the converse setting as well, which obtains the solution to the cooperative Mean-Field Control problem [21].

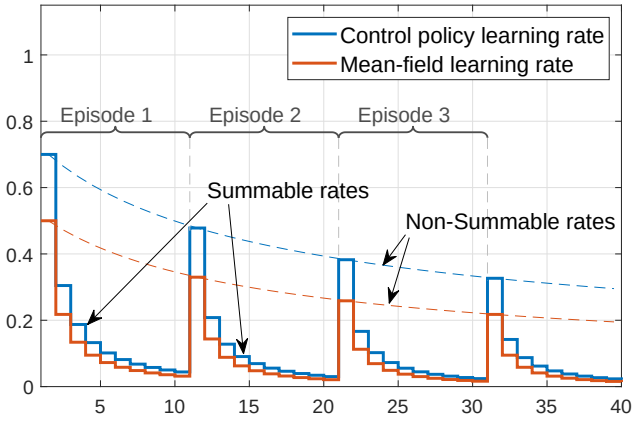


Figure 1: Episodic Two time-scale learning rate

1.3 Organization

Section 2 provides a deep literature dive into RL for MFGs. Section 3 introduces the non-cooperative N-agent game and its mean-field limit. Section 4 introduces our approach, leading to the Sandbox learning algorithm. Section 5 proves convergence of transition dynamics and Q-learning estimation, and guarantees finite sample bounds for the Sandbox learning algorithm. Finally, Section 6 performs empirical analysis for the algorithm on a congestion game.

2 Related Literature

Mean-Field Games (MFGs) originated concurrently in the works of Huang et al. [12, 13] (termed as Nash Certainty Equivalence) and Lasry & Lions [14, 15] (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 [27], scarce interactions [28, 29], risk-sensitive criteria [30, 31, 32] and cooperative equilibria [33, 34]. MFGs have also been applied to a variety of real-world applications such as decentralized charging of EVs [35], economics [36] and congestion dynamics [37], 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 most recently [16, 27].

RL for MFGs was first dealt with in [17] for the finite and in [23] for infinite state and action spaces. The work of [17] 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 (Assumption 3) bound over the ϵ-net. A fictitous play algorithm was proposed [23], 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 [20, 38, 39]. These works have provided finite sample bounds for the LQ-MFG in the stationary [20] and the non-stationary [38, 39] settings, by building on policy gradient [40] and zero-order stochastic optimization methods [41] for the Linear Quadratic Regulator problem. The idea of entropy-regularized MFGs was introduced in [18, 24] along with existence and uniqueness results and RL algorithms to compute the entropy-regularized MFE. The work of [19] 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 [42, 43], where [42] uses Deep RL techniques to learn a flocking model observed in nature and [43] proposes a Neural Network based policy update mechanism. The paper [18] 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 [21] 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 [21] 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 [19]) and fixing the softmax Lipschitz parameter λ in the control policy update (Algorithm 1 line 6).

3 Formulation and 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

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

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 $i = 1, \ldots, N$ is to maximize its total reward discounted by a factor $0 < \rho < 1$, defined as

$$V^i(\pi^i, \pi^{-i}) = \mathbb{E}_{a_t^i \sim \pi_t^i(s_t^i, e_t^N), s_{t+1}^i \sim P(\cdot|s_t^i, a_t^i, e_t^N)} \Big[\sum_{t=1}^{\infty} \rho^t R(s_t^i, a_t^i, e_t^N) \mid s_t^i \sim p_1 \Big], \tag{2}$$

where $\pi^i := (\pi_1^i, \pi_2^i, \ldots)$ 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 ([44]). *A set of policies $\pi^* = \{\pi^{1*}, \ldots, \pi^{N*}\}$ is termed an ϵ-Nash equilibrium if $\forall i \in [N]$, $V^i(\pi^{i*}, \pi^{-i*}) + \epsilon > V^i(\pi^i, \pi^{-i*}), \forall \pi^i \in \Pi^i$, where $\Pi^i := \{\Pi_1^i, \Pi_2^i, \ldots\}$.*

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 [44], 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 [16, 27], 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 ergodicity assumptions [16, 45], which assign positive probability to other agents' traversal of any subset of the state space. 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 $\mu \in \mathcal{P}(\mathcal{S})$:

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

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 [46]. 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 $0 < \rho < 1$, which is

$$V_{\pi,\mu} := \mathbb{E}_{a\sim\pi(s,\mu), s'\sim P(\cdot|s,a,\mu)}\Big[\sum_{t=1}^{\infty} \rho^t R(s_t, a_t, \mu) \mid s_1 \sim p_1\Big]. \tag{4}$$

Next we define the Mean-Field Equilibrium (MFE) by introducing two operators. First define the *optimality* operator $\Gamma_1(\mu) := \operatorname{argmax}_\pi V_{\pi,\mu}$ 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

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

This is also referred to as the Fokker-Planck-Kolmogorov equation in the literature [47], and versions of it appear in the literature on probability flow equations in MDPs [46]. 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 ([16])**.** *The pair (π^*, μ^*) is an MFE of the MFG if $\pi^* = \Gamma_1(\mu^*)$ and $\mu^* = \Gamma_2(\pi^*, \mu^*)$.*

Intuitively this two-part coupled definition can be interpreted as (1) π^* is the optimal policy for the MDP induced by mean-field μ^*, and (2) mean-field μ^* is consistent with the control policy π^*. Next we introduce some background material, needed for our main results. We start with a contraction mapping assumption in MFGs [17, 18].

Assumption 1. *There exist Lipschitz constants d_1, d_2 and d_3 for operators Γ_1 and Γ_2 such that*

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

and $d := d_1d_2 + d_3 < 1$ for policies $\pi, \pi' \in \Pi$ and mean-fields $\mu, \mu' \in \mathcal{P}(\mathcal{S})$.

The Lipschitz conditions in Assumption 1 are widely used in RL for standard MFGs [17, 18, 20]. These conditions follow due to continuity of the optimality and consistency operators, Γ_1 and Γ_2, respectively, and the compactness of their domains. Lemma 5 in [17] provides values for these constants. The $\|\cdot\|_{TV}$ norm used in assumption 1 is the Total variation bound [24] 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 MFE of the MFG has been proven in literature [27, 17, 18] using the standard contraction mapping theorem. Furthermore, this unique MFE is also known to be ϵ-Nash as shown in Theorem 2.3 of [16]. In the next section, we propose an RL algorithm to find the MFE without access to a mean-field oracle, but approximates it by observing the sample path of a generic agent itself.

4 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 – see [17, 20, 18, 24]. In this section, we propose a Sandbox RL algorithm to compute the MFE. Our methodology operates by updating the mean-field and the control policy concurrently using approximations of the optimality and consistency operators, Γ_1 and Γ_2, respectively, defined prior to Definition 2. The approximation to Γ_1 is defined by $\text{softmax}_\lambda(\cdot)$ 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. Therefore, given a reasonable estimate for the consistency operator, the control policy is updated on a faster time-scale. In the following subsection we describe how we can estimate the two operators.

4.1 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

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

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

$$\mu' = P_{\pi,\mu}^\top \mu \tag{6}$$

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 (6). 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:

$$\hat{P}(i, j) = \frac{N(i, j) + 1/S}{N(i) + 1}, \quad N(i, j) = \big|\{l \in [T] : s_l = i, s_{l+1} = j\}\big|, \quad N(i) = \sum_{j\in\mathcal{S}} N_t^k(i, j) \tag{7}$$

where 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. (11)] in (9), 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. Operator Γ_1 returns the policy which is optimal for the MDP induced by a given mean-field μ. From standard results in the literature [48] we know that $\Gamma_1 = \operatorname{argmax} Q_\mu^*$ where $Q_\mu^*(s, a) := \operatorname{argmax}_\pi \mathbb{E}[\sum_{t=1}^\infty R(s_t, a_t, \mu)|s_1 = s, a_1 = a]$ is the optimal Q-function for the MDP induced by the mean-field μ and is the fixed point of the Bellman equation

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

Hence the algorithm uses Q-learning update to approximate the optimal Q-function. The asynchronous Q-learning update [48, 49] can be written as follows,

$$Q_{t+1}(s_t, a_t) = (1 - \beta_t)Q_t(s_t) + \beta_t\big(r_t + \rho\max_{a\in\mathcal{A}} Q_t(s_{t+1}, a)\big), \text{ where } \beta_t := \frac{c_\beta}{(t + 1)^\nu} \tag{8}$$

where $0.5 < \nu \leq 1$. But since the argmax operator does not follow the Lipschitz property in finite action spaces [50], we cannot apply argmax to the estimate of optimal Q-function and expect to have a good estimate of the optimality operator. Instead we resort to $\operatorname{softmax}_\lambda(\cdot)$ operator which does follow Lipschitz property with constant λ. The $\operatorname{softmax}_\lambda(\cdot) : \mathcal{S} \times \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathcal{P}(\mathcal{A})$ is defined as

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

Let us denote the approximate optimality operator as $\hat{\Gamma}_1 := \operatorname{softmax}_\lambda(\cdot, \hat{Q})$ where $\hat{Q}$ is the estimate of the optimal Q-function. The estimation error in $\hat{\Gamma}_1$ is due to (a) the estimation error in the Q-learning, and (b) the difference between the argmax and the $\operatorname{softmax}_\lambda(\cdot)$ operators. Both components of the error have a dependence on the Lipschitz parameter λ. The first component of the estimation error increases with increasing λ. The second component has an inverse relationship with λ and also depends on the action gap in the optimal Q-function. With this technical machinery introduced both for the approximate consistency operator and the Bellman operator and its softmax augmentation, we are now ready to introduce the Sandbox learning algorithm. This is the focus of the following subsection.

