Policy Mirror Ascent for Efficient and Independent Learning in Mean Field Games

Batuhan Yardim [1] **Semih Cayci** [2] **Matthieu Geist** [3] **Niao He** [1]

Abstract

Mean-field games have been used as a theoretical tool to obtain an approximate Nash equilibrium for symmetric and anonymous N-player games. However, limiting applicability, existing theoretical results assume variations of a "population generative model", which allows arbitrary modifications of the population distribution by the learning algorithm. Moreover, learning algorithms typically work on abstract simulators with population instead of the N-player game. Instead, we show that N agents running policy mirror ascent converge to the Nash equilibrium of the regularized game within $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ samples from a single sample trajectory without a population generative model, up to a standard $\mathcal{O}(\frac{1}{\sqrt{N}})$ error due to the mean field. Taking a divergent approach from the literature, instead of working with the best-response map we first show that a policy mirror ascent map can be used to construct a contractive operator having the Nash equilibrium as its fixed point. We analyze single-path TD learning for N-agent games, proving sample complexity guarantees by only using a sample path from the N-agent simulator without a population generative model. Furthermore, we demonstrate that our methodology allows for independent learning by N agents with finite sample guarantees.

1. Introduction

Multiagent reinforcement learning (MARL) is a fundamentally challenging problem, despite this with a wide spectrum of applications, for instance in finance (Shavandi & Khedmati, 2022), civil engineering (Wiering, 2000), multi-player games (Samvelyan et al., 2019), energy markets (Rashedi et al., 2016), robotic control (Matignon et al., 2007), and cloud resource management (Mao et al., 2022).

The mean field game (MFG) framework, first proposed by Lasry & Lions (2007) and Huang et al. (2006), is a useful theoretical tool for analyzing a specific class of MARL problems for the N-player case when N is large. As its main insight, MFG analyzes the limiting game when $N \to \infty$, under the condition that the rewards and transition dynamics of the game are *symmetric* for each agent and depend only on the distribution of agents' states (i.e., the agents are *anonymous*). Conceptually, MFG formulates a representative agent playing against a *distribution* of agents. This abstract framework makes it tractable to theoretically characterize an approximate Nash equilibrium for the N-agent MARL, coinciding with the true Nash equilibrium as N grows (Anahtarci et al., 2022; Saldi et al., 2018). Moreover, one can efficiently learn such "MFG Nash equilibria" from repeated plays (for instance, see Perrin et al., 2020). The MFG formalism is useful in analyzing for instance financial systems, auctions, and city planning.

While sharing similar ideas in principle, there have been various mathematical formalizations of MFGs. In the finite horizon case, the Lasry-Lions conditions have been analyzed to obtain Nash equilibria with time-dependent policies (Perrin et al., 2020; 2021). In the infinite horizon setting, the time-dependent evolution of the population becomes a challenge. One can either consider policies depending on the population distribution (Yang et al., 2018; Carmona et al., 2019), or consider Nash equilibria induced by stationary population distributions (Anahtarci et al., 2022; Xie et al., 2021; Zaman et al., 2022). This paper studies this "stationary MFG" problem, formalized in the following section. The results of our work are juxtaposed with existing theory in Table 1.

In parallel to MFG literature, there exists a plethora of results regarding policy gradient (PG) methods in single agent RL (Agarwal et al., 2021; Lan, 2022; Cen et al., 2022; Mei et al., 2020; Li et al., 2022; Tomar et al., 2021). Such methods are typically on-policy algorithms, as opposed to Q-learning which has been employed in MFG literature (for instance in Anahtarci et al., 2022; Zaman et al.,

[1] Department of Computer Science, ETH Zürich, Zürich, Switzerland [2] Department of Mathematics, RWTH Aachen University, Aachen, Germany [3] Google Research, Brain team. Correspondence to: Batuhan Yardim <alibatuhan.yardim@inf.ethz.ch>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

2022). One recent result in single-agent RL suggests that using a mirror descent style operator can achieve $\mathcal{O}(\varepsilon^{-1})$ sample complexity in regularized MDPs (Lan, 2022). The algorithm builds on conditional TD learning (Kotsalis et al., 2022), which establishes that the value estimation problem can be solved with samples from a single trajectory from the Markov chain. In the context of linear quadratic MFGs (LQ-MFG), policy gradient methods with entropy regularization (Guo et al., 2022b) as well as actor-critic methods Fu et al. (2019) have been analyzed. For general MFGs, policy-based methods combined with TD learning have been also analyzed as a method for approximating the optimal Q-value (Guo et al., 2022a). A similar mirror descent operator to ours was considered for finite horizon MFGs with continuous time analysis in the exact case by Pérolat et al. (2022), with a heuristic extension to deep learning by Laurière et al. (2022). We use insights from MFG and PG theory to obtain a $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ time step complexity in the infinite horizon, stationary MFG setting.

Moreover, existing methods for solving stationary MFGs have strong oracle assumptions regarding the population distribution. For instance, Anahtarci et al. (2022) assume a generative model oracle that can produce samples from the game dynamics for any population distribution. Similarly, the N-player weak simulator oracle proposed by Guo et al. (2022a) can produce samples of state transitions and rewards for *any* population distribution. The oracle-free Sandbox Learning algorithm by Zaman et al. (2022) (as well as the actor-critic method of Mao et al. (2022)) uses a two-timescale update of the population distribution based on model-based estimates of the state dynamics. However, this result still requires that the population distribution at each time step can be manipulated by the algorithm. In real-world problems, however, the N-player simulator might not allow the algorithm to modify or control the population of agents explicitly. For example, when learning to control a transportation infrastructure with a large number of drivers, it might be challenging and costly to force the drivers into a particular configuration before simulating the system. Or it might be impossible to run simulations starting from arbitrary population configurations, as is typical in strategic video games with unknown/complicated dynamics. One of our main goals is to propose an algorithm that interacts with the simulator only by controlling the *policies* of the agents, as done in the single-agent RL literature.

Finally, in this paper, we tackle the question of independent learning. Existing algorithms for MFGs rely on a centralized controller running simulations and coordinating the policies of agents. However, we show that our policy mirror ascent based algorithm can be extended to an independent learning algorithm run by N agents without additional knowledge of the population and environment, only observing their own actions, state transitions, and rewards. Independent learning in N-player mean-field games have been studied by Yongacoglu et al. (2022), although they analyze asymptotic convergence to subjective (rather than objective) equilibria without finite sample bounds. A similar idea of independent learning has been analyzed in the two agent setting by Daskalakis et al. (2020) and Sayin et al. (2021). In the more restrictive context of localized learning on graphs, Gu et al. (2021) analyze independent learning where agent states are only affected by neighboring nodes on an undirected graph. Independent learning in the case of large-scale Markov potential games has been analyzed by Ding et al. (2022). For two-player zero-sum games, independent learning algorithms have been reviewed under various settings by Ozdaglar et al. (2021).

To summarize, our contributions are the following ones.

Policy mirror ascent operators. To the best of our knowledge, practically all results in the *stationary* MFG literature with finite state-action spaces rely on the conditions developed by Anahtarci et al. (2022) or Cui & Koeppl (2021) for the best response map (from the set of populations to the set of policies) to be Lipschitz continuous or take it as a blanket assumption, yielding an operator contracting to the Nash equilibrium. We take a different approach and *prove* that policy mirror ascent (PMA) can yield a contraction under comparable conditions on the strong concavity of the regularizer. This difference in the operators allows us to develop an algorithm that does not need to compute the best response at each step.

No population manipulation. In this work, we only assume that we can simulate a single episode of N agents playing a symmetric and anonymous game. The proposed algorithm can only interact with the environment by manipulating the policies of agents and can not directly manipulate the states of each agent. This is in contrast with past work where the empirical state distribution of agents can be arbitrarily set (Anahtarci et al., 2022; Guo et al., 2019), mixed (Zaman et al., 2022; Mao et al., 2022) or projected (Zaman et al., 2022; Guo et al., 2019).

TD learning with population. We show that by extending the single-agent conditional TD learning results of Kotsalis et al. (2022), one can efficiently perform TD learning in N-player games without a generative model. In the absence of a population oracle, TD learning is conducted by simulating N agents, introducing several complexities including a population bias and non-homogeneous dynamics due to the evolving population. We establish that TD learning can be performed with single path simulations with $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ samples, up to a standard $\mathcal{O}(\frac{1}{\sqrt{N}})$ error.

Sample efficiency. Our algorithms differ from several past works in that the best response does not need to be recomputed for each population distribution (unlike for instance

	No population manipulation	Single path	N-agent simulator	Independent learning
(Guo et al., 2019)	No	No	No	No
(Anahtarci et al., 2022)	No	No	No	No
(Subramanian & Mahajan, 2019)	No	No	No	No
(Xie et al., 2021)	No	Yes	No	No
(Zaman et al., 2022)	No	Yes	No	No
This work-1	**Yes**	**Yes**	**Yes**	**No**
This work-2	**Yes**	**Yes**	**Yes**	**Yes**

Table 1. Theoretical results in the literature for computing stationary MFG-NE in discrete state-action spaces and their requirements.

(Guo et al., 2019; Anahtarci et al., 2022)), and only a value function estimation is necessary for policy mirror ascent at each iteration. This approach yields a time step complexity of $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ as opposed to for example $\mathcal{O}(\varepsilon^{-4})$ in (Zaman et al., 2022) and $\mathcal{O}(\varepsilon^{-4|\mathcal{A}|})$ in (Anahtarci et al., 2022).

Independent learning. A fundamental question in competitive multi-agent learning is whether independent learning can be achieved as opposed to learning in the presence of a centralized controller. This question is especially significant for mean-field games, where the added complexity of having a very large number of agents might not allow centralized learning in practice. To the best of our knowledge, we establish the first MFG algorithm for independent learning with N agents with finite sample bounds.

2. Mean Field Game Formalization

Firstly, we introduce the (stationary) MFG problem. We assume $\mathcal{S}$ is a finite state space and $\mathcal{A}$ is a finite action space. We denote the set of probability measures on a finite set $\mathcal{X}$ by $\Delta_{\mathcal{X}}$. We denote the set of policies as $\Pi := \{\pi : \mathcal{S} \to \Delta_{\mathcal{A}}\}$. Let $h : \Delta_{\mathcal{A}} \to \mathbb{R}_{\geq 0}$ be a given function.

We formally define the symmetric anonymous game with states (SAGS) with N players, which is the main object of interest of this work. In SAGS, the state and reward dynamics depend only on the empirical distribution of states among agents (hence anonymity) and are the same for each agent (hence the symmetry).

Definition 2.1 (Symmetric anonymous games)**.** A N-player symmetric anonymous game is a tuple $(N, \mathcal{S}, \mathcal{A}, P, R, \gamma)$ for which any initial states of players $\{s_0^i\}_{i=1}^N \in \mathcal{S}^N$ and policies $\{\pi^i\}_{i=1}^N \in \Pi^N$ induce a random sequence of states and rewards $\{s_t^i\}_{i,t}, \{r_t^i\}_{i,t}$, so that at each timestep $t \geq 0, \forall i = 1, \ldots, N$,

$$a_t^i \sim \pi^i(s_t^i),\ r_t^i = R(s_t^i, a_t^i, \widehat{\mu}_t),\ s_{t+1}^i \sim P(\cdot|s_t^i, a_t^i, \widehat{\mu}_t),$$

where $P : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}} \to \Delta_{\mathcal{S}}$ and $R : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}} \to [0, 1]$ map state, action and population distribution tuples to transition probabilities and (bounded) rewards, and $\widehat{\mu}_t = \frac{1}{N}\sum_{i=1}^N \sum_{s \in \mathcal{S}} \mathbb{1}_{s_t^i = s} \mathbf{e}_s \in \Delta_{\mathcal{S}}$ is the $\mathcal{F}(\{s_t^i\}_i)$-measurable random vector of the empirical state distribution at time t of the agents.