4.2 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 7), and hence the algorithm involves a single sample path (of the generic agent) without re-initialization. Before stating the update rules for the mean-field and policy, we introduce the set $S(\epsilon^{\text{net}})$ which is a set of mean-field distributions. This set (also termed ϵ-net [17] over $\mathcal{P}(\mathcal{S})$) defined as $S(\epsilon^{\text{net}}) = \{\mu^1, \ldots, \mu^{N_{\text{net}}}\} \subset \mathcal{P}(\mathcal{S})$ is a finite set of simplexes over $\mathcal{S}$ with the property that for any $\mu \in \mathcal{P}(\mathcal{S}), \exists \mu' \in S(\epsilon^{\text{net}})$ such that $\|\mu - \mu'\|_1 \leq \epsilon^{\text{net}}$. The existence of the set is guaranteed due to the compactness of $\mathcal{P}(\mathcal{S})$. The mean-field μ_t^k and control policy π_t^k are updated at each timestep (lines 5-6) using

$$\mu_t^k = \mathbb{P}_{S(\epsilon^{\text{net}})}\big[(1 - c_{\mu,t}^k)\mu_{t-1}^k + c_{\mu,t}^k\big((\hat{P}_t^k)^\top \mu_{t-1}^k\big), 1 - \psi_t\big], \tag{9}$$

$$\pi_t^k = (1 - c_{\pi,t}^k)\pi_{t-1}^k + c_{\pi,t}^k\big((1 - \psi_t)\text{softmax}_{\lambda^k}(\cdot, Q_t^k) + \psi_t \mathbb{1}_{|\mathcal{A}|}\big), \tag{10}$$

The update of mean-field involves the operation $\mathbb{P}_{S(\epsilon^{\text{net}})}[\mu, x]$, which projects μ onto the ϵ-net $S(\epsilon^{\text{net}})$ only if $x = 1$. This projection step is performed on the first time-step of each episode k. The set $S(\epsilon^{\text{net}})$ is finite due to the compactness of $\mathcal{P}(\mathcal{S})$. In the analysis (Section 5) we show that $\epsilon^{\text{net}} = \mathcal{O}(\epsilon^2)$, where $\epsilon > 0$ is the approximation error in the 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 $\text{softmax}_{\lambda^k}(\cdot, Q_t^k)$. The Lipschitz parameter λ^k is increased at a logarithmic rate so that $\text{softmax}_{\lambda^k}(\cdot)$ approximates argmax in the limit $k \to \infty$. The control policy updates also involve a uniform exploration noise $\psi_t \mathbb{1}_{|\mathcal{A}|}$ to avoid degenerate (non-totally mixed) policies [44] which might invalidate the sufficient exploration condition. By scheduling $\psi_t = 0$ for time $t = 1$ and $\psi_t = 0.2$ for $t > 1$, the exploration noise is added to all timesteps after the first one (in each episode) to decrease the effect of exploration noise on convergence analysis. The learning rates for the update, $c_{\mu,t}^k$ and $c_{\pi,t}^k$, are episodic two time-scale

$$c_{\mu,t}^k = \frac{c_\mu}{k^\gamma}\frac{1}{t^\zeta}, \;\; c_{\pi,t}^k = \frac{c_\pi}{k^\theta}\frac{1}{t^\zeta}, \text{ where } 0 < \theta < \gamma < 1 < \zeta < \infty. \tag{11}$$

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 MFE of the MFG as shown in the next section.

5 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 mean-field and 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 (9)-(10) is more amenable to analysis and leads to good approximations of Γ_1 and Γ_2 operators. In contrast, earlier works [17, 18, 19] deal with approximating just Γ_1 under a time invariant MC. Finally we show that the two time-scale updates of the control policy and the mean-field using these approximate operators will converge to the MFE.

We first prove convergence of the transition probability and Q-learning estimation inside each episode $k \in [K]$. These results are worthy of interest independent of the Sandbox learning algorithm, due to the slowly time-varying MC setting. Lemma 1 presents error bounds on transition probability estimation (7) for a slowly time-varying MC, under a sufficient exploration condition as given below. Before discussing Assumption 2, we would like to point out that the main contribution of this paper is removing the assumption of access to a mean-field oracle. We contend that access to an oracle generating the mean field realizations [17, 18] is stronger relative to Assumption 2, as it implicitly assumes the knowledge of the distribution of all other agents, which never holds in practice. In particular, in the case of financial asset pricing where one uses a mean field to represent the distribution of institutional versus retail investors, it is impossible to have knowledge of the mean field, and instead one must estimate it based on market indicators – see [51].

Assumption 2. (Sufficient Exploration) *For any mean-field $\mu \in \mathcal{P}(\mathcal{S})$ and policy of the form $(1 - \psi)\pi + \psi \mathbb{1}_{\{\mathcal{A}\}}$ where $\pi : \mathcal{S} \times \mathcal{P}(\mathcal{S}) \to \mathcal{P}(\mathcal{A})$ and $\psi \in (0, 1)$, there exists a $\sigma \in (0, 1)$ and a positive integer τ such that for any $(s, a) \in |\mathcal{S} \times \mathcal{A}|$ and $t \geq \tau$, $\mathbb{P}((s_t, a_t) = (s, a) \mid \mathcal{F}_{t-\tau}) \geq \sigma$.*

Q-learning for N-player stochastic games also requires a similar assumption (Assumption 1 [52]). Assumption 2 generalizes the covering time assumption in RL-MFG literature (Lemma 8 [17]). We also note that it is similar to

Algorithm 1: Sandbox RL for MFG

1: **Initialize**: initial state $s_1^1 \sim p_1$, policy π_0^1 and mean-field μ_0^1
2: **Step-sizes**:

$$c_{\mu,t}^k = c_\mu^k \frac{1}{t^\zeta}, c_{\pi,t}^k = c_\pi^k \frac{1}{t^\zeta}, c_\mu^k = \frac{c_\mu}{k^\gamma}, c_\pi^k = \frac{c_\pi}{k^\theta}, 0 < \theta < \gamma < 1 < \zeta < \infty,$$
$$\lambda^k = \log(\max\{k, 3\} - 1), \beta_t = \frac{c_\beta}{(t+1)^\nu}, \nu \in (0.5, 1], \psi_1 = 0 \text{ and } \psi_t = 0.2, t > 1$$

3: **for** $k \in \{1, 2, \ldots, K\}$ **do**
4: **for** $t \in \{1, 2, \ldots, T\}$ **do**
5: $\mu_t^k = \mathbb{P}_{S(\epsilon^{\text{net}})}\big[(1 - c_{\mu,t}^k)\mu_{t-1}^k + c_{\mu,t}^k\big((\hat{P}_t^k)^\top \mu_{t-1}^k\big), 1 - \psi_t\big]$
6: $\pi_t^k = (1 - c_{\pi,t}^k)\pi_{t-1}^k + c_{\pi,t}^k\big((1 - \psi_t)\text{softmax}_{\lambda^k}(\cdot, Q_t^k) + \psi_t \mathbb{1}_{|\mathcal{A}|}\big)$
7: Generate a single transition $s_{t+1}^k \sim P(\cdot \mid s_t^k, a_t^k, \mu_t^k)$ and reward $r_t^k = R(s_t^k, a_t^k, \mu_t^k)$ where $a_t^k \sim \pi_t^k(s_t^k, \mu_t^k)$.
8: **Transition probability estimation:** For $(i, j) \in \mathcal{S} \times \mathcal{S}$

$$\hat{P}_{t+1}^k(i, j) = \frac{N_t^k(i, j) + 1/S}{N_t^k(i + 1)}, \tag{12}$$

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

Assumption 3 in [22] and is more general than an ergodicity assumption used in the stochastic optimization literature [53]. Indeed, using Proposition 3 in [22] we can show that for a given mean-field μ and policy π if the MC of the agent is ergodic then it also satisfies Assumption 2. A setting where sufficient exploration will hold is crowd dynamics [54], where the agent (pedestrian) takes an action and with some probability p lands in the desired state but with probability $1 - p$ it may be 'jostled' by the crowd and might land in a neighboring state. This ensures that the agent sufficiently explores its state space given a long enough time horizon. A recent work [55] presents the empirical finding that greedy policy update as in (line 6 Algorithm 1) causes a policy churn which induces sufficient exploration of the state space.

Failure to satisfy this condition may lead to uneven exploration of the state and action spaces, resulting in bad transition probability estimates. The sufficient exploration condition has been used before in two time-scale settings in the literature [26] and is similar to the uniform finite concentrability assumption in RL in the MFG literature [18]. The transition probability estimation was shown to converge in [56] for the time invariant MC. Lemma 1 quantifies the error in transition probability estimation under sufficient exploration, Lipschitz conditions on transition probability P [21] and a slowly time-varying MC. 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 mean-field at the first timestep (π_1^k, μ_1^k, respectively).

Lemma 1. *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*

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

with probability at least $1 - \delta_P$ where $\tilde{\mathcal{O}}$ hides logarithmic terms and $P_{\pi_1^k, \mu_1^k}$ 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 [57] and crowd management [54], continuity of transition probability w.r.t. mean-field and policy is ensured. Lemma 1 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. This drift can be decreased by increasing the inter-episodic learning parameter ζ. But increasing ζ to ∞ will stop inter-episodic learning and the algorithm might get stuck with a degenerate control policy (non-totally mixed control policy [44]) for the duration of the episode. This will contradict Assumption 2 and cause the transition probability estimation to fail. We avoid degenerate control policies by adding exploration noise in the control policy update $\mathbb{1}_A$ (Algorithm 1 line 6) given inter-episodic learning ($\zeta < \infty$). 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 Appendix and relies on Freedman's inequality [58].

Next we move on to analyzing the error in Q-learning estimation (8) for each episode $k \in [K]$. This update has been shown to converge to the optimal Q function under the sufficient exploration condition (Assumption 2) for a time invariant MC [49, 22]. In Lemma 2 we show that this update converges (albeit with a drift) under the sufficient exploration condition even 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 1, a drift term $\mathcal{O}(2^{1-\zeta})$ creeps in due to the slowly time-varying MC setting.

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

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

with probability at least $1 - \delta_Q$, where $\tilde{\mathcal{O}}$ hides logarithmic terms.

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 1 and 2 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.

Before proving the convergence bound of the Sandbox RL algorithm, we need to introduce an action gap assumption. We use $\text{softmax}_\lambda(\cdot)$ instead of argmax in the control policy update, as the Lipschitzness of softmax ensures a good approximate optimality operator, given that ϵ_Q^k is small. But since we use $\text{softmax}_\lambda(\cdot)$ instead of argmax, an error which depends on the action gap of Q_μ^* function [50] creeps into our analysis. Q_μ^* is the optimal Q-function for MDP induced by mean-field μ. In order to bound this error, we need to know the lower bound on action-gaps for Q_μ^* for $\mu \in S(\epsilon^{\text{net}})$.

Assumption 3. (Action Gaps) *For any mean-field $\mu \in S(\epsilon^{net})$, the optimal Q-function for the MDP induced by μ, Q_μ^*, has action gap $\Delta(Q_\mu^*) \geq \bar{\Delta} > 0$, where action gap for any function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^+$ is defined as in [17],* $\Delta(Q) := \min_{s\in\mathcal{S}} \left(\max_{a\in\mathcal{A}} Q(s,a) - \max_{a\in\mathcal{A}\setminus \text{argmax}_{a\in\mathcal{A}} Q(s,a)} Q(s,a)\right)$.

This assumption is not restrictive since action gaps are strictly positive by definition and the $\min$ over a finite set of strictly positive numbers is also strictly positive. Under this assumption, by increasing λ^k logarithmically with k (so that $\text{softmax}_{\lambda^k}(\cdot) \to \text{argmax}$) the overall error in estimating the MFE is shown to decrease in Theorem 1. Now we are in a position to present the main result of the paper, relying on good approximations of the consistency and optimality operators. Theorem 1 below bounds the average error in policy $e_\pi^k := \|\pi_1^k - \Gamma_1(\mu_1^k)\|_{TV}$ 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 $e_\pi^k := \|\pi_1^k - \Gamma_1(\mu_1^k)\|_{TV}$ 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 and 3, with the estimation errors satisfying $\epsilon_P^k \leq \epsilon_P, \epsilon_Q^k \leq \epsilon_Q/\log(K)$, for some $\epsilon_P, \epsilon_Q > 0$, the average approximation errors decrease at the following rates:*

$$\frac{1}{K}\sum_{k=1}^{K-1} e_\pi^k = \mathcal{O}(K^{\theta-1}) + \mathcal{O}(\epsilon_Q) + \mathcal{O}(K^{-\bar{\Delta}}) + \mathcal{O}(K^{\theta-\gamma}) + \mathcal{O}(K^{-1}) + \mathcal{O}(2^{1-\zeta}),$$

$$\frac{1}{K}\sum_{k=1}^{K-1} e_\mu^k = \mathcal{O}(K^{\gamma-1}) + \mathcal{O}(\epsilon) + \mathcal{O}(\epsilon_P) + \frac{1}{K}\sum_{k=1}^{K-1} e_\pi^k$$

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. The error term $\mathcal{O}(K^{-\bar{\Delta}})$ is due to the usage of $\text{softmax}_{\lambda^k}(\cdot)$ instead of argmax in the algorithm. The rationale for that has been provided in 4.2. By choosing λ^k to increase at a rate of $\log(k)$, the error can be bounded by $\mathcal{O}(K^{-\bar{\Delta}})$. This rate might seem restrictive since the uniform action gap $\bar{\Delta}$ might be small. But $\bar{\Delta}$ can be scaled up by an arbitrary value by scaling up the instantaneous rewards in line 9 of Algorithm 1, rendering the $\mathcal{O}(K^{-\bar{\Delta}})$ term *nice*. 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 MFE of the MFG. The proof of Corollary 1 depends on the result of Theorem 1 and is provided in the Appendix.

Corollary 1. *As a corollary to Theorem 1, we have*

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

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 quantity $\mathcal{O}(K^{-\bar{\Delta}})$ is due to the lack of regularization term in standard MFG setting but it is also monotonically decreasing with K and the rate of decrease can be accelerated by scaling up the rewards by a constant factor, hereby increasing the action gap $\bar{\Delta}$. Reward scaling would not affect the analysis due to bounded reward function. 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 as discussed after Lemma 1. The error introduced by the projection step in the mean-field update line 5 Algorithm 1 is $\mathcal{O}(\epsilon)$.

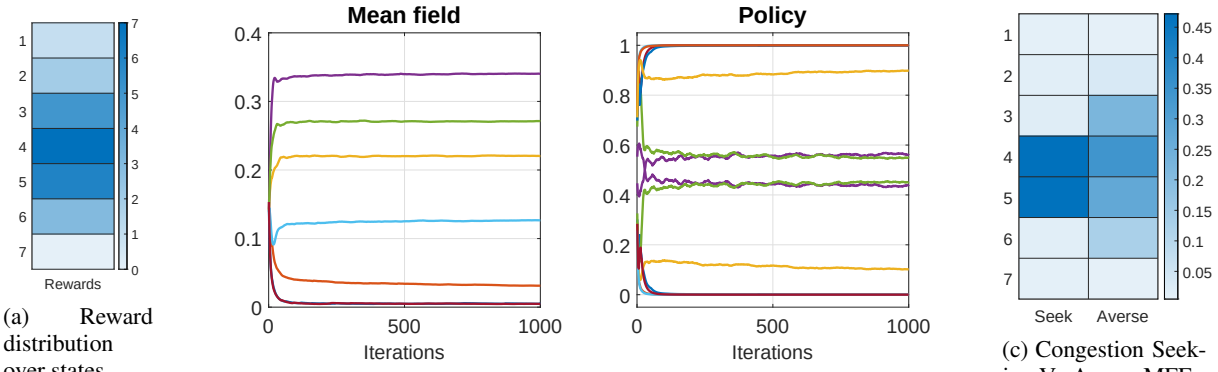
If the learning rates are chosen such that $\theta = 0.01, \gamma = 0.5, \nu = 1, \zeta = \Omega(\log(1/\epsilon))$ then the output of Algorithm 1 will be ϵ close to the MFE with high probability (for small enough $\epsilon > 0$), given that episode length is $T = \Omega(\epsilon^{-2})$ and the number 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 $\mathcal{O}(\epsilon^{-4})$. In the next section we apply the algorithm to a congestion game.

6 Numerical results

We simulate Sandbox learning for a simple congestion MFG [59] on one axis, with state space $\mathcal{S} = \{1, \ldots, 7\}$, action space $\mathcal{A} = \{-1, 1\}$, and discount factor $\rho = 0.9$. The initial condition of the agent is concentrated at state 1, and thus the agent is initialized at $s_1^1 = 1$. The generic agent's dynamics are assumed to be independent of the mean-field μ. If the agent takes action $a = -1$ it moves one state down (and if $a = 1$ it moves up) with probability 0.9. With probability 0.1 the agent's next state is chosen uniformly from the state space. The agent can be congestion seeking or averse based on reward function's dependence on μ. The agent's reward is $r(s, \mu) = (1 - c \cdot \mu(s)) \cdot R(s)$ where c is the congestion seeking/averse parameter and $R = \{1, 2, 5, 7, 6, 3, 0\}$. If $c < 0$ the agent is congestion-seeking, and if $0 < c < 1$ the agent is congestion-averse. The state-dependent rewards $R(s)$ are concentrated around favorable states $\{3, 4, 5\}$.

Now we introduce the parameters in the Sandbox learning algorithm. 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 $T = 10000$ we observe that Q-learning and transition probability estimation converge very well to their true values, inside each episode k. Furthermore, by choosing $c_\mu = c_\pi = 0.5$ and the two timescale learning rates as $\theta = 0.55 < \gamma = 0.6$ we see that the control policy and mean-field estimates approach their true values after $K = 1000$ many episodes.

Figure 2a shows the state space of the agent with the reward distribution. The agent aims to go to favorable states $\{4, 5\}$; but might be averse to $(0 < c < 1)$ or seeking of $(c < 0)$ high congestion in those states. Figure 2b shows fast convergence of control policy and mean-field estimates using the Sandbox learning algorithm, for the congestion



(a) Reward distribution over states

(b) Convergence for Congestion Averse MFGs

(c) Congestion Seeking Vs Averse MFE

Figure 2: Numerical Analysis of Sandbox learning

averse setting. In Figure 2c we the contrast the performance of the Sandbox learning algorithm for congestion seeking (higher rewards for states with higher congestion) and congestion averse (lower rewards for states with higher congestion) settings. As shown in Figure 2c a congestion seeking MFG leads to very higher concentration of agents around the favorable states $\{4, 5\}$, whereas a congestion averse MFG leads to a distribution which is more spread out as favorable states appear less appealing due to congestion.

7 Conclusion and future work

In this paper we have developed the Sandbox learning algorithm with finite-time guarantees to learn the stationary MFE of a finite-state finite-action MFG without access to a mean-field oracle. The sample complexity of the Sandbox learning algorithm is $\mathcal{O}(\epsilon^{-4})$. 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 timescale learning rates and approximate consistency and optimality operators. This work opens up several interesting research directions. It would be worthwhile to investigate whether the exploration noise can be crafted to relax the sufficient exploration condition. Another important research direction would be to explore how feature embeddings can improve the scalability beyond the tabular MFG setting.

References

[1] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep exploration via bootstrapped dqn," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[2] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1889–1897.