With SAGS dynamics induced by sampling initial states from an initial distribution $\mu_0 \in \Delta_{\mathcal{S}}$ and policies $\boldsymbol{\pi} := (\pi^1, \ldots, \pi^N) \in \Pi^N$, we can formalize the expected discounted returns of each agent.

Definition 2.2 (N-player discounted reward)**.** For the N-player SAGS $(N, \mathcal{S}, \mathcal{A}, P, R, \gamma)$, we define the (regularized) discounted return of player i for initial state distribution $\mu_0 \in \Delta_{\mathcal{S}}$ and Markov policies $\boldsymbol{\pi} \in \Pi^N$ as $J_h^i(\boldsymbol{\pi}, \mu_0) := \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \left(R(s_t^i, a_t^i, \widehat{\mu}_t) + h(\pi^i(s_t^i))\right) | s_0^j \sim \mu_0, a_t^j \sim \pi^j(s_t^j), s_{t+1}^j \sim P(\cdot|s_t^j, a_t^j, \widehat{\mu}_t), \forall t \geq 0, j \in 1, \ldots, N\right]$.

Based on the above definition, we introduce the notion of a Nash equilibrium (NE) to the N-player SAGS. The NE is the natural solution concept for an N-player competitive game, therefore our main objective will be to compute an approximate NE for the SAGS.

Definition 2.3 (δ-Nash equilibrium)**.** For $\delta \geq 0$, an N-tuple of policies $\boldsymbol{\pi} = (\pi^1, \ldots, \pi^N) \in \Pi^N$ and initial distribution $\mu_0 \in \Delta_{\mathcal{S}}$ constitute a δ-Nash equilibrium $(\boldsymbol{\pi}, \mu_0)$ if for all $i = 1, \ldots, N$ we have $J_h^i(\boldsymbol{\pi}, \mu_0) \geq \max_{\pi \in \Pi} J_h^i\left((\pi, \boldsymbol{\pi}^{-i}), \mu_0\right) - \delta$, where $(\pi, \boldsymbol{\pi}^{-i}) := (\pi^1, \ldots, \pi^{i-1}, \pi, \pi^{i+1}, \ldots \pi^N) \in \Pi^N$.

As a useful theoretical analysis tool one can approximate the N player game with the infinite agent limit ($N \to \infty$). This is the main insight of MFG: At the limit, one observes a single representative agent's policy playing against a population of infinitely many infinitesimal opponents characterized with a fixed distribution $\mu \in \Delta_{\mathcal{S}}$. Observing the limit $N \to \infty$, one can define the discounted expected reward with respect to the single-agent MDP parameterized (or induced) by the population. We next define the discounted expected reward of a representative agent against such a population. We use the letter V to distinguish the expected "infinite-agent" reward from the expected rewards of the N-agent SAGS, denoted with the letter J.

Definition 2.4 (Mean-field discounted reward)**.** We define the expected mean-field reward for a population-policy pair $(\pi, \mu) \in \Pi \times \Delta_{\mathcal{S}}$ as $V_h(\pi, \mu) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t, \mu) + h(\pi(s_t))\right) | s_0 \sim \mu, a_t \sim \pi(s_t), s_{t+1} \sim P(\cdot|s_t, a_t, \mu)\right]$.

A Nash equilibrium concept can be also introduced at the MFG limit. The so-called stationary MFG-NE is formulated with an optimality condition and a stability condition on π, μ. Intuitively, the stability condition ensures that the population distribution remains consistent, removing the need to consider time-varying distributions (see Guo et al., 2019).

Definition 2.5 (MFG-NE)**.** A policy $\pi^* \in \Pi$ and population distribution $\mu^* \in \Delta_{\mathcal{S}}$ pair (π^*, μ^*) is called an MFG-NE if the following conditions are satisfied:

$$\text{Stability: } \mu^*(s) = \sum_{s', a'} \mu^*(s') \pi^*(a'|s') P(s|s', a', \mu^*),$$
$$\text{Optimality: } V_h(\pi^*, \mu^*) = \max_{\pi} V_h(\pi, \mu^*).$$

If the optimality condition is replaced with $V_h(\pi^*_\delta, \mu^*_\delta) \geq \max_\pi V_h(\pi, \mu^*_\delta) - \delta$, we call $(\pi^*_\delta, \mu^*_\delta)$ a δ-MFG-NE.

While the optimality condition above corresponds to the objective of single-agent RL, incorporating a mean field game dynamic in Markov decision processes yields an *evolving* MDP depending on the population distribution (in other words, induces an infinite family of MDPs to be solved). Conceptually this challenge in MFGs mirrors that in the case of multi-agent MDPs, where each agent plays against an evolving environment. We also comment on the effect of regularization (due to h) on the MFG-NE in the appendix; see Section C.9.

The main motivation to study the abstract MFG-NE concept is that it corresponds to an approximate Nash equilibrium of the N-player SAGS. We re-iterate this well-known result.

Proposition 2.6 (MFG-NE and NE (Theorem 1 of Anahtarci et al. (2022)))**.** *Assume that the pair (π^*, μ^*) is an MFG-NE. Under technical conditions, for each $\delta > 0$, there exists an $N = N(\delta) \in \mathbb{N}_{>0}$ such that (π^*, μ^*) is a δ-Nash equilibrium for the N-player SAGS.*

In fact, it can be shown by standard techniques (used also in our paper) that (π^*, μ^*) is a $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$-Nash equilibrium for the N-player SAGS. With this reduction, the main objective of this paper will be to learn the MFG-NE. Our goal is to expand on (and relax) this existential result and show that a similar bound on bias can be shown when also *learning* completely occurs in the finite agent setting, removing the abstract MFG formalism completely from the algorithm.

3. Operators and the Exact PMA Case

In this section, we show that finding the MFG-NE in the infinite agent limit can be formulated as a fixed point iteration of an appropriately defined policy mirror ascent operator. The results of this section will ultimately show convergence in the so-called "exact" setting, where value functions can be computed exactly. These results will also be instrumental in establishing convergence later in the N-agent stochastic case. We state the operators and major ideas, postponing the proofs to the appendices. Theorems with omitted constants are restated in the appendix. This section can be best compared to the work of Anahtarci et al. (2022). We also compare in greater detail existing assumptions in the literature to establish a contraction to the MFG-NE in the appendix (Section C.1).

Definitions. We equip sets $\mathcal{S}, \mathcal{A}$ with the discrete metric $d(x, y) = \mathbb{1}_{x \neq y}$. We denote the set of state-action value functions as $\mathcal{Q} := \{q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$. We equip both of the spaces $\Delta_{\mathcal{S}} \subset \mathbb{R}^{\mathcal{S}}$ and $\Delta_{\mathcal{A}} \subset \mathbb{R}^{\mathcal{A}}$ with the norm $\|\cdot\|_1$. For $\pi, \pi' \in \Pi, q, q' \in \mathcal{Q}$ we use the norms $\|\pi - \pi'\|_1 := \sup_{s \in \mathcal{S}} \|\pi(s) - \pi'(s)\|_1, \|q - q'\|_\infty := \sup_{s \in \mathcal{S}, a \in \mathcal{A}} |q(s, a) - q'(s, a)|, \|q - q'\|_2 := \sqrt{\sum_{s,a} |q(s, a) - q'(s, a)|^2}$. Finally, we assume that $h : \Delta_{\mathcal{A}} \to \mathbb{R}_{\geq 0}$ is a ρ-strongly concave function on $\Delta_{\mathcal{A}}$ with respect to norm $\|\cdot\|_1$ (see Definition B.4 in the appendix). We define $u_{max} := \arg\max_{u \in \Delta_{\mathcal{A}}} h(u)$, $h_{max} := h(u_{max})$, $\pi_{max} \in \Pi$ such that $\pi_{max}(s) = u_{max}$ for all $s \in \mathcal{S}$ and $Q_{max} := \frac{1 + h_{max}}{1 - \gamma}$. We further assume h is continuously differentiable for simplicity, although the results yield a straightforward generalization to the case h is not everywhere differentiable. Finally, for any $\Delta h \in \mathbb{R}_{>0}$ we define the convex sets

$$\mathcal{U}_{\Delta h} := \{u \in \Delta_{\mathcal{A}} : h(u) \geq h_{max} - \Delta h\}, \tag{1}$$
$$\Pi_{\Delta h} := \{\pi \in \Pi : \pi(s) \in \mathcal{U}_{\Delta h}, \forall s \in \mathcal{S}\}. \tag{2}$$

As standard in previous work, we require the following smoothness assumptions on P, R.

Assumption 1 (Lipschitz continuity of P, R)**.** There exists constants $K_\mu, K_s, K_a, L_\mu, L_s, L_a \in \mathbb{R}_{\geq 0}$ such that $\forall s, s' \in \mathcal{S}, a, a' \in \mathcal{A}, \mu, \mu' \in \Delta_{\mathcal{S}}$,

$$\begin{aligned}\|P(\cdot|s, a, \mu) - P(\cdot|s', a', \mu')\|_1 \leq & K_\mu \|\mu - \mu'\|_1 + K_s d(s, s') \\ & + K_a d(a, a'), \\ |R(s, a, \mu) - R(s', a', \mu')| \leq & L_\mu \|\mu - \mu'\|_1 + L_s d(s, s') \\ & + L_a d(a, a').\end{aligned}$$

Without loss of generality, we can assume $K_s, K_a \leq 2$ and $L_s, L_a \leq 1$ since it holds that $\|P(\cdot|s, a, \mu) - P(\cdot|s', a', \mu')\|_1 \leq 2$ and $|r(s, a, \mu) - r(s', a', \mu')| \leq 1$.

3.1. Population Update Operators

One critical goal in the MFG framework is understanding the evolution of the population. We define an operator to characterize the single-step change of the population.

Definition 3.1 (Population update)**.** The population update operator $\Gamma_{pop} : \Delta_{\mathcal{S}} \times \Pi \to \Delta_{\mathcal{S}}$ is defined as

$$\Gamma_{pop}(\mu, \pi)(s) := \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} \mu(s')\pi(a'|s')P(s|s', a', \mu),$$

for all $s \in \mathcal{S}$. We also introduce the shorthand notation $\Gamma^n_{pop}(\mu, \pi) := \underbrace{\Gamma_{pop}(\ldots \Gamma_{pop}(\Gamma_{pop}(\mu, \pi), \pi), \ldots \pi)}_{n \text{ times}}$.

Re-iterating known results (for instance from Anahtarci et al. (2022) or Guo et al. (2019)) in our notation and definition of constants, we can show that Γ_{pop} is Lipschitz.

Lemma 3.2 (Lipschitz population updates)**.** *The population update operator Γ_{pop} is Lipschitz with $\|\Gamma_{pop}(\mu, \pi) - \Gamma_{pop}(\mu', \pi')\|_1 \leq L_{pop,\mu}\|\mu - \mu'\|_1 + \frac{K_a}{2}\|\pi - \pi'\|_1$, where $L_{pop,\mu} := (\frac{K_s}{2} + \frac{K_a}{2} + K_\mu)$, for all $\pi \in \Pi, \mu \in \Delta_{\mathcal{S}}$.*

In the stationary MFG framework, we would like to compute a unique population distribution that is stable with respect to a policy. This requires $\Gamma_{pop}(\cdot, \pi)$ to be contractive for all π. Hence, we state our second assumption, also implied by assumptions in past work (see Anahtarci et al. (2022); Zaman et al. (2022); Guo et al. (2019), also Section C.1).