[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[4] B. O'Donoghue, R. Munos, K. Kavukcuoglu, and V. Mnih, "Combining policy gradient and q-learning," *arXiv preprint arXiv:1611.01626*, 2016.

[5] X. Xiang and S. Foo, "Recent advances in deep reinforcement learning applications for solving partially observable markov decision processes (pomdp) problems: Part 1—fundamentals and applications in games, robotics and natural language processing," *Machine Learning and Knowledge Extraction*, vol. 3, no. 3, pp. 554–581, 2021.

[6] V. Petrenko and M. Gurchinskiy, "Multi-agent deep reinforcement learning concept for mobile cyber-physical systems control," in *E3S Web of Conferences*, vol. 270. EDP Sciences, 2021, p. 01036.

[7] J. Lee, R. Kim, S.-W. Yi, and J. Kang, "Maps: multi-agent reinforcement learning-based portfolio management system," *arXiv preprint arXiv:2007.05402*, 2020.

[8] A. Ghasemi, A. Shojaeighadikolaei, K. Jones, M. Hashemi, A. G. Bardas, and R. Ahmadi, "A multi-agent deep reinforcement learning approach for a distributed energy marketplace in smart grids," in *2020 IEEE In-*

ternational Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm). IEEE, 2020, pp. 1–6.

[9] Y. Shoham, R. Powers, and T. Grenager, "If multi-agent learning is the answer, what is the question?" *Artificial Intelligence*, vol. 171, no. 7, pp. 365–377, 2007.

[10] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021.

[11] E. Sonu, Y. Chen, and P. Doshi, "Decision-theoretic planning under anonymity in agent populations," *Journal of Artificial Intelligence Research*, vol. 59, pp. 725–770, 2017.

[12] M. Huang, R. P. Malhamé, and P. E. Caines, "Large population stochastic dynamic games: Closed-loop Mckean-Vlasov systems and the Nash certainty equivalence principle," *Communications in Information & Systems*, vol. 6, no. 3, pp. 221–252, 2006.

[13] M. Huang, P. E. Caines, and R. P. Malhamé, "Large-population cost-coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized ε-Nash equilibria," *IEEE Transactions on Automatic Control*, vol. 52, no. 9, pp. 1560–1571, 2007.

[14] J.-M. Lasry and P.-L. Lions, "Jeux à champ moyen. i–le cas stationnaire," *Comptes Rendus Mathématique*, vol. 343, no. 9, pp. 619–625, 2006.

[15] ——, "Mean field games," *Japanese Journal of Mathematics*, vol. 2, no. 1, pp. 229–260, 2007.

[16] N. Saldi, T. Başar, and M. Raginsky, "Markov–Nash equilibria in mean-field games with discounted cost," *SIAM Journal on Control and Optimization*, vol. 56, no. 6, pp. 4256–4287, 2018.

[17] X. Guo, A. Hu, R. Xu, and J. Zhang, "Learning mean-field games," in *Advances in Neural Information Processing Systems*, 2019.

[18] Q. Xie, Z. Yang, Z. Wang, and A. Minca, "Learning while playing in mean-field games: Convergence and optimality," in *International Conference on Machine Learning*. PMLR, 2021, pp. 11 436–11 447.

[19] B. Anahtarcı, C. D. Karıksız, and N. Saldi, "Fitted Q-learning in mean-field games," *arXiv preprint arXiv:1912.13309*, 2019.

[20] Z. Fu, Z. Yang, Y. Chen, and Z. Wang, "Actor-critic provably finds Nash equilibria of linear-quadratic mean-field games," in *International Conference on Learning Representation*, 2020.

[21] A. Angiuli, J.-P. Fouque, and M. Laurière, "Unified reinforcement Q-learning for mean field game and control problems," *Mathematics of Control, Signals, and Systems*, pp. 1–55, 2022.

[22] G. Qu and A. Wierman, "Finite-time analysis of asynchronous stochastic approximation and Q-learning," in *Conference on Learning Theory*. PMLR, 2020, pp. 3185–3205.

[23] R. Elie, J. Pérolat, M. Laurière, M. Geist, and O. Pietquin, "Approximate fictitious play for mean field games," *arXiv preprint arXiv:1907.02633*, 2019.

[24] K. Cui and H. Koeppl, "Approximately solving mean field games via entropy-regularized deep reinforcement learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 1909–1917.

[25] B. Anahtarci, C. D. Kariksiz, and N. Saldi, "Q-learning in regularized mean-field games," *Dynamic Games and Applications*, pp. 1–29, 2022.

[26] Y. F. Wu, W. Zhang, P. Xu, and Q. Gu, "A finite-time analysis of two time-scale actor-critic methods," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 617–17 628, 2020.

[27] J. Moon and T. Başar, "Discrete-time LQG mean field games with unreliable communication," in *53rd IEEE Conference on Decision and Control*. IEEE, 2014, pp. 2697–2702.

[28] P. E. Caines and M. Huang, "Graphon mean field games and the GMFG equations: ε-Nash equilibria," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 286–292.

[29] M. A. U. Zaman, S. Bhatt, and T. Başar, "Adversarial linear-quadratic mean-field games over multigraphs," in *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2021, pp. 209–214.

[30] H. Tembine, Q. Zhu, and T. Başar, "Risk-sensitive mean-field games," *IEEE Transactions on Automatic Control*, vol. 59, no. 4, pp. 835–850, 2013.

[31] J. Moon and T. Başar, "Linear quadratic risk-sensitive and robust mean field games," *IEEE Transactions on Automatic Control*, vol. 62, no. 3, pp. 1062–1077, 2016.

[32] N. Saldi, T. Başar, and M. Raginsky, "Approximate markov-nash equilibria for discrete-time risk-sensitive mean-field games," *Mathematics of Operations Research*, vol. 45, no. 4, pp. 1596–1620, 2020.

[33] A. Bensoussan, T. Huang, and M. Laurière, "Mean field control and mean field game models with several populations," *arXiv preprint arXiv:1810.00783*, 2018.

[34] J. Barreiro-Gomez and H. Tembine, *Mean-field-type Games for Engineers*. CRC Press, 2021.

[35] Z. Ma, D. S. Callaway, and I. A. Hiskens, "Decentralized charging control of large populations of plug-in electric vehicles," *IEEE Transactions on Control Systems Technology*, vol. 21, no. 1, pp. 67–78, 2011.

[36] R. Carmona, "Applications of mean field games in financial engineering and economic theory," *arXiv preprint arXiv:2012.05237*, 2020.

[37] A. Lachapelle and M.-T. Wolfram, "On a mean field game approach modeling congestion and aversion in pedestrian crowds," *Transportation Research Part B: Methodological*, vol. 45, no. 10, pp. 1572–1589, 2011.

[38] M. A. U. Zaman, K. Zhang, E. Miehling, and T. Başar, "Reinforcement learning in non-stationary discrete-time linear-quadratic mean-field games," in *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 2278–2284.

[39] M. A. U. Zaman, E. Miehling, and T. Başar, "Reinforcement learning for non-stationary discrete-time linear–quadratic mean-field games in multiple populations," *Dynamic Games and Applications*, pp. 1–47, 2022.

[40] M. Fazel, R. Ge, S. M. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *International Conference on Machine Learning*, 2018, pp. 1467–1476.

[41] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. Bartlett, and M. Wainwright, "Derivative-free methods for policy optimization: Guarantees for linear quadratic systems," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 2916–2925.

[42] S. Perrin, M. Laurière, J. Pérolat, M. Geist, R. Élie, and O. Pietquin, "Mean field games flock! the reinforcement learning way," *arXiv preprint arXiv:2105.07933*, 2021.

[43] J. Subramanian and A. Mahajan, "Reinforcement learning in stationary mean-field games," in *International Conference on Autonomous Agents and Multiagent Systems*, 2019, pp. 251–259.

[44] T. Başar and G. J. Olsder, *Dynamic Noncooperative Game Theory*. SIAM, 1998.

[45] B. Jovanovic and R. W. Rosenthal, "Anonymous sequential games," *Journal of Mathematical Economics*, vol. 17, no. 1, pp. 77–87, 1988.

[46] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[47] A. Bensoussan, J. Frehse, and S. C. P. Yam, "The master equation in mean field theory," *Journal de Mathématiques Pures et Appliquées*, vol. 103, no. 6, pp. 1441–1474, 2015.

[48] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal Control*. John Wiley & Sons, 2012.

[49] E. Even-Dar, Y. Mansour, and P. Bartlett, "Learning rates for Q-learning." *Journal of Machine Learning Research*, vol. 5, no. 1, 2003.

[50] B. Gao and L. Pavel, "On the properties of the softmax function with application in game theory and reinforcement learning," *arXiv preprint arXiv:1704.00805*, 2017.

[51] N. Vadori, S. Ganesh, P. Reddy, and M. Veloso, "Calibration of shared equilibria in general sum partially observable markov games," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 118–14 128, 2020.

[52] J. Hu and M. P. Wellman, "Nash Q-learning for general-sum stochastic games," *Journal of Machine Learning Research*, vol. 4, no. Nov, pp. 1039–1069, 2003.

[53] R. Srikant and L. Ying, "Finite-time error bounds for linear stochastic approximation and TD learning," in *Conference on Learning Theory*. PMLR, 2019, pp. 2803–2830.

[54] F. S. Priuli, "First order mean field games in crowd dynamics," *arXiv preprint arXiv:1402.7296*, 2014.

[55] T. Schaul, A. Barreto, J. Quan, and G. Ostrovski, "The phenomenon of policy churn," *arXiv preprint arXiv:2206.00730*, 2022.

[56] D. Hsu, A. Kontorovich, D. A. Levin, Y. Peres, C. Szepesvári, and G. Wolfer, "Mixing time estimation in reversible Markov chains from a single sample path," *The Annals of Applied Probability*, vol. 29, no. 4, pp. 2439–2480, 2019.

[57] G. d. Reis and V. Platonov, "Forward utilities and mean-field games under relative performance concerns," in *From Particle Systems to Partial Differential Equations*. Springer, 2019, pp. 227–251.

[58] D. A. Freedman, "On tail probabilities for martingales," *the Annals of Probability*, pp. 100–118, 1975.

[59] N. Toumi, R. Malhamé, and J. 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)*. IEEE, 2020, pp. 1020–1025.

Supplementary Material for "Oracle-free Reinforcement Learning in Mean-Field Games along a Single Sample Path"

A Proofs of Results in Section 5

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. In this proof we provide finite sample convergence bounds for the transition probability estimation (7) under the slowly time-varying MC setting. The proof of Lemma 1 relies on Freedman's inequality [58] similar to the analysis of Theorem 4 in [56]. 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 (Assumption 2), 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 [56]. 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 $\{s_1, \mu_1, \pi_1, \ldots, s_t, \mu_t, \pi_t\}$. 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 $t \in [T]$,

$$\begin{aligned}S_T := \sum_{t=1}^{T} Y_t &= \sum_{t=2}^{T} \mathbb{1}\{s_{t-1} = i\}\big(\mathbb{1}\{s_t = j\} - P_{t-1}(i, j)\big), \\ &= \sum_{t=2}^{T} \mathbb{1}\{s_{t-1} = i\}\big(\mathbb{1}\{s_t = j\} - P_1(i, j) + P_1(i, j) - P_{t-1}(i, j)\big), \\ &= N_{i,j} - N_i P_1(i, j) + \sum_{t=2}^{T} \mathbb{1}\{s_{t-1} = i\}\tilde{P}_{t-1}(i, j), \end{aligned} \tag{13}$$

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

$$V_T := \sum_{t=2}^{T} \mathbb{E}[Y_t^2 \mid \mathcal{F}_{t-1}] = N_i P_1(i, j)\big(1 - P_1(i, j)\big) + \sum_{t=2}^{T} \mathbb{1}\{s_{t-1} = i\}\tilde{P}_{t-1}(i, j)\big(2P_1(i, j) - 1 - \tilde{P}_{t-1}(i, j)\big). \tag{14}$$

Again both parts of the above expression are positive. Recalling (12) 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 (13). Applying Corollary 1 from [56], which is based on Freedman's inequality, we get

$$\begin{aligned}&|\hat{P}_T(i,j) - P_1(i,j)|\\ &\leq \sqrt{\frac{2cV_T\tau_{T,\delta_P}}{(N_i+1)^2}} + \frac{4\tau_{T,\delta_P} + \sum_{t=2}^{T}\mathbb{1}\{s_{t-1}=i\}|\tilde{P}_{t-1}(i,j)| + |1/S - P_1(i,j)|}{N_i+1},\\ &= \Bigg(\frac{2cN_iP_1(i,j)\big(1-P_1(i,j)\big)\tau_{T,\delta_P}}{(N_i+1)^2}\\ &\qquad + \frac{2c\sum_{t=2}^{T}\mathbb{1}\{s_{t-1}=i\}\tilde{P}_{t-1}(i,j)\big(2P_1(i,j) - 1 - \tilde{P}_{t-1}(i,j)\big)\tau_{T,\delta_P}}{(N_i+1)^2}\Bigg)^{\frac{1}{2}}\\ &\qquad + \frac{4\tau_{T,\delta_P} + \sum_{t=2}^{T}\mathbb{1}\{s_{t-1}=i\}|\tilde{P}_{t-1}(i,j)| + |1/S - P_1(i,j)|}{N_i+1},\\ &\leq \sqrt{\frac{2cN_iP_1(i,j)\big(1-P_1(i,j)\big)\tau_{T,\delta_P}}{(N_i+1)^2}} + \frac{\sqrt{2c\sum_{t=2}^{T}|\tilde{P}_{t-1}(i,j)|\tau_{T,\delta_P}}}{N_i+1}\\ &\qquad + \frac{4\tau_{T,\delta_P} + \sum_{t=2}^{T}|\tilde{P}_{t-1}(i,j)| + |1/S - P_1(i,j)|}{N_i+1},\\ &\leq \sqrt{\frac{2c\tau_{T,\delta_P}}{N_i+1}} + \frac{\sqrt{2c\sum_{t=2}^{T}|\tilde{P}_{t-1}(i,j)|\tau_{T,\delta_P}}}{N_i+1} + \frac{4\tau_{T,\delta_P} + \sum_{t=2}^{T}|\tilde{P}_{t-1}(i,j)| + |1/S - P_1(i,j)|}{N_i+1},\end{aligned} \tag{15}$$

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 (14) to obtain the second inequality. Analyzing $|\tilde{P}_t(i,j)|$,

$$\begin{aligned}|\tilde{P}_t(i,j)| &\leq \|P_1 - P_t\|_F = \|P_{\pi_1,\mu_1} - P_{\pi_t,\mu_t}\|_F,\\ &\leq \sum_{l=1}^{t-1}\big(\|P_{\pi_l,\mu_l} - P_{\pi_l,\mu_{l+1}}\|_F + \|P_{\pi_l,\mu_{l+1}} - P_{\pi_{l+1},\mu_{l+1}}\|_F\big),\\ &\leq \sum_{l=1}^{t-1}\big(L_P^{\mu}\|\mu_{l+1}-\mu_l\|_1 + L_P^{\pi}\|\pi_{l+1}-\pi_l\|_{TV}\big)\\ &\leq \sum_{l=2}^{t}\big(L_P^{\mu}c_{\mu,l} + L_P^{\pi}c_{\pi,l}\big),\\ &\leq (L_P^{\mu}c_{\mu} + L_P^{\pi}c_{\pi})\sum_{l=2}^{t} l^{-\zeta},\\ &\leq \frac{L_P^{\mu}c_{\mu} + L_P^{\pi}c_{\pi}}{\zeta-1}2^{1-\zeta} = \tilde{L}_P 2^{1-\zeta}.\end{aligned} \tag{16}$$

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 1 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 (15) and (16):

$$\begin{aligned}&|\hat{P}_T(i,j) - P_1(i,j)| \\ &\leq \sqrt{\frac{2c\tau_{T,\delta_P}}{N_i+1}} + \frac{\sqrt{2c\tau_{T,\delta_P}\tilde{L}_P T}}{N_i+1} 2^{\frac{1-\zeta}{2}} + \frac{4\tau_{T,\delta_P} + \tilde{L}_P T 2^{1-\zeta} + |1/S - P_1(i,j)|}{N_i+1},\end{aligned} \tag{17}$$

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 3. *Under Assumption 2, $N_i \geq T/T_e$ with probability at least $1 - \delta_P/(2S^2)$, where*

$$T_e := \mathcal{O}\Big(\frac{1}{\sigma}\log\Big(\frac{2S^3}{\delta_P}\Big)\Big). \tag{18}$$

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 (Assumption 2). 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 Assumption 2 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 (17) and Lemma 3 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.2 Proof of Lemma 2

Proof. In this proof we provide finite sample convergence bounds for the Q-learning update (8) within the slowly time-varying MC setting. The proof of Lemma 2 follows an approach similar to proof of Theorem 4 in [22] and extends the results to a slowly time-varying MC and learning exponent $0.5 < \nu \leq 1$, which is empirically observed to have better convergence properties. We also find that convergence cannot be guaranteed for $0 < \nu \leq 0.5$. 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 (8) is lower bounded by $\frac{1}{\sigma}\max\Big\{\nu + \zeta - 1, \frac{1}{(1-\sqrt{\rho})}\Big\}$. 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:

$$Q_{t+1} = Q_t + \beta_t[e_{i_t}^T[F(\mu_t, Q_t) - Q_t] + w(t, \mu_t)]e_{i_t} \tag{19}$$

where

$$\begin{aligned} w(t,\mu_t) &= \rho[\max_{a\in\mathcal{A}} Q_t(s_{t+1}, a) - \mathbb{E}_{s'\sim P(\cdot|s_t,a_t,\mu_t)}[\max_{a'\in\mathcal{A}} Q_t(s', a')], \\ F(\mu_t, Q_t)(s,a) &= r(s,a,\mu_t) + \rho\mathbb{E}_{s'\sim P(\cdot|s_t,a_t,\mu_t)}[\max_{a'\in\mathcal{A}} Q_t(s', a')] \end{aligned} \tag{20}$$

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 $D_t := \mathbb{E}[e_{i_t}^T e_{i_t} \mid \mathcal{F}_{t-\tau}]$. 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 assumption (Assumption 2) we know that $d_{t,i} \geq \sigma > 0$.

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

Let us define

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

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

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

Hence, ϵ_t is a shifted Martingale Difference Sequence. Writing down the Q-function as a sum from τ (Assumption 2) to t, we get

$$Q_{t+1} = \tilde{B}_{\tau-1,t}Q_\tau + \sum_{k=\tau}^{t} B_{k,t}F(\mu_k, Q_k) + \sum_{k=\tau}^{t} \beta_t \tilde{B}_{k,t}(\epsilon_k + \phi_k), \tag{21}$$

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 4.

$$\begin{aligned} e_{t+1}^Q \leq \tilde{B}_{\tau-1,t}e_\tau^Q + \rho\max_i \sum_{k=\tau}^{t} b_{k,t,i}e_k^Q + \Big\| \sum_{k=\tau}^{t} \beta_k \tilde{B}_{k,t}(\epsilon_k + \phi_k)\Big\|_\infty \\ + L_Q^\mu\Big[\tilde{\beta}_{\tau-1,t}\sum_{l=\tau}^{t} c_{\mu,l} + \sum_{k=\tau}^{t}\beta_{k,t}\sum_{l=k}^{t} c_{\mu,l}\Big] \end{aligned} \tag{22}$$

Proof. Using (21) 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 [48] 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 (22).