Assumption 2 (Stable population)**.** Population updates are stable, i.e., $L_{pop,\mu} < 1$.

We can now formalize the operator that maps policies to their stable distributions, which is well-defined under Assumption 2.

Definition 3.3 (Stable population operator Γ^∞_{pop})**.** Under Assumption 2, the stable population operator $\Gamma^\infty_{pop} : \Pi \to \Delta_{\mathcal{S}}$ is defined as the unique population distribution such that $\Gamma_{pop}(\Gamma^\infty_{pop}(\pi), \pi) = \Gamma^\infty_{pop}(\pi)$, that is, the (unique) fixed point of $\Gamma_{pop}(\cdot, \pi) : \Delta_{\mathcal{S}} \to \Delta_{\mathcal{S}}$.

It is straightforward to see that $\Gamma^\infty_{pop} = \lim_{n\to\infty} \Gamma^n_{pop}$ and that Γ^∞_{pop} is Lipschitz with constant $L_{pop,\infty} := \frac{K_a}{2(1-L_{pop,\mu})}$, proven in the appendices (Lemma C.10). The operators $\Gamma_{pop}, \Gamma^\infty_{pop}$ are sometimes called population oracles.

3.2. Policy Mirror Ascent Operator

For policy updates, we take a diverging approach from literature and demonstrate that a policy mirror ascent (PMA) operator is Lipschitz, similar to the best response operator. We define the Q_h and q_h functions for each state-action pair as $Q_h(s, a|\pi, \mu) := \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \left(R(s_t, a_t, \mu) + h(\pi(s_t))\right) | s_0 = s, a_0 = a, s_{t+1} \sim P(\cdot|s_t, a_t, \mu) a_{t+1} \sim \pi(\cdot|s_{t+1}), \forall t \geq 0\right]$ and $q_h(s, a|\pi, \mu) := R(s, a, \mu) + \gamma \sum_{s',a'} P(s'|s, a, \mu)\pi(a'|s')Q_h(s', a'|\pi, \mu)$. We also define $V_h(s|\pi, \mu) := \sum_a \pi(a|s)Q_h(s, a|\pi, \mu)$. Note that $Q_h(s, a|\pi, \mu) = q_h(s, a|\pi, \mu) + h(\pi(s))$. With these definitions in place, we analyze the operator that maps population-policy pairs to Q-functions. This operator is crucial for the online learning algorithm, as it can be approximated purely from trajectories of the current policy.

Definition 3.4 (Γ_q operator)**.** We define $\Gamma_q : \Pi \times \Delta_{\mathcal{S}} \to \mathcal{Q}$ as $\Gamma_q(\pi, \mu) = q_h(\cdot, \cdot|\pi, \mu) \in \mathcal{Q}, \quad \forall \pi \in \Pi, \mu \in \Delta_{\mathcal{S}}$.

As expected, the map Γ_q is also Lipschitz continuous. The Lipschitz properties of Γ_q are shown in the appendix, Lemma C.9. Next, we define the policy improvement operator.

Definition 3.5 (Policy mirror ascent operator)**.** Let $\eta > 0$ and $L_h := L_a + \gamma\frac{L_s K_a}{2-\gamma K_s}$. We define the PMA operator $\Gamma^{md}_\eta : \mathcal{Q} \times \Pi \to \Pi$ as $\forall s \in \mathcal{S}, q \in \mathcal{Q}, \pi \in \Pi$,

$$\Gamma^{md}_\eta(q, \pi)(s) = \arg\max_{u \in \mathcal{U}_{L_h}} \langle u, q(s, \cdot)\rangle + h(u) - \frac{1}{2\eta}\|u - \pi(s)\|^2_2.$$

We establish the Lipschitz continuity of Γ^{md}_η as a result of Fenchel duality and the strong monotonicity of the gradient operator of a strongly convex function, the full proof presented in Lemma C.3.

With the building blocks defined above, we now define the main learning operator of interest. For a learning rate $\eta > 0$, we define $\Gamma_\eta : \Pi \to \Pi$ as

$$\Gamma_\eta(\pi) := \Gamma^{md}_\eta\left(\Gamma_q(\pi, \Gamma^\infty_{pop}(\pi)), \pi\right).$$

Intuitively, the operator Γ_η takes a PMA step with respect to the MDP induced by the stationary distribution $\Gamma^\infty_{pop}(\pi)$. This operator will be used in the fixed-point iteration process, to find π such that $\Gamma_\eta(\pi) = \pi$. The main justification for this is the following lemma, which demonstrates that the fixed points of Γ_η are MFG-NE policies for any $\eta > 0$.

Lemma 3.6 (Fixed points of Γ_η are MFG-NE)**.** *Let $\eta > 0$ be arbitrary. A pair (π^*, μ^*) is a MFG-NE if and only if $\pi^* = \Gamma_\eta(\pi^*)$ and $\mu^* = \Gamma^\infty_{pop}(\pi^*)$.*

Finally, we present the Lipschitz continuity result of Γ_η and establish the conditions that make it contractive.

Lemma 3.7 (Lipschitz continuity of Γ_η)**.** *For any $\eta > 0$, the operator $\Gamma_\eta : \Pi \to \Pi$ is Lipschitz with constant L_{Γ_η} on $(\Pi, \|\cdot\|_1)$, where*

$$L_{\Gamma_\eta} := \frac{L_{\Gamma,q}\eta|\mathcal{A}|}{1 + \rho\eta|\mathcal{A}|} + \frac{1}{|\mathcal{A}|^{-1} + \eta\rho} < \frac{L_{\Gamma,q}}{\rho} + \frac{1}{\eta\rho},$$

where $L_{\Gamma,q}$ is a problem-dependent constant.

Proof. See Lemma C.12. □

We point out that $L_{\Gamma,q}$ only depends on $L_\mu, L_a, L_s, K_\mu, K_s, K_a, \gamma$. While L_{Γ_η} depends on the dimensionality of the problem (namely $|\mathcal{A}|$), this dependence will disappear when the learning rate η is large enough since at the limit $\eta \to \infty$ we obtain $L_{\Gamma_\eta} \to \frac{L_{\Gamma,q}}{\rho}$. Moreover, for any SAGS, if we take the regularizer to be $\lambda h(\cdot)$, for sufficiently large $\lambda > 0$ it will always be possible to obtain $L_{\Gamma_\eta} < 1$. We also point out that for the contraction condition $L_{\Gamma_\eta} < 1$ to hold, it must hold that $\rho > L_{\Gamma,q}$. Conversely, whenever $\rho > L_{\Gamma,q}$ holds, a contraction can be obtained by a sufficiently large learning rate, for instance, $\eta > (\rho - L_{\Gamma,q})^{-1}$. These results mirror the conditions for contraction developed for the best response operator in (Anahtarci et al., 2022) without the need of computing the best response at each stage.

It is worth noting that the mirror ascent operator Γ_η^{md} only considers the Bregman divergence induced by the squared ℓ_2 norm. A natural alternative would be the divergence induced by h, leading to closed form solutions for Γ_η^{md} in certain cases (e.g. when h is the entropy regularizer). However, it is difficult to establish Lipschitz continuity for general divergences with respect to π and we postpone this consideration as future work. Finally, as expected, computing Γ_η^{md} in practice will require approximating the solution of a strongly concave maximization problem, which can be solved efficiently with linear convergence. The computational aspects of Γ_η^{md} are discussed further by Lan (2022).

3.3. Learning in the Exact Policy Mirror Ascent Case

Concluding this section, we prove the linear convergence to the MFG-NE policy in the exact case, assuming that value functions are known exactly (i.e., we can compute Γ_q).

Proposition 3.8 (Learning MFG-NE, exact case)**.** *Assume that (π^*, μ^*) is the MFG-NE and $L_{\Gamma_\eta} < 1$ for the learning rate $\eta > 0$. Assume $\pi_0 = \pi_{max}$ and consider the updates $\pi_{t+1} = \Gamma_\eta(\pi_t)$ for all $t \geq 0$. For any $T \geq 1$, we have $\|\pi_T - \pi^*\|_1 \leq L_{\Gamma_\eta}^T \|\pi_0 - \pi^*\|_1 \leq 2L_{\Gamma_\eta}^T$.*

In case we have access to generative-model based samples of the SAGS as in literature (i.e., at any time t we can sample from $P(\cdot|s_t, a_t, \mu)$ and $R(s_t, a_t, \mu)$ for arbitrary μ), Proposition 3.8 readily implies a sample complexity by plugging in a method to estimate value functions and the mean field $\Gamma_{pop}^\infty(\cdot)$.

Remark 3.9. The contraction property of Γ_η requires the MFG to be sufficiently regularized, as in the case for best-response operators (Anahtarci et al., 2022). A natural question is if learning of stationary MFG is possible for unregularized games. While we do not have a universal proof of intractability, we show in the appendices (Section C.8) that a large class of best response operators employed in literature can not be continuous in μ, unless they are trivial (i.e., the same best-response for all μ). This result is similar to (Cui & Koeppl, 2021) with a different characterization.

Remark 3.10. Since the results do not allow $\rho \to 0$, the regularized MFG-NE will have non-vanishing bias when we are interested in the unregularized NE as discussed in Section C.9 (see also (Geist et al., 2019)). The idea of regularization to obtain true NE has been used by (Perolat et al., 2021) to obtain the unbiased NE of a two player imperfect information game by updating a reference policy. We leave it as an open question if unregularized stationary MFG-NE can be computed by a similar scheme, for instance using a Bregman divergence regularizer $-D_{-h}(\pi(s)||\pi_{old}(s))$ where π_{old} is periodically updated.

4. Sample-Based Learning with N Agents

After establishing the deterministic operator to be computed, we now move to the case where we learn from samples/simulations. The main goal is to show that Γ_η can be repeatedly estimated from the simulation of a single path with N agents. The difficulty will be establishing that the map Γ_q can be approximated by taking simulation steps in a single trajectory and that the accumulation of error from past iterations can be controlled.

Definitions. We provide the probabilistic setup for single-path learning with N-agents. Let $\{s_0^i\}_{i=1}^N \subset \mathcal{S}^N$ be arbitrary initial states. At each timestep t, we denote the policy followed by agent i as π_t^i for $i = 1, \ldots, N$, yielding the transitions $a_t^i \sim \pi_t^i(s_t^i), s_{t+1}^i \sim P(\cdot|s_t^i, a_t^i, \widehat{\mu}_t)$, for $i = 1, \ldots, N$, where $\widehat{\mu}_t$ is the $\mathcal{F}(\{s_t^i\}_{i=1}^N)$-measurable empirical state distribution as before. For any $T > 0$, we denote by $\mathcal{F}_T$ the sigma algebra $\mathcal{F}_T := \mathcal{F}(\{s_t^i, a_t^i, r_t^i\}_{i=1,t=0}^{N,T})$. We note that any learning algorithm in our context can only specify π_t^i at time t for any agent i, and only using past observations (i.e., π_t^i must be $\mathcal{F}_{t-1}$-measurable). We define $\mathcal{Z} := \mathcal{S} \times \mathcal{A} \times [0,1] \times \mathcal{S} \times \mathcal{A}$, and set ζ_t^i to be the random transition observations of agent i at time t, given by $\zeta_t^i = (s_t^i, a_t^i, r_t^i, s_{t+1}^i, a_{t+1}^i)$ in set $\mathcal{Z}$.