Lemma 5.

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

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 (20) and Q-update (19). Similarly for ϕ we get

$$\begin{aligned}
\|\phi_t\|_\infty &= \|(e_{i_t}^T e_{i_t} - D_t)(F(\mu_t, Q_t) - F(\mu_{t-\tau}, Q_{t-\tau}) - Q_t + Q_{t-\tau})\|_\infty, \\
&\leq \|F(\mu_t, Q_t) - F(\mu_{t-\tau}, Q_{t-\tau})\|_\infty + \|Q_{t-\tau} - Q_t\|_\infty, \\
&\leq \sum_{k=1}^{\tau}\|F(\mu_{t-k+1}, Q_{t-k+1}) - F(\mu_{t-k}, Q_{t-k})\|_\infty + \sum_{k=1}^{\tau}\|Q_{t-k+1} - Q_{t-k}\|_\infty. \qquad (23)
\end{aligned}$$

We first analyze the first summand in equation (23)

$$\begin{aligned}
&\|F(\mu_{t-k+1}, Q_{t-k+1}) - F(\mu_{t-k}, Q_{t-k})\|_\infty \\
&\leq \|F(\mu_{t-k+1}, Q_{t-k+1}) - F(\mu_{t-k+1}, Q_{t-k})\|_\infty + \|F(\mu_{t-k+1}, Q_{t-k}) - F(\mu_{t-k}, Q_{t-k})\|_\infty, \\
&\leq \rho\|Q_{t-k+1} - Q_{t-k}\|_\infty + \max_{s,a}|R(s,a,\mu_{t-k+1}) - R(s,a,\mu_{t-k})| \\
&\qquad\qquad + \max_{s,a}|P(\cdot \mid s,a,\mu_{t-k+1}) - P(\cdot \mid s,a,\mu_{t-k})|\frac{1}{1-\rho}, \\
&\leq \rho\|Q_{t-k+1} - Q_{t-k}\|_\infty + \Big(L_R^\mu + \frac{L_P^\mu}{1-\rho}\Big)\|\mu_{t-k+1} - \mu_{t-k}\|_1 \qquad (24)
\end{aligned}$$

Similarly the second summand in (23) is

$$\begin{aligned}
\|Q_{t-k+1} - Q_{t-k}\|_\infty &= \beta_{t-k}\|e_{i_{t-k}}^T\big(F(\mu_{t-k}, Q_{t-k}) - Q_{t-k} + w(t-k, \mu_{t-k})\big)e_{i_{t-k}}\|_\infty, \\
&\leq \beta_{t-k}\|F(\mu_{t-k}, Q_{t-k})\|_\infty + \beta_{t-k}\|Q_{t-k}\|_\infty + \bar{w} \leq \beta_{t-k}\bar{\epsilon}. \qquad (25)
\end{aligned}$$

Substituting (24), (25) into (23)

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

□

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

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

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

Substituting into the sum

$$\sum_{k=1}^{t}\beta_{k,t}^2 \leq \sum_{k=1}^{t}\frac{2c_\beta^2}{(t+2)^{2c_\beta\sigma}}(k+1)^{2c_\beta\sigma-2\nu} \leq \frac{2c_\beta^2}{(t+2)^{2c_\beta\sigma}}\int_1^{t+1}(y+1)^{2c_\beta\sigma-2\nu}dy,$$
$$\leq \frac{2c_\beta^2}{(t+2)^{2c_\beta\sigma}}\frac{(t+2)^{2c_\beta\sigma-2\nu+1}}{2c_\beta\sigma-2\nu+1} = \frac{2c_\beta^2}{2c_\beta\sigma-2\nu+1}\frac{1}{(t+2)^{2\nu-1}}$$

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

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

Hence the inequalities have been proved. □

Having proved some properties of the learning rates in Lemma 6 we are now able to bound the two parts of the quantity $\left\|\sum_{k=\tau}^{t}\beta_k\tilde{B}_{k,t}(\epsilon_k+\phi_k)\right\|_\infty$ 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 7.

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

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

$$C_\phi^1 = \frac{4c_\beta^2(1+\tau)^\nu\tau}{1+c_\beta\sigma-2\nu}\bar{\epsilon}, \tag{26}$$

$$C_\phi^2 = \left(L_R^\mu + \frac{L_P^\mu}{1-\rho}\right)\frac{2c_\mu c_\beta\tau(1+\tau)^\zeta}{c_\beta\sigma-\nu-\zeta+1}, \tag{27}$$

$$C_\epsilon = \frac{10\bar{\epsilon}}{\sqrt{2c_\beta\sigma-2\nu+1}}\sqrt{(\tau+1)c_\beta^2\log\left(\frac{2(\tau+1)T^2SA}{\delta_Q}\right)}. \tag{28}$$

Proof. We start with the first inequality

$$\begin{aligned}
\Big\| \sum_{k=\tau}^{t} \beta_k \tilde{B}_{k,t} \phi_k \Big\|_\infty &\leq \sum_{k=\tau}^{t} \beta_{k,t} \|\phi_k\|_\infty \\
&\leq \sum_{k=\tau}^{t} \beta_{k,t} \Big(\Big(L_R^\mu + \frac{L_P^\mu}{1-\rho} \Big) \sum_{l=1}^{\tau} \|\mu_{k-l+1} - \mu_{k-l}\|_\infty + 2\bar{\epsilon} \sum_{l=1}^{\tau} \beta_{k-l} \Big), \\
&\leq 2\bar{\epsilon} \sum_{k=\tau}^{t} \beta_{k,t} \sum_{l=k-\tau}^{k-1} \beta_l + \Big(L_R^\mu + \frac{L_P^\mu}{1-\rho} \Big) \sum_{k=tau}^{t} \beta_{k,t} \sum_{l=k-\tau}^{k-1} \|\mu_{l+1} - \mu_l\|_\infty, \\
&\leq C_\phi^1 \frac{1}{(t+2)^{2\nu-1}} + \Big(L_R^\mu + \frac{L_P^\mu}{1-\rho} \Big) \sum_{k=tau}^{t} \beta_{k,t} \sum_{l=k-\tau}^{k-1} c_{\mu,l}, \qquad (29)
\end{aligned}$$

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 $c_{\mu,l} = \frac{c_\mu}{(l+1)^\zeta} \leq \frac{c_\mu (\tau+1)^\zeta}{(k+1)^\zeta}$

$$\begin{aligned}
\sum_{k=tau}^{t} \beta_{k,t} \sum_{l=k-\tau}^{k-1} c_{\mu,l} &\leq \sum_{k=\tau}^{t} \beta_{k,t} \frac{c_\mu (\tau+1)^\zeta \tau}{(k+1)^\zeta} \leq \sum_{k=\tau}^{t} \frac{c_\mu \tau}{(k+1)^\zeta} \left(\frac{k+2}{t+2} \right)^{c_\beta \sigma} \frac{c_\beta (\tau+1)^\zeta}{(k+1)^\nu}, \\
&\leq 2 \sum_{k-\tau}^{t} \frac{c_\mu c_\beta \tau (\tau+1)^\zeta}{(t+2)^{c_\beta \sigma}} \frac{1}{(k+1)^{\nu+\zeta-c_\beta\sigma}}, \\
&\leq \frac{c_\mu c_\beta \tau (\tau+1)^\zeta}{(t+2)^{c_\beta\sigma}} \frac{(t+2)^{c_\beta\sigma-\nu-\zeta+1}}{c_\beta\sigma-\nu-\zeta+1} \leq \frac{c_\mu c_\beta \tau (\tau+1)^\zeta}{c_\beta\sigma-\nu-\zeta+1} \frac{1}{(t+2)^{\nu+\zeta-1}}, \qquad (30)
\end{aligned}$$

since $c_\beta \sigma \geq \nu + \zeta - 1$. Substituting (30) into (29) 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 [22]. 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 [22] we get

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

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

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

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

Now we aim to bound the last term in (22)

Lemma 8. *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 (22) 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 9. *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

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

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

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

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

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

which proves the recursion step and completes the proof. □

Now we prove the main result using Lemma 9. Recalling (22),

$$\begin{aligned}
e^Q_{t+1} \leq \tilde{B}_{\tau-1,t} e^Q_\tau + \rho \sup_i \sum_{k=\tau}^t b_{k,t,i} e^Q_k + \Big\| \sum_{k=\tau}^t \beta_k \tilde{B}_{k,t} (\epsilon_k + \phi_k) \Big\|_\infty \\
+ L^\mu_Q \Big[\tilde{\beta}_{\tau-1,t} \sum_{l=\tau}^t c_{\mu,l} + \sum_{k=\tau}^t \beta_{k,t} \sum_{l=k}^t c_{\mu,l} \Big] \qquad (31)
\end{aligned}$$

Using Lemmas 7 and 8 and using $C_\mu := 10 L^\mu_Q c_\mu$ for $\zeta \geq 1.1$,

$$e^Q_{t+1} \leq \tilde{B}_{\tau-1,t} e^Q_\tau + \rho \sup_i \sum_{k=\tau}^t b_{k,t,i} e^Q_k + \frac{C^1_\phi}{(t+2)^{2\nu-1}} + \frac{C^2_\phi}{(t+2)^{\zeta+\nu-1}} + \frac{C_\epsilon}{(t+2)^{\nu-1/2}} + \frac{C_\mu}{\tau^{\zeta-1}} \qquad (32)$$

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

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

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

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

Finally

$$\begin{aligned}
\epsilon_Q &= \|Q_T - Q_1^*\|_\infty, \\
&\leq \|Q_T - Q_T^*\|_\infty + \|Q_T^* - Q_1^*\|_\infty, \\
&\leq e_T^Q + \sum_{t=1}^{T-1} \|Q_{t+1}^* - Q_t^*\|_\infty, \\
&\leq \frac{\bar{C}_1}{(t+2)^{2\nu-1}} + \frac{\bar{C}_2}{(t+2)^{\zeta+\nu-1}} + \frac{\bar{C}_3}{(t+2)^{\nu-1/2}} + \frac{\bar{C}_4}{\tau^{\zeta-1}} + L_Q^\mu \sum_{k=1}^{T-1} \|\mu_{t+1} - \mu_t\|_1, \\
&\leq \frac{\bar{C}_1}{(t+2)^{2\nu-1}} + \frac{\bar{C}_2}{(t+2)^{\zeta+\nu-1}} + \frac{\bar{C}_3}{(t+2)^{\nu-1/2}} + \frac{\bar{C}_4}{\tau^{\zeta-1}} + L_Q^\mu \sum_{k=1}^{T-1} c_{\mu,t}, \\
&= \mathcal{O}(T^{1-2\nu}) + \mathcal{O}(T^{1-\zeta-\nu}) + \tilde{\mathcal{O}}(T^{1/2-\nu}) + \mathcal{O}(2^{1-\zeta})
\end{aligned}$$

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

A.3 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_π^k and e_μ^k, respectively. We start by characterizing the approximation errors in control policy and mean field e_π^k and e_μ^k 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_π^k which is evolving at a faster learning rate compared to the approximation error in the mean-field e_μ^k. This error is shown to converge due to the good approximation of the Q-function (Lemma 2), 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 1), the contraction mapping property (Assumption 1) and the convergence of e_π^k.

First we recall the update rules in Algorithm 1

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

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

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

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

$$\begin{aligned}
\mu^{k+1} &= \mathbb{P}_{S(\epsilon^{\text{net}})}\big[(1-c_{\mu}^{k+1})(\mu^k+\Delta_\mu^k) + c_{\mu}^{k+1}(\hat{P}^{k})^\top (\mu^k+\Delta_\mu^k), 1\big],\\
\pi^{k+1} &= (1-c_{\pi}^{k+1})(\pi^k+\Delta_\pi^k) + c_{\pi}^{k+1}\text{softmax}_{\lambda^{k+1}}(\cdot, Q^{k}).
\end{aligned} \tag{33}$$

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 $e_\pi^k := \|\pi^k - \Gamma_1(\mu^k)\|_{TV}$:

$$\begin{aligned}
e_\pi^{k+1} &= \|\pi^{k+1} - \Gamma_1(\mu^{k+1})\|_{TV},\\
&\le \|\pi^{k+1} - \Gamma_1(\mu^{k})\|_{TV} + \|\Gamma_1(\mu^{k}) - \Gamma_1(\mu^{k+1})\|_{TV},\\
&\le \|(1-c_\pi^{k+1})(\pi^k+\Delta_\pi^k) + c_\pi^{k+1}\text{softmax}_{\lambda^{k+1}}(\cdot,Q^k) - \text{argmax}(Q^*_{\mu^k})\|_{TV} + d_1\|\mu^{k+1}-\mu^k\|_1,\\
&\le (1-c_\pi^{k+1})\|\pi^k - \Gamma_1(\mu^k)\|_{TV} + (1-c_\pi^{k+1})\|\Delta_\pi^k\|_{TV}\\
&\qquad + c_\pi^{k+1}\|\text{softmax}_{\lambda^{k+1}}(\cdot,Q^k) - \text{argmax}(Q^*_{\mu^k})\|_{TV} + d_1\|\mu^{k+1}-\mu^k\|_1,\\
&\le (1-c_\pi^{k+1})e_\pi^k + \|\Delta_\pi^k\|_{TV} + c_\pi^{k+1}\|\text{softmax}_{\lambda^{k+1}}(\cdot,Q^k) - \text{softmax}_{\lambda^{k+1}}(\cdot,Q^*_{\mu^k})\|_{TV}\\
&\qquad + c_\pi^{k+1}\|\text{softmax}_{\lambda^{k+1}}(\cdot,Q^*_{\mu^k}) - \text{argmax}(Q^*_{\mu^k})\|_{TV} + d_1\|\mu^{k+1}-\mu^k\|_1.
\end{aligned} \tag{34}$$

To simplify the above expression we introduce some properties of the $\text{softmax}_\lambda(\cdot, Q)$ operator.

Lemma 10. *The $\text{softmax}_\lambda(\cdot,Q)$ satisfies the following properties for $\lambda > 0$ and $Q : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^+$,*

$$\begin{aligned}
\|\textit{softmax}_\lambda(\cdot,Q) - \textit{softmax}_\lambda(\cdot,Q')\|_{TV} &\le \lambda S\sqrt{A}\|Q-Q'\|_\infty,\\
\|\textit{softmax}_\lambda(\cdot,Q) - \text{argmax}(Q)\|_{TV} &\le 2SA\exp(-\lambda\bar{\Delta})
\end{aligned}$$

where $\bar{\Delta}$ is the action gap for Q as defined in Assumption 3.

Proof. The Lipschitzness of softmax can be obtained using Proposition 4 in [50]. Let us denote the policy under $\text{softmax}_\lambda(\cdot,Q)$ 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

$$\begin{aligned}
\|\text{softmax}_\lambda(\cdot,Q) - \text{softmax}_\lambda(\cdot,Q')\|_{TV} &= \|\pi_Q^\lambda - \pi_{Q'}^\lambda\|_{TV},\\
&= \max_{a\in\mathcal{A}}\sum_{s\in\mathcal{S}}\big|\pi_Q^\lambda(a|s) - \pi_{Q'}^\lambda(a|s)\big|
\end{aligned} \tag{35}$$

From Proposition 4 in [50] we know that for any $s\in\mathcal{S}$

$$\begin{aligned}
\|\pi_Q^\lambda(\cdot|s) - \pi_{Q'}^\lambda(\cdot|s)\|_2 = \sqrt{\sum_{a\in\mathcal{A}}\big(\pi_Q^\lambda(a|s) - \pi_{Q'}^\lambda(a|s)\big)^2} &\le \lambda\|Q(s,\cdot) - Q(s,\cdot)\|_2,\\
&= \lambda\sqrt{\sum_{a\in\mathcal{A}}\big(Q(s,a)-Q'(s,a)\big)^2},\\
&\le \lambda\sqrt{A}\|Q(s,\cdot)-Q'(s,\cdot)\|_\infty,\\
&\le \lambda\sqrt{A}\|Q-Q'\|_\infty.
\end{aligned} \tag{36}$$

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

$$\|\pi_Q^\lambda(\cdot|s) - \pi_{Q'}^\lambda(\cdot|s)\|_2 = \sqrt{\sum_{a\in\mathcal{A}} \left(\pi_Q^\lambda(a|s) - \pi_{Q'}^\lambda(a|s)\right)^2} \geq \max_{a\in\mathcal{A}} \left|\pi_Q^\lambda(a|s) - \pi_{Q'}^\lambda(a|s)\right| \tag{37}$$

Recalling (35),

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

where the last inequality is obtained using (36) and (37). The second inequality can be proved using Lemma 7 in [17]. Let us denote the policy under $\text{argmax}(Q)$ as π_Q where the probability is spread evenly among all maximizing actions $a \in \mathcal{A}$ for any given state. Recalling Lemma 7 from [17] for any $s \in \mathcal{S}$,

$$\|\pi_Q^\lambda(\cdot|s) - \pi_Q(\cdot|s)\|_2 \leq 2A\exp(-\lambda\bar{\Delta}) \implies \max_{a\in\mathcal{A}} |\pi_Q^\lambda(a|s) - \pi_Q(a|s)| \leq 2A\exp(-\lambda\bar{\Delta}) \tag{38}$$

due to the equivalence between 2 and ∞ norms.

$$\begin{aligned}
\|\text{softmax}_{\lambda^{k+1}}(\cdot, Q) - \text{argmax}(Q)\|_{TV} &= \|\pi_Q^\lambda - \pi_Q\|_{TV}, \\
&= \max_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \left|\pi_Q^\lambda(a|s) - \pi_Q(a|s)\right|, \\
&\leq \sum_{a\in\mathcal{A}} \max_{s\in\mathcal{S}} \left|\pi_Q^\lambda(a|s) - \pi_Q(a|s)\right|, \\
&\leq 2SA\exp(-\lambda\bar{\Delta})
\end{aligned}$$

where the last inequality is due to (38), thus concluding the proof. □

Now we can further simplify (34) as:

$$\begin{aligned}
e_\pi^{k+1} &\leq (1 - c_\pi^{k+1})e_\pi^k + \|\Delta_\pi^k\|_{TV} + c_\pi^{k+1}\lambda^{k+1}S\sqrt{A}\|Q^k - Q_{\mu^k}^*\|_\infty \\
&\qquad + 2c_\pi^{k+1}SA\exp(-\lambda^{k+1}\bar{\Delta}) + d_1\|\mu^{k+1} - \mu^k\|_1, \\
&\leq (1 - c_\pi^{k+1})e_\pi^k + \|\Delta_\pi^k\|_{TV} + c_\pi^{k+1}\lambda^{k+1}S\sqrt{A}\epsilon_Q^k \\
&\qquad + 2c_\pi^{k+1}SA\exp(-\lambda^{k+1}\bar{\Delta}) + c_\mu^{k+1}d_1(2 + \|\Delta_\mu^k\|_1) + \|\Delta_\mu^k\|_1.
\end{aligned} \tag{39}$$

The first inequality is due to Lemma 10 and the second inequality is due to (33) 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

$$\|\Delta_\pi^k\|_{TV} \leq c_\pi^k \sum_{t=2}^{T-1} t^{-\zeta} \leq c_\pi^k \frac{2^{1-\zeta}}{\zeta - 1}, \qquad \|\Delta_\mu^k\|_1 \leq c_\mu^k \sum_{t=2}^{T-1} t^{-\zeta} \leq c_\mu^k \frac{2^{1-\zeta}}{\zeta - 1} \tag{40}$$