In general, a mixing assumption is required for online learning on a single trajectory. The mixing condition can be reduced to a persistence of excitation condition coupled with technical conditions on the probability transition function. Rather than a blanket assumption, we formalize this as a "persistence of excitation" and "sufficient mixing" assumption. The first assumption imposes that the policies throughout the entire training process take each action in each state with probability bounded away from zero.

Assumption 3 (Persistence of excitation)**.** Assume there exists $p_{inf} > 0$ such that

1. $\pi_{max}(a|s) \geq p_{inf}$ for any $s \in \mathcal{S}, a \in \mathcal{A}$,
2. For any $\pi \in \Pi, q \in \mathcal{Q}$ satisfying $\pi(a|s) \geq p_{inf}, 0 \leq q(s,a) \leq Q_{max}, \forall(s,a) \in \mathcal{S} \times \mathcal{A}$, it holds that $\Gamma_\eta^{md}(q,\pi)(a|s) \geq p_{inf}, \forall(s,a) \in \mathcal{S} \times \mathcal{A}$.

Assumption 3 is a property of the policy update operator

Γ^{md}_{η} and therefore of the regularizer h. That is, the regularizer h must ensure that all actions at all states are explored with non-zero probability.

Equivalent conditions on h. The persistence of excitation (PE) assumption above is in fact achieved for a large class of strongly concave h, for instance with entropy regularization. Sufficient conditions on the gradient ∇h at the boundary of $\Delta_{\mathcal{A}}$ for PE to hold are characterized in Section D.1.

Assumption 4 (Sufficient mixing)**.** For any $\pi \in \Pi$ satisfying $\pi(a|s) \geq p_{inf} > 0, \forall s \in \mathcal{S}, a \in \mathcal{A}$, and any initial states $\{s^i_0\}_i \in \mathcal{S}^N$, there exists $T_{mix} > 0, \delta_{mix} > 0$ such that $\mathbb{P}(s^j_{T_{mix}} = s' | \{s^i_0\}_i) \geq \delta_{mix}, \quad \forall s' \in \mathcal{S}, j \in [N]$.

Assumptions 3 and 4 are equivalent to Assumption 2 of (Zaman et al., 2022), where a policy is mixed with a uniform policy to achieve PE. However, since we employ a strongly concave regularizer h, we would not typically need to explicitly mix policies, under certain conditions on h. The mixing condition is implied by aperiodicity and irreducability, and have also been explored outside of MFG, for instance by Lan (2022); Tsitsiklis & Van Roy (1996).

4.1. TD Learning Under Population Dynamics

One of the main results of this paper is that a variant of TD learning has a $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ sample complexity in SAGS. Specifically, we extend the results of conditional TD learning (CTD) (Kotsalis et al., 2022), where the learner waits several time steps between TD learning updates to ensure sufficient convergence to the steady state. Following the variational inequality approach of (Kotsalis et al., 2022), we introduce the relevant operators.

Definition 4.1 (Operators for CTD)**.** Let $\pi \in \Pi$ and $\mu_\pi := \Gamma^\infty_{pop}(\pi)$. We define the Bellman operator $T^\pi : \mathcal{Q} \to \mathcal{Q}$ as $(T^\pi Q)(s,a) := R(s,a,\mu_\pi) + h(\pi(s)) + \gamma \sum_{s',a'} P(s' \mid s,a,\mu_\pi)\, \pi(a' \mid s)\, Q(s',a')$ for each $Q \in \mathcal{Q}$. We also define the corresponding TD learning operator as $F^\pi(Q) := \mathbf{M}^\pi (Q - T^\pi Q)$, where $\mathbf{M}^\pi := \operatorname{diag}(\{\mu_\pi(s)\pi(a|s)\}_{s,a}) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ is the state-action distribution matrix induced by π at the limiting mean field distribution. Finally, we define the stochastic TD learning operator $\widetilde{F}^\pi : \mathcal{Q} \times \mathcal{Z} \to \mathcal{Q}$ with

$$\widetilde{F}^\pi(Q,\zeta) = \Big(Q(s,a) - r - h(\pi(s)) - \gamma Q(s',a')\Big)\mathbf{e}_{s,a},$$

for any $Q \in \mathcal{Q}, \zeta := (s,a,r,s',a') \in \mathcal{Z}$.

To keep the notation clean, we omit the dependence of $T^\pi, F^\pi, \widetilde{F}^\pi$ on the regularizer h. Intuitively, Definition 4.1 defines F^π with respect to the *abstract* (i.e., non-observable) MDP induced by the limiting stable distribution $\Gamma^\infty_{pop}(\pi)$; however, at time t we can only observe $\widetilde{F}^\pi\left(Q, \zeta^i_{t-2}\right)$ for $i \in 1,\ldots,N$ from simulations with N agents. The main challenge is establishing that $\widetilde{F}^\pi$ can estimate F^π well after waiting for the population and the (population-dependent) Markov chain to mix between each evaluation of $\widetilde{F}^\pi$. CTD (presented in Algorithm 1) waits several time steps between each TD step to utilize this mixing. Note that the algorithm performs TD learning for the first agent ($i = 1$).

Algorithm 1 Conditional TD learning with population

```
Require: $M$, $M_{td}$, learning rates $\{\beta_m\}_m$, policies $\{\pi^i\}_i$,
  initial states $\{s^i_0\}_i$.
  Set $t \leftarrow 0$ and $\widehat{Q}_0(s,a) = Q_{max}, \forall s \in \mathcal{S}, a \in \mathcal{A}$
  for $m \in 0, 1, \ldots, M-1$ do
    for $M_{td}$ iterations do
      Take simulation step ($\forall i$): $a^i_t \sim \pi^i(s^i_t)$, $r^i_t = R(s^i_t, a^i_t, \widehat{\mu}_t)$, $s^i_{t+1} \sim P(\cdot|s^i_t, a^i_t, \widehat{\mu}_t)$.
      $t \leftarrow t + 1$
    end for
    Set $\zeta^1_{t-2} = (s^1_{t-2}, a^1_{t-2}, r^1_{t-2}, s^1_{t-1}, a^1_{t-1})$.
    CTD Update: $\widehat{Q}_{m+1} = \widehat{Q}_m - \beta_m \widetilde{F}^{\pi^1}\left(\widehat{Q}_m, \zeta^1_{t-2}\right)$.
  end for
  Return $\widehat{Q}_M$.
```

It is known that T^π is contractive with γ and F^π is Lipschitz with $L_F := (1+\gamma)$ with respect to the $\|\cdot\|_2$ norm on $\mathcal{Q}$ for any $\pi \in \Pi$. F^π is also generalized strongly monotone (Kotsalis et al., 2022) with modulus $\mu_F := (1-\gamma)\delta_{mix}p_{inf}$, that is, $\forall Q \in \mathcal{Q}$, it holds that

$$\langle F^\pi(Q), Q - Q_h(\cdot,\cdot|\pi,\mu_\pi)\rangle \geq \mu_F \|Q - Q_h(\cdot,\cdot|\pi,\mu_\pi)\|_2^2,$$

where $Q_h(\cdot,\cdot|\pi,\mu_\pi)$ is the true value function of the policy π at the mean-field $\mu_\pi = \Gamma^\infty_{pop}(\pi)$ as defined in Section 3.

Our analysis consists of two steps: (1) we prove that the CTD algorithm of (Kotsalis et al., 2022) can be used when the samples are biased and state the explicit bound (see Theorem D.2 in the appendices), and (2) quantify the bias and mixing rate to the limiting distribution in N-player SAGS (in Section D.3) to prove the main result, Theorem 4.2.

Theorem 4.2 (CTD learning with population)**.** *Assume Assumption 4 holds and let policies $\{\pi^i\}_i$ be given so that $\pi^i(a|s) \geq p_{inf}$ for all i. Assume Algorithm 1 is run with policies $\{\pi^i\}_i$,* arbitrary *initial agent states $\{s^i_0\}_i$, learning rates $\beta_m = \frac{2}{(1-\gamma)(t_0+m-1)}, \forall m \geq 0$ and $M > \mathcal{O}(\varepsilon^{-2})$, $M_{td} > \mathcal{O}(\log \varepsilon^{-1})$. If $\bar{\pi} \in \Pi$ is an arbitrary policy, $\Delta_{\bar{\pi}} := \frac{1}{N}\sum_i \|\pi^i - \bar{\pi}\|_1$ and $Q^* := Q_h(\cdot,\cdot|\bar{\pi},\mu_{\bar{\pi}})$, then the (random) output $\widehat{Q}_M$ of Algorithm 1 satisfies*

$$\mathbb{E}[\|\widehat{Q}_M - Q^*\|_\infty] \leq \varepsilon + \mathcal{O}\left(\frac{1}{\sqrt{N}} + \Delta_{\bar{\pi}} + \|\pi^1 - \bar{\pi}\|_1\right).$$

Proof. See Theorem D.7. □

If the policies of all agents are equal (i.e., $\pi^i = \bar{\pi}$ for all i), then Theorem 4.2 suggests an expected ℓ_∞ accuracy of $\varepsilon > 0$ can be achieved with respect to the "limiting mean field" $\Gamma^\infty_{pop}(\bar{\pi})$ in $\mathcal{O}(\varepsilon^{-2}\log\varepsilon^{-1})$ time steps, up to a finite population bias $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ as expected.

4.2. Main Results: Learning MFG-NE from Samples

In this section, we analyze two algorithms to estimate MFG-NE from a single sample path using N agents. In the first (Algorithm 2), the policies are synchronized by a centralized controller, allowing each agent to follow the same policy yielding a superior sampling complexity. In the second (Algorithm 3), learning is performed completely independently by each agent utilizing their local observations only. The algorithms are based on repeating the following iteration:

1. **Estimate values:** Each agent, keeping their policies constant, performs CTD learning for $M_{pg} > 0$ steps, each time waiting M_{td} steps between TD updates,

2. **Policy update:** Afterwards, agents simultaneously perform *one* policy mirror ascent update using Q-value estimates (the same update in the centralized case).

Both algorithms use samples from the SAGS in the form $s^i_{t+1} \sim P(\cdot|s^i_t, a^i_t, \widehat{\mu}_t)$, and $r^i_t = R(s^i_t, a^i_t, \widehat{\mu}_t)$, where $\widehat{\mu}_t$ is the *empirical* state distribution of the N agents at time t which the algorithm can not control, hence they are single-path and generative-model-free algorithms.