Rearranging the inequality (39),

$$\begin{aligned}
e_\pi^k &\leq \frac{1}{c_\pi^{k+1}}(e_\pi^k - e_\pi^{k+1}) + \frac{2^{1-\zeta}}{\zeta - 1} + \lambda^{k+1}S\sqrt{A}\epsilon_Q^k + 2SA\exp(-\lambda^{k+1}\bar{\Delta}) \\
&\qquad + \frac{c_\mu^{k+1}}{c_\pi^{k+1}} d_1\left(2 + \frac{2^{1-\zeta}}{\zeta - 1}\right) + \frac{c_\mu^k}{c_\pi^{k+1}}\frac{2^{1-\zeta}}{\zeta - 1}, \\
&\leq \frac{e_\pi^k - e_\pi^{k+1}}{c_\pi^{k+1}} + 10 \cdot 2^{1-\zeta} + \lambda^{k+1}S\sqrt{A}\epsilon_Q^k + 2SA\exp(-\lambda^{k+1}\bar{\Delta}) + 12\frac{c_\mu^{k+1}}{c_\pi^{k+1}}d_1 + 10\frac{c_\mu^k}{c_\pi^{k+1}},
\end{aligned}$$

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

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K-1} e_\pi^k \\
&\leq \frac{1}{K}\sum_{k=1}^{K-1}\left(\frac{e_\pi^k - e_\pi^{k+1}}{c_\pi^{k+1}} + \lambda^{k+1}S\sqrt{A}\epsilon_Q^k + 2SA\exp(-\lambda^{k+1}\bar{\Delta}) + 12\frac{c_\mu^{k+1}}{c_\pi^{k+1}}d_1 + 10\frac{c_\mu^k}{c_\pi^{k+1}}\right) \\
&\qquad\qquad + 10\cdot 2^{1-\zeta}, \\
&\leq \frac{1}{K}\sum_{k=2}^{K-1}\Big(\frac{1}{c_\pi^{k+1}} - \frac{1}{c_\pi^k}\Big)e_\pi^{k+1} + \frac{1}{K}\sum_{k=1}^{K-1}\left(\lambda^{k+1}S\sqrt{A}\epsilon_Q^k + 2SA\exp(-\lambda^{k+1}\bar{\Delta})\right. \\
&\qquad\qquad \left. + (12d_1+20)\frac{c_\mu^{k+1}}{c_\pi^{k+1}}\right) + \frac{1}{c_\pi^2 K}e_\pi^1 - \frac{1}{c_\pi^{K+1}K}e_\pi^K + 10\cdot 2^{1-\zeta}, \\
&\leq \frac{2}{K}\sum_{k=2}^{K-1}\Big(\frac{1}{c_\pi^{k+1}} - \frac{1}{c_\pi^k}\Big) + \frac{1}{K}\sum_{k=1}^{K-1}\left(\lambda^{k+1}S\sqrt{A}\epsilon_Q^k + 2SA\exp(-\lambda^{k+1}\bar{\Delta})\right. \\
&\qquad\qquad \left. + (24d_1+40)\frac{c_\mu}{c_\pi}k^{\theta-\gamma}\right) + \frac{2}{c_\pi^2 K} + 10\cdot 2^{1-\zeta}, \\
&\leq \frac{2}{Kc_\pi^K} + S\sqrt{A}\log(K)\epsilon_Q^k + 2SAK^{-\bar{\Delta}} + \frac{(24d_1+40)c_\mu}{(1+\theta-\gamma)c_\pi}K^{\theta-\gamma} + \frac{2}{c_\pi^2 K} + 10\cdot 2^{1-\zeta}, \qquad (41)
\end{aligned}$$

where the second to last inequality is due to the fact that $e_\pi^k \leq 2$ and the last inequaliy is due to the fact that λ^k increases logarithmically with k. Since $\epsilon_Q^k \leq \epsilon_Q/\log(K)$, where $\epsilon_Q > 0$, then

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K-1} e_\pi^k &\leq \frac{2}{Kc_\pi^K} + S\sqrt{A}\epsilon_Q + 2SAK^{-\bar{\Delta}} + \frac{(24d_1+40)\bar{\mu}c_\mu}{(1+\theta-\gamma)c_\pi}K^{\theta-\gamma} + \frac{2}{c_\pi^2 K} + 10\cdot 2^{1-\zeta}, \\
&\leq \mathcal{O}(K^{\theta-1}) + \mathcal{O}(\epsilon_Q) + \mathcal{O}(K^{-\bar{\Delta}}) + \mathcal{O}(K^{\theta-\gamma}) + \mathcal{O}(K^{-1}) + \mathcal{O}(2^{1-\zeta}) \qquad (42)
\end{aligned}$$

where $\bar{\Delta}$ is the uniform action gap and 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,

$$\begin{aligned}
e_\mu^{k+1} &= \|\mu^{k+1} - \mu^*\|_1 = \|\mathbb{P}_{S(\epsilon^{\text{net}})}[\hat{\mu}^{k+1}, 1] - \Gamma_2(\Gamma_1(\mu^*), \mu^*)\|_1 \\
&\leq \|\mathbb{P}_{S(\epsilon^{\text{net}})}[\hat{\mu}^{k+1}, 1] - \hat{\mu}^{k+1}\|_1 + \|\hat{\mu}^{k+1} - \Gamma_2(\Gamma_1(\mu^*), \mu^*)\|_1, \\
&\leq (1 - c_\mu^{k+1})\|\mu^k - \mu^*\|_1 + (1 - c_\mu^{k+1})\|\Delta_\mu^k\|_1 + c_\mu^{k+1}\big[\|(\hat{P}^k)^\top(\mu^k + \Delta_\mu^k) - \Gamma_2(\pi^k, \mu^k)\|_1 \\
&\qquad + \|\Gamma_2(\pi^k, \mu^k) - \Gamma_2(\Gamma_1(\mu^*), \mu^*)\|_1\big] + \epsilon^{\text{net}}, \\
&\leq (1 - c_\mu^{k+1})e_\mu^k + \|\Delta_\mu^k\|_1 + c_\mu^{k+1}\big[\|(\hat{P}^k)^\top(\mu^k + \Delta_\mu^k) - \Gamma_2(\pi^k, \mu^k)\|_1 \\
&\qquad + \|\Gamma_2(\pi^k, \mu^k) - \Gamma_2(\Gamma_1(\mu^k), \mu^k)\|_1 + \|\Gamma_2(\Gamma_1(\mu^k), \mu^k) - \Gamma_2(\Gamma_1(\mu^*), \mu^*)\|_1\big] + \epsilon^{\text{net}}, \\
&\leq (1 - c_\mu^{k+1})e_\mu^k + \|\Delta_\mu^k\|_1 + c_\mu^{k+1}\big[\|(\hat{P}^k)^\top(\mu^k + \Delta_\mu^k) - P_{\pi^k,\mu^k}^\top\mu^k\|_1 + d_2\|\pi^k - \Gamma_1(\mu^k)\|_1 \\
&\qquad + (d_1 d_2 + d_3)\|\mu^k - \mu^*\|_1\big] + \epsilon^{\text{net}}, \\
e_\mu^{k+1} &\leq (1 - c_\mu^{k+1}\bar{d})e_\mu^k + (1 + c_\mu^k)\|\Delta_\mu^k\|_1 + c_\mu^{k+1}\|(\hat{P}^k)^\top - P_{\pi^k,\mu^k}^\top\|_1 + c_\mu^{k+1}d_2 e_\pi^k + \epsilon^{\text{net}}, \\
&\leq (1 - c_\mu^{k+1}\bar{d})e_\mu^k + (1 + c_\mu^k)\|\Delta_\mu^k\|_1 + c_\mu^{k+1}\sqrt{S}\|\hat{P}^k - P_{\pi^k,\mu^k}\|_F + c_\mu^{k+1}d_2 e_\pi^k + \epsilon^{\text{net}}, \\
&\leq (1 - c_\mu^{k+1}\bar{d})e_\mu^k + 11c_\mu^k 2^{1-\zeta} + c_\mu^{k+1}\sqrt{S}\epsilon_P^k + c_\mu^{k+1}d_2 e_\pi^k + \epsilon^{\text{net}}
\end{aligned}$$

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

$$\begin{aligned}
e_\mu^k &\leq \frac{1}{c_\mu^{k+1}\bar{d}}(e_\mu^k - e_\mu^{k+1}) + 11\frac{c_\mu^k}{c_\mu^{k+1}\bar{d}}2^{1-\zeta} + \frac{\sqrt{S}\epsilon_P^k}{\bar{d}} + \frac{d_2 e_\pi^k}{\bar{d}} + \frac{\epsilon^{\text{net}}}{c_\mu^{k+1}\bar{d}}, \\
&\leq \frac{1}{c_\mu^{k+1}\bar{d}}(e_\mu^k - e_\mu^{k+1}) + 22\frac{2^{1-\zeta}}{\bar{d}} + \frac{\sqrt{S}\epsilon_P^k}{\bar{d}} + \frac{d_2 e_\pi^k}{\bar{d}} + \frac{\epsilon^{\text{net}}}{c_\mu^{k+1}\bar{d}}
\end{aligned}$$

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

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

where the second inequality is obtained using steps similar to (41) and the fact that $\epsilon_P^k \le \epsilon_P$. The last inequality is obtained using (42) and the fact that $\epsilon^{\text{net}} \le c_\mu \bar{d}\epsilon/K^\gamma$. The proof is thus concluded. □

A.4 Proof of Corollary 1

Proof. This is a corollary to Theorem 1:

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

where the last inequality follows from Theorem 1. □