Algorithm 2 Centralized MFG-PMA

```
Require: parameters $K, M_{td}, M_{pg}, \eta, \{\beta_m\}_m$.
Require: initial states $\{s^i_0\}_i$.
  Set $\pi_0 = \pi_{max}$ and $t \leftarrow 0$.
  for $k \in 0, \dots, K-1$ do
    $\forall s, a : \widehat{Q}_0(s,a) = Q_{max}$
    for $m \in 0, \dots, M_{pg}-1$ do
      for $M_{td}$ iterations do
        Take step $\forall i$: $a^i_t \sim \pi_k(s^i_t)$, $r^i_t = R(s^i_t, a^i_t, \widehat{\mu}_t)$,
        $s^i_{t+1} \sim P(\cdot|s^i_t, a^i_t, \widehat{\mu}_t)$.
        $t \leftarrow t+1$
      end for
      Set $\zeta^1_{t-2} = (s^1_{t-2}, a^1_{t-2}, r^1_{t-2}, s^1_{t-1}, a^1_{t-1})$.
      CTD step: $\widehat{Q}_{m+1} = \widehat{Q}_m - \beta_m \widetilde{F}^{\pi_k}\left(\widehat{Q}_m, \zeta^1_{t-2}\right)$
    end for
    PMA step: $\pi_{k+1} = \Gamma^{md}_\eta(\widehat{Q}_{M_{pg}}, \pi_k)$
  end for
  Return policy $\pi_K$
```

Algorithm 3 Independent MFG-PMA

```
Require: parameters $K, M_{td}, M_{pg}, \eta, \{\beta_m\}_m$.
Require: initial states $\{s^i_0\}_i$.
  Set $\pi^i_0 = \pi_{max}, \forall i$ and $t \leftarrow 0$.
  for $k \in 0, \dots, K-1$ do
    $\forall s, a, i : \widehat{Q}^i_0(s,a) = Q_{max}$
    for $m \in 0, \dots, M_{pg}-1$ do
      for $M_{td}$ iterations do
        Take step $\forall i$: $a^i_t \sim \pi^i_k(s^i_t)$, $r^i_t = R(s^i_t, a^i_t, \widehat{\mu}_t)$,
        $s^i_{t+1} \sim P(\cdot|s^i_t, a^i_t, \widehat{\mu}_t)$.
        $t \leftarrow t+1$
      end for
      Set $\forall i : \zeta^i_{t-2} = (s^i_{t-2}, a^i_{t-2}, r^i_{t-2}, s^i_{t-1}, a^i_{t-1})$.
      CTD step $\forall i$: $\widehat{Q}^i_{m+1} = \widehat{Q}^i_m - \beta_m \widetilde{F}^{\pi^i_k}\left(\widehat{Q}^i_m, \zeta^i_{t-2}\right)$
    end for
    PMA step $\forall i : \pi^i_{k+1} = \Gamma^{md}_\eta(\widehat{Q}^i_{M_{pg}}, \pi^i_k)$
  end for
  Return policies $\{\pi^i_K\}_i$
```

We first present the main result for centralized learning. The centralized algorithm uses samples from the path of a single agent among N (for the first agent $i = 1$) and updates the policies of all agents accordingly. For simplicity, algorithms are presented in a form that requires the number of iterations to be set beforehand but can be trivially converted into algorithms that take the sample budget as input.

Theorem 4.3 (Centralized learning)**.** *Assume that $\eta > 0$ an arbitrary learning rate which satisfies $L_{\Gamma_\eta} < 1$, Assumptions 1, 2, 3 and 4 hold and π^* is the unique MFG-NE. Let $\varepsilon > 0$ be arbitrary. If the learning rates $\{\beta_m\}$ are as defined in Lemma 4.2 and $K > \mathcal{O}(\log\varepsilon^{-1}), M_{td} > \mathcal{O}(\log\varepsilon^{-1}), M_{pg} > \mathcal{O}(\varepsilon^{-2})$ then the (random) output π_K of Algorithm 2 satisfies $\mathbb{E}\left[\|\pi_K - \pi^*\|_1\right] \leq \varepsilon + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$.*

Proof. (sketch) The proof builds up on Theorem 4.2 to build estimates of the true Q-values $Q(\cdot,\cdot|\pi_k, \Gamma^\infty_{pop}(\pi_k))$ at each outer loop. The estimates allow the approximate evaluation of the Γ_η operator which was shown to contract to the MFG-NE. See Theorem D.8 for the full statement and proof. □

Remark 4.4. In Algorithm 2, the total number of time steps simulated from the SAGS is equal to $K \times M_{pg} \times M_{td}$; hence, Theorem 4.3 implies a $\mathcal{O}(\varepsilon^{-2}\log^2\varepsilon^{-1})$ sample complexity. In fact, since MFG-NE are $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$-NE of the N-agent SAGS, Theorem 4.3 shows that a Algorithm 2 finds a $\varepsilon + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$-NE of the SAGS in $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ time steps.

Finally, we present the case where each agent learns separate policies only by observing their own state-action transitions, presented in Algorithm 3. It only uses the observations of agent i to update policy π^i_t, hence learning is independent. Moreover, the algorithm does not employ multiple timescales and it has the desirable *symmetry* property that each learner follows the same protocol, meaning no handshake procedure between learners is required before learning starts (other than agreement on the number of epochs, K). The proof utilizes the bounds in Theorem 4.2 in terms of policy deviations $\Delta_{\bar{\pi}}$ of agents.

Theorem 4.5 (Independent learning)**.** *Assume that* $\eta > 0$ *satisfies* $L_{\Gamma_\eta} < 1$*, Assumptions 1, 2, 3 and 4 hold and* π^* *is the unique MFG-NE. Let* $\varepsilon > 0$ *be arbitrary. Let the learning rates* $\{\beta_m\}$ *for CTD be as defined in Lemma 4.2. There exists a problem dependent constant* $a \in [0, \infty)$ *such that if* $K = \lceil \frac{\log 8\varepsilon^{-1}}{\log L_{\Gamma_\eta}^{-1}} \rceil$*,* $M_{td} > \mathcal{O}(\log^2 \varepsilon^{-1})$*, and* $M_{pg} > \mathcal{O}(\varepsilon^{-2-a})$*, then the (random) output* $\{\pi_K^i\}_i$ *of Algorithm 3 satisfies for all agents* $i = 1, \dots, N$*,* $\mathbb{E}\left[\|\pi_K^i - \pi^*\|_1\right] \leq \varepsilon + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$.

Proof. (sketch) The proof in the independent learning case also builds up on Theorem 4.2, however, as the policies of agents diverge due to stochasticity, the terms $\Delta_{\bar{\pi}}$ become significant as they do not vanish. Therefore, the choice of K and consequently M_{pg} ensure the final error can be made to be less than ε up to finite population bias, which disappears as $N \to \infty$. See Theorem D.9 for the full statement and proof. □

Remark 4.6. The constant a in Theorem 4.5 is 0 for smooth enough problems. Hence the theorem implies a sample complexity of $\widetilde{\mathcal{O}}(\varepsilon^{-2})$ for SAGS with sufficiently smooth dynamics (e.g., large enough ρ, small enough constants K_a, L_μ), even though these smoothness conditions might be too restrictive. In the general case however, the sampling complexity will be $\widetilde{\mathcal{O}}(\varepsilon^{-2-a})$, where a is affected by the learning rate η and has a logarithmic dependency in general on the parameters $T_{mix}, \delta_{mix}^{-1}$, and p_{inf}^{-1}.

Concluding this section, we comment on the differences in sample complexity upper bounds between the centralized and independent learning cases. In the independent learning case, the worse $\widetilde{\mathcal{O}}(\varepsilon^{-2-a})$ sample complexity arises due to the fact that the MFG-NE formalism requires the policies of all N agents to be *exactly* the same. As variances in the agents' updates accumulate a non-vanishing bias in the later stages of the outer loop (due to the terms $\Delta_{\bar{\pi}}$), the CTD learning loop must be executed for a long time to keep the variance sufficiently low. It is possible that this is an artifact of the analysis technique, and we leave possible improvements as future work.

5. Discussion

We prove that policy mirror ascent can achieve an improved sample complexity for computing the approximate Nash equilibrium of an N-player SAGS. Our approach is complementary to existing MFG literature in that we analyze mirror ascent type operators allowing us to remove strong generative model assumptions to obtain an algorithm for the truly N-player game. Future work could involve a finer (e.g., potential-based) analysis using results in policy mirror descent literature.

Acknowledgements

This project has been carried out with funding from the Swiss National Science Foundation under the framework of NCCR Automation.

References

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

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

Carmona, R., Laurière, M., and Tan, Z. Model-free mean-field reinforcement learning: mean-field mdp and mean-field q-learning. *arXiv preprint arXiv:1910.12802*, 2019.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Resesarch*, 70: 2563–2578, 2022.

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

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

Ding, D., Wei, C.-Y., Zhang, K., and Jovanovic, M. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*, pp. 5166–5220. PMLR, 2022.

Fu, Z., Yang, Z., Chen, Y., and Wang, Z. Actor-critic provably finds nash equilibria of linear-quadratic mean-field games. In *International Conference on Learning Representations*, 2019.

Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169. PMLR, 2019.

Georgii, H.-O. Gibbs measures and phase transitions. In *Gibbs Measures and Phase Transitions*. de Gruyter, 2011.

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

Guo, X., Hu, A., Xu, R., and Zhang, J. Learning mean-field games. *Advances in Neural Information Processing Systems*, 32, 2019.

Guo, X., Hu, A., Xu, R., and Zhang, J. A general framework for learning mean-field games. *Mathematics of Operations Research*, 2022a.

Guo, X., Xu, R., and Zariphopoulou, T. Entropy regularization for mean field games with learning. *Mathematics of Operations Research*, 2022b.

Huang, M., Malhamé, R. P., and Caines, P. E. Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the nash certainty equivalence principle. *Communications in Information & Systems*, 6(3): 221–252, 2006.

Kontorovich, L. A. and Ramanan, K. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126–2158, 2008.

Kotsalis, G., Lan, G., and Li, T. Simple and optimal methods for stochastic variational inequalities, ii: Markovian noise and policy evaluation in reinforcement learning. *SIAM Journal on Optimization*, 32(2):1120–1155, 2022.

Lan, G. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, pp. 1–48, 2022.

Lasry, J.-M. and Lions, P.-L. Mean field games. *Japanese journal of mathematics*, 2(1):229–260, 2007.

Laurière, M., Perrin, S., Girgin, S., Muller, P., Jain, A., Cabannes, T., Piliouras, G., P'erolat, J., Elie, R., Pietquin, O., and Geist, M. Scalable deep reinforcement learning algorithms for mean field games. In *International Conference on Machine Learning*, 2022.

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

Li, Y., Zhao, T., and Lan, G. Homotopic policy mirror descent: Policy convergence, implicit regularization, and improved sample complexity. *arXiv preprint arXiv:2201.09457*, 2022.

Mao, W., Qiu, H., Wang, C., Franke, H., Kalbarczyk, Z., Iyer, R., and Basar, T. A mean-field game approach to cloud resource management with function approximation. In *Advances in Neural Information Processing Systems*, 2022.

Matignon, L., Laurent, G. J., and Le Fort-Piat, N. Hysteretic q-learning: an algorithm for decentralized reinforcement learning in cooperative multi-agent teams. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 64–69. IEEE, 2007.

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

Nesterov, Y. et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Ozdaglar, A., Sayin, M. O., and Zhang, K. Independent learning in stochastic games. *arXiv preprint arXiv:2111.11743*, 2021.

Perolat, J., Munos, R., Lespiau, J.-B., Omidshafiei, S., Rowland, M., Ortega, P., Burch, N., Anthony, T., Balduzzi, D., De Vylder, B., et al. From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *International Conference on Machine Learning*, pp. 8525–8535. PMLR, 2021.

Pérolat, J., Perrin, S., Elie, R., Laurière, M., Piliouras, G., Geist, M., Tuyls, K., and Pietquin, O. Scaling mean field games by online mirror descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 1028–1037, 2022.

Perrin, S., Pérolat, J., Laurière, M., Geist, M., Elie, R., and Pietquin, O. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in Neural Information Processing Systems*, 33:13199–13213, 2020.

Perrin, S., Laurière, M., Pèrolat, J., Geist, M., Elie, R., and Pietquin, O. Mean field games flock! the reinforcement learning way. In *IJCAI*, 2021.

Rashedi, N., Tajeddini, M. A., and Kebriaei, H. Markov game approach for multi-agent competitive bidding strategies in electricity market. *IET Generation, Transmission & Distribution*, 10(15):3756–3763, 2016.

Saldi, N., Basar, T., and Raginsky, M. Markov–nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization*, 56(6):4256–4287, 2018.

Samvelyan, M., Rashid, T., Schroeder de Witt, C., Farquhar, G., Nardelli, N., Rudner, T. G. J., Hung, C.-M., Torr, P. H. S., Foerster, J., and Whiteson, S. The starcraft multi-agent challenge. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*, AAMAS '19, pp. 2186–2188, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems.

Sayin, M., Zhang, K., Leslie, D., Basar, T., and Ozdaglar, A. Decentralized q-learning in zero-sum markov games. *Advances in Neural Information Processing Systems*, 34: 18320–18334, 2021.

Shalev-Shwartz, S. and Singer, Y. *Online learning: Theory, algorithms, and applications*. PhD thesis, Hebrew University, 2007.

Shavandi, A. and Khedmati, M. A multi-agent deep reinforcement learning framework for algorithmic trading in financial markets. *Expert Systems with Applications*, 208: 118124, 2022.

Subramanian, J. and Mahajan, A. Reinforcement learning in stationary mean-field games. In *Proceedings of the 18th International Conference on Autonomous Agents and Multi Agent Systems*, pp. 251–259, 2019.

Tomar, M., Shani, L., Efroni, Y., and Ghavamzadeh, M. Mirror descent policy optimization. In *International Conference on Learning Representations*, 2021.

Tsitsiklis, J. and Van Roy, B. Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.

Wiering, M. A. Multi-agent reinforcement learning for traffic light control. In *Machine Learning: Proceedings of the Seventeenth International Conference (ICML'2000)*, pp. 1151–1158, 2000.

Xie, Q., Yang, Z., Wang, Z., and Minca, A. Learning while playing in mean-field games: Convergence and optimality. In *International Conference on Machine Learning*, pp. 11436–11447. PMLR, 2021.

Yang, J., Ye, X., Trivedi, R., Xu, H., and Zha, H. Learning deep mean field games for modeling large population behavior. In *International Conference on Learning Representations (ICLR)*, 2018.

Yongacoglu, B., Arslan, G., and Yüksel, S. Independent learning in mean-field games: Satisficing paths and convergence to subjective equilibria. *arXiv preprint arXiv:2209.05703*, 2022.

Zaman, M. A. u., Koppel, A., Bhatt, S., and Başar, T. Oracle-free reinforcement learning in mean-field games along a single sample path. *arXiv preprint arXiv:2208.11639*, 2022.

A. Clarification on Notation and Constants

The topological interior and boundary of a set $\mathcal{X} \subset \mathbb{R}^D$ are denoted by $\mathcal{X}^\circ$ and $\partial\mathcal{X}$ respectively. We define the point-set distance $d(x, \mathcal{X}) := \inf_{x' \in \mathcal{X}} \|x - x'\|_2$. $\mathcal{X}^{m \times n}$ denotes the set of $m \times n$ matrices with entries from the set $\mathcal{X}$. For $\mathbf{A} \in \mathcal{X}^{m\times n}$, the i-th row vector is denoted as $\mathbf{A}_{i\cdot}$ and the j-th column vector as $\mathbf{A}_{\cdot j}$. Stochastic matrices are defined as matrices with positive entries and with row sums equal to 1. $\mathbf{e}_x \in \Delta_{\mathcal{X}}$ for $x \in \mathcal{X}$ is the vector with only the entry corresponding to x set to 1. $\|\cdot\|_*$ denotes the dual of a norm $\|\cdot\|$, given by the standard definition

$$\|\mathbf{x}\|_* := \sup\{\mathbf{y}^\top \mathbf{x} : \mathbf{y} \in \mathbb{R}^D, \|\mathbf{y}\| \leq 1\}, \forall \mathbf{x} \in \mathbb{R}^D.$$

A note on notation for functions on $\mathcal{A}$. In certain cases, it will be useful to alternate between (equivalent) definitions of certain functions on $\Delta_{\mathcal{A}}$ and $\mathcal{A}$ (see Anahtarci et al., 2022). For instance, one can define state transition probabilities as functions $P : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}} \to \Delta_{\mathcal{S}}$ taking state-action pairs as input, or equivalently, as a function $\bar{P} : \mathcal{S} \times \Delta_{\mathcal{A}} \times \Delta_{\mathcal{S}} \to \Delta_{\mathcal{S}}$ taking state-probability distribution of actions with the relation $\bar{P}(s'|s, u, \mu) := \sum_{a \in \mathcal{A}} u(a) P(s'|s, a, \mu)$. To ease reading we follow the convention that for a function on f on $\mathcal{A}$, we denote the equivalent function on $\Delta_{\mathcal{A}}$ with $\bar{f}$.

Notation for constants. Our analysis will involve multiple Lipschitz constants. For simplicity and readability, we use the convention that for a function of multiple variables say $f(x, y, \mu)$, the Lipschitz constant of f with respect to x, y, μ are denoted by $L_{f,x}, L_{f,y}, L_{f,\mu}$ respectively.

B. Basic Lemmas

In this section, we present several general lemmas re-used throughout the paper.

Lemma B.1 (p. 141, (Georgii, 2011))**.** *Assume E a finite set, $F : E \to \mathbb{R}$ a real valued function, and ν, μ two probability measures on E. Then,*

$$\left|\sum_e F(e)\mu(e) - \sum_e F(e)\nu(e)\right| \leq \frac{\sup_e F(e) - \inf_e F(e)}{2} \|\mu - \nu\|_1.$$

We provide two generalizations of this lemma, the first one adapted from Kontorovich & Ramanan (2008, Lemma A2).

Lemma B.2. *Assume E a finite set, $g : E \to \mathbb{R}^p$ a vector value function, and ν, μ two probability measures on E. Then,*

$$\left\|\sum_e g(e)\mu(e) - \sum_e g(e)\nu(e)\right\|_1 \leq \frac{\lambda_g}{2}\|\mu - \nu\|_1,$$

where $\lambda_g := \sup_{e,e'} \|g(e) - g(e')\|_1$.

Proof. We use the fact that $\|u\|_1 = \sup_{\|v\|_\infty \leq 1} u^\top v$ by duality of the norms. Applying this on the vector $\sum_e g(e)\left(\mu(e) - \nu(e)\right)$, we obtain

$$\left\|\sum_e g(e)\left(\mu(e) - \nu(e)\right)\right\|_1 = \sup_{\|v\|_\infty \leq 1} \sum_e \left(\mu(e) - \nu(e)\right) g(e)^\top v.$$

Take any $v \in \mathbb{R}^d$ such that $\|v\|_\infty \leq 1$ and apply Lemma B.1,

$$\begin{aligned}\left|\sum_e \left(\mu(e) - \nu(e)\right) g(e)^\top v\right| &\leq \|\mu - \nu\|_1 \frac{\sup_e g(e)^\top v - \inf_e g(e)^\top v}{2} & (3)\\ &\leq \|\mu - \nu\|_1 \frac{\sup_{e,e'} \|g(e') - g(e)\|_1}{2} & (4)\end{aligned}$$

where the last line follows from the fact that E is finite and for any $e, e' \in E$,

$$|g(e)^\top v - g(e')^\top v| \leq \|g(e') - g(e)\|_1 \|v\|_\infty = \|g(e') - g(e)\|_1,$$

by an application of Hölder's inequality. Then taking the supremum of the left hand side in (3), the result follows. □

We further generalize this lemma below.

Lemma B.3. *Assume $\mathbf{v}_1, \mathbf{v}_2 \in \Delta_D$ and $\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(K)}, \mathbf{B}^{(1)}, \dots, \mathbf{B}^{(K)} \in \mathbb{R}^{D\times D}_{\geq 0}$ are stochastic matrices. Then, it holds that*

$$\|\mathbf{v}_1^\top \mathbf{A}^{(1)}\mathbf{A}^{(2)}\dots\mathbf{A}^{(K)} - \mathbf{v}_2^\top \mathbf{B}^{(1)}\mathbf{B}^{(2)}\dots\mathbf{B}^{(K)}\|_1 \leq \|\mathbf{v}_1 - \mathbf{v}_2\|_1 + \sum_{k=1}^{K} \sup_j \|\mathbf{A}^{(k)}_{j\cdot} - \mathbf{B}^{(k)}_{j\cdot}\|_1.$$

Proof. We prove by induction on k. For $k = 1$, we have

$$\begin{aligned}\|\mathbf{v}_1^\top \mathbf{A}^{(1)} - \mathbf{v}_2^\top \mathbf{B}^{(1)}\|_1 &\leq \|(\mathbf{v}_1 - \mathbf{v}_2)^\top \mathbf{A}^{(1)}\|_1 + \|\mathbf{v}_2^\top(\mathbf{A}^{(1)} - \mathbf{B}^{(1)})\|_1 \\ &\overset{\text{Lemma B.2}}{\leq} \|\mathbf{v}_1 - \mathbf{v}_2\|_1 + \|\mathbf{v}_2^\top(\mathbf{A}^{(1)} - \mathbf{B}^{(1)})\|_1 \\ &\overset{\text{Jensen's}}{\leq} \|\mathbf{v}_1 - \mathbf{v}_2\|_1 + \sum_{j=1}^{n} \|\mathbf{A}^{(1)}_{j\cdot} - \mathbf{B}^{(1)}_{j\cdot}\|_1 \mathbf{v}_2(j) \\ &\leq \|\mathbf{v}_1 - \mathbf{v}_2\|_1 + \sup_j \|\mathbf{A}^{(1)}_{j\cdot} - \mathbf{B}^{(1)}_{j\cdot}\|_1.\end{aligned}$$

Assuming it holds for some $K \geq 1$, the statement also holds for $K + 1$ since

$$\begin{aligned}\|\mathbf{v}_1^\top \mathbf{A}^{(1)}\mathbf{A}^{(2)}\dots\mathbf{A}^{(K+1)} - \mathbf{v}_2^\top \mathbf{B}^{(1)}\mathbf{B}^{(2)}\dots\mathbf{B}^{(K+1)}\|_1 \leq& \|\mathbf{v}_1^\top \mathbf{A}^{(1)}\dots\mathbf{A}^{(K)} - \mathbf{v}_2^\top \mathbf{B}^{(1)}\dots\mathbf{B}^{(K)}\|_1 \\ &+ \sup_j \|\mathbf{A}^{(K+1)}_{j\cdot} - \mathbf{B}^{(K+1)}_{j\cdot}\|_1.\end{aligned}$$

Hence the general statement follows. □

Finally, we use the widely known Fenchel duality to establish certain Lipschitz regularity results. We re-iterate the following definition of strong convexity and strong concavity with respect to arbitrary norms.

Definition B.4 (Strong convexity)**.** Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be a convex differentiable function with domain $S := \{x \in \mathbb{R}^d : f(x) \in \mathbb{R}\}$ and let $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be an arbitrary norm on $\mathbb{R}^d$. If for any $x, y \in S$ it holds that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\rho\|y - x\|^2,$$

then f is called a strongly convex function with modulus ρ with respect to norm $\|\cdot\|$.

As expected, if $-f$ is a strongly convex function, f is called a strongly concave function with respect to norm $\|\cdot\|$. We list standard properties of strong convexity, which hold with respect to arbitrary norms. If f_1, f_2 are ρ_1, ρ_2 strongly convex with respect to $\|\cdot\|$ and $\alpha_1, \alpha_2 > 0$, then $\alpha_1 f_1 + \alpha_2 f_2$ is $\alpha_1\rho_1 + \alpha_2\rho_2$ strongly convex with respect to $\|\cdot\|$. If $|||\cdot|||$ and $\|\cdot\|$ are equivalent norms so that $|||\cdot||| \geq c\|\cdot\|$ for some constant c, and if f is ρ-strongly convex with respect to $|||\cdot|||$, then it is $c^2\rho$-strongly convex with respect to $\|\cdot\|$. Finally, ρ-strong convexity of f with respect to $\|\cdot\|$ implies that for all x_1, x_2 in the domain of f:

$$\left[\nabla f(x_1) - \nabla f(x_2)\right]^\top (x_1 - x_2) \geq \rho\|x_1 - x_2\|^2.$$

We will also need the following standard concept of a Fenchel conjugate.

Definition B.5 (Fenchel conjugate)**.** Assume that $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a convex function, with domain $S \subset \mathbb{R}^d$. The Fenchel conjugate $f^* : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is defined as

$$f^*(y) = \sup_{x \in S}\langle x, y\rangle - f(x).$$

For further details regarding the Fenchel conjugate, see Nesterov et al. (2018). The Fenchel conjugate is useful due to the following well-known duality result.

Lemma B.6. *Assume that $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is differentiable and τ-strongly convex with respect to a norm $\|\cdot\|$ and has domain $S \subset \mathbb{R}^d$. Then,*

1. *f^* is differentiable on $\mathbb{R}^d$,*
2. *$\nabla f^*(y) = \arg\max_{x \in S} \langle x, y \rangle - f(x)$,*
3. *f^* is $\frac{1}{\tau}$-smooth with respect to $\| \cdot \|_*$, i.e., $\| \nabla f^*(y) - \nabla f^*(y') \| \leq \frac{1}{\tau} \|y - y'\|_*, \forall y, y' \in \mathbb{R}^d$.*

Proof. See Lemma 15 of (Shalev-Shwartz & Singer, 2007) or Lemma 6.1.2 of (Nesterov et al., 2018). □

C. Operators and Lipschitz Continuity

In this section, we provide the proofs regarding the Lipschitz continuity of operators mentioned in the main text. Some of these results are organized and generalized from (Anahtarci et al., 2022) to our setting, while others relevant to mirror ascent are unique to our case.

C.1. Discussion of Assumptions for Contractivity

The construction or assumption of an operator that contracts to the MFG-NE is a common strategy in stationary MFG literature. Before we move on to presenting our results and proofs for obtaining a contractive operator, we compare our assumptions with several other relevant works in literature. We present Table 2 as a summary.

Contractivity assumption	**Additional/implied assumption on Γ_{pop}**	**Works**
Blanket contractivity	$\Gamma_{pop}(\cdot, \pi)$ contraction (Assumption 2)	Guo et al. (2019), Guo et al. (2022a), Zaman et al. (2022), Xie et al. (2021)
Lipschitz continuous P, R (Assumption 1) + regularization	Access to mean-field oracle Γ_{pop}^{∞} + Lipschitz continuous Γ_{pop}^{∞}	Cui & Koeppl (2021)
Lipschitz continuous P, R (Assumption 1) + regularization	$\Gamma_{pop}(\cdot, \pi)$ contraction (Assumption 2)	**Our work**, Anahtarci et al. (2022)

Table 2. A summary of existing assumptions in stationary MFGs in literature for obtaining a contraction. We note that the works by Xie et al. (2021) and Zaman et al. (2022) refer to past works for justification of the contraction assumption.

A note on assumptions on Γ_{pop}. The assumption on the population distribution $\Gamma_{pop}(\cdot, \pi)$ being a contraction for all π is implicit in most past works as shown in Table 2. For instance the assumption in Theorem 4 of in Guo et al. (2019) and Theorem 4 of Guo et al. (2022a) that $d_1 d_2 + d_3 < 1$ for positive Lipschitz constants d_1, d_2, d_3 implies $d_3 < 1$, where d_3 satisfies (in the notation of the mentioned paper)

$$\mathcal{W}_1 \left(\Gamma_2 \left(\pi, \mathcal{L}^1 \right), \Gamma_2 \left(\pi, \mathcal{L}^2 \right) \right) \leq d_3 \mathcal{W}_1 \left(\mathcal{L}^1, \mathcal{L}^2 \right),$$

where Γ_2 is equivalent to Γ_{pop} in our notation, $\mathcal{L}^1, \mathcal{L}^2 \in \Delta_{\mathcal{S}}$ and $\mathcal{W}_1$ is the ℓ_1 Wasserstein distance. Likewise, Assumption 2 of Xie et al. (2021) imposes that there exists d_3 satisfying $\|\Gamma_2(\pi, \mu) - \Gamma_2(\pi, \mu')\|_{\mathcal{H}} \leq d_3 \|\mu - \mu'\|_{\mathcal{H}}$ for all π, μ with the additional restriction $d_1 d_2 + d_3 < 1$, where $\| \cdot \|_{\mathcal{H}}$ is norm associated with the RHKS of a bounded and universal kernel k. Similarly, Assumption 1 of Zaman et al. (2022) states that (once again, Γ_2 being equivalent to Γ_{pop} in our notation):

$$\|\Gamma_2(\pi, \mu) - \Gamma_2(\pi, \mu')\|_1 \leq d_3 \|\mu - \mu'\|_1,$$

with the additional restriction $d_1 d_2 + d_3 < 1$ for positive d_1, d_2, d_3 also implying $d_3 < 1$. In the case of Anahtarci et al. (2022), it is implied that $\Gamma_{pop}(\cdot, \pi)$ is a contraction for all $\pi \in \Pi' \subset \Pi$, where Π' is a set of possible best response policies which satisfy a smoothness condition as their Lemma 3 suggests. This is not immediate since they do not define Γ_{pop} explicitly, but a careful analysis of their proof of Proposition 3, their assumption that $K_H < 1$ and their population update operator $H_2 : \Delta_{\mathcal{S}} \times \mathcal{Q} \to \Delta_{\mathcal{S}}$ defined on value functions rather than policies yields this result. Finally, Cui & Koeppl (2021) assume instead access to the mean-field population oracle Γ_{pop}^{∞} (in the notation of their paper, Ψ). This is in general different from assuming $\Gamma_{pop}(\cdot, \pi)$ is contractive for all π, as this instead implies that the fixed point equation $\Gamma_{pop}(\mu, \pi) = \mu$ can be solved for μ with a solution Lipschitz continuous in π.

In fact, we point out that it is not granted (to the best of our knowledge) that the infinite agent game is a good approximation for the N-agent game when $\Gamma_{pop}(\cdot, \pi)$ is not a contraction in some metric space for all π, for instance, Theorem 1 of Anahtarci et al. (2022) (summarized in Proposition 2.6 in our paper) requires a similar contraction constraint (in the notation of the paper) that $C_1 < 1$. This might be intuitive: if $\Gamma_{pop}(\cdot, \pi)$ is not a contraction, in general, the finite population bias might be amplified between time steps of the N-agent SAGS. This would make the stationarity condition of the population in the definition of MFG-NE irrelevant in the case of N agents and when there is stochasticity in the transition dynamics: the empirical distributions $\widehat{\mu}_t$ will not be close to μ^* for the MFG-NE population distribution μ^*. We leave it as an interesting question for future work if the finite agent SAGS is well approximated by the MFG without explicitly assuming contraction of $\Gamma_{pop}(\cdot, \pi)$ (for instance, when $\Gamma_{pop}(\cdot, \pi)$ is only non-expansive).

C.2. Lipschitz Continuity in $(\Delta_{\mathcal{A}}, \|\cdot\|_1)$ of P, R

Firstly, we define $\bar{R} : \mathcal{S} \times \Delta_{\mathcal{A}} \times \Delta_{\mathcal{S}} \to [0, 1]$ and $\bar{P} : \mathcal{S} \times \Delta_{\mathcal{A}} \times \Delta_{\mathcal{S}} \to \Delta_{\mathcal{S}}$ as the rewards and action probabilities on probability distributions over actions as $\bar{R}(s, u, \mu) := \sum_{a \in \mathcal{A}} u(a) R(s, a, \mu)$ and $\bar{P}(\cdot|s, u, \mu) := \sum_{a \in \mathcal{A}} u(a) P(\cdot|s, a, \mu)$ for all $s \in \mathcal{S}, u \in \Delta_{\mathcal{A}}, \mu \in \Delta_{\mathcal{S}}$. These alternative definitions will be practical when establishing certain Lipschitz identities later. As expected, $\bar{P}, \bar{R}$ are Lipschitz in their arguments.

Lemma C.1 (Lipschitz continuity of $\bar{P}, \bar{R}$)**.** *For all $s, s' \in \mathcal{S}, u, u' \in \Delta_{\mathcal{A}}, \mu, \mu' \in \Delta_{\mathcal{S}}$,*

$$\begin{aligned}
|\bar{R}(s, u, \mu) - \bar{R}(s', u', \mu')| \leq& L_\mu \|\mu - \mu'\|_1 + L_s d(s, s') + \frac{L_a}{2} \|u - u'\|_1 \\
\|\bar{P}(\cdot|s, u, \mu) - \bar{P}(\cdot|s', u', \mu')\|_1 \leq& K_\mu \|\mu - \mu'\|_1 + K_s d(s, s') + \frac{K_a}{2} \|u - u'\|_1
\end{aligned}$$

Proof. The lemma simply follows from the Lemma B.1, with the inequalities,

$$\begin{aligned}
|\bar{R}(s, u, \mu) - \bar{R}(s', u', \mu')| \leq& \left| \sum_{a \in \mathcal{A}} u(a) R(s, a, \mu) - \sum_{a \in \mathcal{A}} u(a) R(s', a, \mu') \right| \\
&+ \left| \sum_{a \in \mathcal{A}} u(a) R(s', a, \mu') - \sum_{a \in \mathcal{A}} u'(a) R(s', a, \mu') \right| \\
\leq& \sum_{a \in \mathcal{A}} u(a) |R(s, a, \mu) - R(s', a, \mu')| \\
&+ \left| \sum_{a \in \mathcal{A}} u(a) R(s', a, \mu') - \sum_{a \in \mathcal{A}} u'(a) R(s', a, \mu') \right|,
\end{aligned}$$

where the last line follows from Jensen's inequality. Using Lemma B.1 on the second summand we obtain the statement of the lemma.

Similarly for $\bar{P}$, we have

$$\begin{aligned}
\|\bar{P}(\cdot|s, u, \mu) - \bar{P}(\cdot|s', u', \mu')\|_1 \leq& \left\| \sum_{a \in \mathcal{A}} u(a) P(\cdot|s, a, \mu) - \sum_{a \in \mathcal{A}} u(a) P(\cdot|s', a, \mu') \right\|_1 \\
&+ \left\| \sum_{a \in \mathcal{A}} u(a) P(\cdot|s', a, \mu') - \sum_{a \in \mathcal{A}} u'(a) P(\cdot|s', a, \mu') \right\|_1 .
\end{aligned}$$

Using Jensen's inequality for the first term and Lemma B.2 for the second, we conclude. □

C.3. Lipschitz Continuity of Policy Mirror Ascent

As opposed to (Anahtarci et al., 2022), in this work we focus on a policy mirror ascent scheme, instead of a best response operator. This section includes relevant proofs for the continuity of the policy mirror descent response. Firstly, we prove the following useful lemma.

Lemma C.2 (Lipschitz continuity of mirror ascent)**.** *Assume that* $u_0 \in \Delta_{\mathcal{A}}, q \in \mathbb{R}^{\mathcal{A}}$*,* $\eta > 0$*,* $K \subset \Delta_{\mathcal{A}}$ *a convex closed set and define* $g_\eta : \mathbb{R}^{\mathcal{A}} \times \Delta_{\mathcal{A}} \to K$ *as*

$$g_\eta(q, u_0) = \arg\max_{u \in K} q^\top u + h(u) - \frac{1}{2\eta}\|u - u_0\|_2^2.$$

Then, g is Lipschitz in both q, u_0*, that is,*

$$\|g_\eta(q, u_0) - g_\eta(q', u_0')\|_1 \leq L_{g,q}\|q - q'\|_\infty + L_{g,u}\|u_0 - u_0'\|_1,$$

where $L_{g,q} = \frac{\eta|\mathcal{A}|}{1+\rho\eta|\mathcal{A}|}, L_{g,u} = \frac{1}{|\mathcal{A}|^{-1}+\eta\rho}$*.*

Proof. The Lipschitz continuity with respect to q simply follows from Lemma B.6, and the fact that $-h(u) + \frac{1}{\eta}\|u - u_0\|_2^2$ is $\rho + \frac{1}{\eta|\mathcal{A}|}$ strongly convex in u with respect to the norm $\|\cdot\|_1$. For the continuity with respect to u_0, we first define the function

$$f(u) = q^\top u + h(u) - \frac{1}{2\eta}\|u - u_0\|_2^2.$$

Computing the gradient, we obtain:

$$\nabla f(u) = q + \nabla h(u) - \frac{1}{\eta}(u - u_0).$$

By strong concavity of h, $-\nabla h$ is a strongly monotone operator with parameter ρ. Define by N_K the normal cone operator corresponding to the set K. Then N_K is (maximal) monotone since K is closed and convex. By first order optimality conditions, we have

$$\begin{aligned}
u^* = g_\eta(q, u_0) &\iff 0 \in \nabla f(u^*) - N_K(u^*) \\
&\iff 0 \in q + \nabla h(u^*) - \frac{1}{\eta}(u^* - u_0) - N_K(u^*) \\
&\iff 0 \in q + \frac{1}{\eta}u_0 - \left(\frac{1}{\eta}I - \nabla h + N_K\right)(u^*) \\
&\iff u^* = \left(\frac{1}{\eta}I - \nabla h + N_K\right)^{-1}(q + \frac{1}{\eta}u_0),
\end{aligned}$$

where the last line follows since the resolvent $\left(\frac{1}{\eta}I - \nabla h + N_K\right)^{-1}$ is a function. Since N_K is monotone, $\frac{1}{\eta}I$ is $\frac{1}{|\mathcal{A}|\eta}$-strongly monotone in $\|\cdot\|_1$ and $-\nabla h$ is ρ-strongly monotone in $\|\cdot\|_1$, $G := \left(\frac{1}{\eta}I - \nabla h + N_K\right)$ is a (set-valued) strongly monotone operator with parameter $\frac{1}{|\mathcal{A}|\eta} + \rho$. Hence by definition, it holds that for any $y_1 \in G(x_1), y_2 \in G(x_2)$

$$\begin{aligned}
\|y_1 - y_2\|_1\|x_1 - x_2\|_1 &\geq \|y_1 - y_2\|_\infty\|x_1 - x_2\|_1 \\
&\geq (y_1 - y_2)^\top(x_1 - x_2) \\
&\geq \left(\frac{1}{|\mathcal{A}|\eta} + \rho\right)\|x_1 - x_2\|_1^2
\end{aligned}$$

Then, G^{-1} is a Lipschitz function with constant $\frac{\eta}{|\mathcal{A}|^{-1}+\eta\rho}$ (between normed spaces $(\mathbb{R}^{\mathcal{A}}, \|\cdot\|_1) \to (\mathbb{R}^{\mathcal{A}}, \|\cdot\|_1)$), hence $G^{-1}(q + \frac{1}{\eta}u_0)$ is $\frac{1}{|\mathcal{A}|^{-1}+\eta\rho}$ Lipschitz continuous in u_0. It follows that g is Lipschitz continuous in u_0 with respect to norm $\|\cdot\|_1$ with constant $L_{g,u}$ defined in the theorem. □

With the above, we can prove Lemma C.3 below that states Lipschitz continuity in terms of the norms defined on Π, $\mathcal{Q}$ in the main text.

Lemma C.3 (Lipschitz continuity of Γ_η^{md})**.** *Γ_η^{md} is Lipschitz continuous in both of its arguments, so that for all $q, q' \in \mathcal{Q}, \pi, \pi' \in \Pi$, it holds that $\|\Gamma_\eta^{md}(q, \pi) - \Gamma_\eta^{md}(q', \pi')\|_1 \leq L_{md,\pi}\|\pi - \pi'\|_1 + L_{md,q}\|q - q'\|_\infty$, where $L_{md,q} = \frac{\eta|\mathcal{A}|}{1+\eta\rho|\mathcal{A}|}$ and $L_{md,\pi} = \frac{1}{|\mathcal{A}|^{-1}+\eta\rho}$.*

Proof. For each state $s \in \mathcal{S}$, we have $\Gamma_\eta^{md}(q, \pi)(s) = g_\eta(q(\cdot|s), \pi(s))$ with the set $K := \mathcal{U}_{L_h}$ in Lemma C.2. By the Lipschitz continuity of g_η shown in Lemma C.2, we can write

$$\|\Gamma_\eta^{md}(q, \pi)(\cdot|s) - \Gamma_\eta^{md}(q', \pi')(\cdot|s)\|_1 \leq L_{g,q}\|q(s, \cdot) - q'(s, \cdot)\|_\infty + L_{g,u}\|\pi(\cdot|s) - \pi'(\cdot|s)\|_1.$$

So taking the supremum of both sides with respect to s and using the definition of norms on Π, we obtain

$$\|\Gamma_\eta^{md}(q, \pi) - \Gamma_\eta^{md}(q', \pi')\|_1 \leq L_{g,q}\|q - q'\|_\infty + L_{g,u}\|\pi - \pi'\|_1.$$

So the desired Lipschitz constants are given by $L_{md,q} = L_{g,q}, L_{md,\pi} = L_{g,u}$. □

C.4. Definitions of Value Functions

In this section, we define various value functions that depend on the population distribution. The definitions are standard from single-agent RL literature.

Definition C.4 ($V_h(\cdot|\pi, \mu), Q_h(\cdot|\pi, \mu), q_h(\cdot|\pi, \mu)$)**.** The value function V and the Q-functions Q, q are respectively defined as

$$V_h(s|\pi, \mu) := \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t \left(R(s_t, a_t, \mu) + h(\pi(s_t))\right) \middle| \substack{s_0=s \\ s_{t+1}\sim P(\cdot|s_t,\pi_t^i(s_t),\mu) \\ a_t\sim\pi(\cdot|s_t)}\right],$$

$$Q_h(s, a|\pi, \mu) := \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t \left(R(s_t, a_t, \mu) + h(\pi(s_t))\right) \middle| \substack{s_0=s, a_0=a \\ s_{t+1}\sim P(\cdot|s_t,\pi_t^i(s_t),\mu) \\ a_t\sim\pi(\cdot|s_t)}\right],$$

$$q_h(s, a|\pi, \mu) := R(s, a, \mu) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} P(s'|s, a, \mu)\pi(a'|s')Q_h(s', a'|\pi, \mu).$$

Similarly, Q functions with argument in $u \in \Delta_{\mathcal{S}}$ in are defined as for all $s \in \mathcal{S}$, $\bar{Q}_h(s, u|\pi, \mu) := \sum_a Q_h(s, a|\pi, \mu)u(a)$ and $\bar{q}_h(s, u|\pi, \mu) := \sum_a q_h(s, a|\pi, \mu)u(a)$.

Likewise, we define the standard optimal value functions for the MDP as follows.

Definition C.5 ($V_h^*(\cdot|\mu), Q_h^*(\cdot|\mu), q_h^*(\cdot|\mu)$)**.** The optimal value function V and the Q-functions Q, q are respectively defined as

$$V_h^*(s|\mu) := \max_{\pi \in \Pi} V_h(s|\pi, \mu), \quad Q_h^*(s, a|\mu) := \max_{\pi \in \Pi} Q_h(s, a|\pi, \mu),$$
$$q_h^*(s, a|\mu) := \max_{\pi \in \Pi} q_h(s, a|\pi, \mu).$$

Similarly, Q functions with argument in $u \in \Delta_{\mathcal{S}}$ are defined as for all $s \in \mathcal{S}$, $\bar{Q}_h^*(s, u|\mu) := \sum_a Q_h^*(s, a|\mu)u(a)$ and $\bar{q}_h^*(s, u|\pi, \mu) := \sum_a q_h^*(s, a|\mu)u(a)$.

Finally, we state the very useful characterization of value functions as the fixed points of a certain Bellman operator. The next lemma states this standard result without proof.

Lemma C.6 (Value functions as fixed points)**.** *For any $\pi \in \Pi, \mu \in \Delta_{\mathcal{S}}$, the value functions V_h, Q_h, q_h (uniquely) satisfy*

$$V_h(s|\pi, \mu) = \sum_a \pi(a|s) \left(R(s, a, \mu) + h(\pi(s)) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a, \mu)V_h(s|\pi, \mu)\right),$$
$$Q_h(s, a|\pi, \mu) = R(s, a, \mu) + h(\pi(s)) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} P(s'|s, a, \mu)\pi(a'|s')Q_h(s', a'|\pi, \mu),$$
$$q_h(s, a|\pi, \mu) := R(s, a, \mu) + \gamma \sum_{s' \in \mathcal{S}} \sum_{a' \in \mathcal{A}} P(s'|s, a, \mu)\pi(a'|s')\left(q_h(s', a'|\pi, \mu) + h(\pi(s'))\right).$$

Likewise, the optimal value functions are uniquely defined as the fixed points satisfying

$$V_h^*(s|\mu) = \max_{u\in\Delta_{\mathcal{A}}} \left[\sum_a u(a) \left(R(s,a,\mu) + h(u) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a,\mu) V_h^*(s|\mu)\right)\right],$$

$$q_h^*(s,a|\mu) = R(s,a,\mu) + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,a,\mu) \max_{u\in\Delta_{\mathcal{A}}} \left[h(u) + \sum_{a'} u(a') q_h^*(s',a'|\mu)\right].$$

C.5. Lipschitz Continuity of Value Functions

In this section, we establish that Γ_q is Lipschitz continuous on a well-defined convex subset of Π. The main difficulty will be avoiding dependence on the Lipschitz continuity of h. We first prove two technical lemmas.

Lemma C.7. *Assume that $\pi \in \Pi_{\Delta h}$ and $\mu \in \Delta_{\mathcal{S}}$ arbitrary. Then, for any $s_1, s_2 \in \mathcal{S}$,*

$$|V_h(s_1|\pi,\mu) - V_h(s_2|\pi,\mu)| \leq L_{V,s} := \frac{L_s + L_a + \Delta h}{1 - \gamma \min\{1, \frac{K_s + K_a}{2}\}}.$$

Proof. Using the fixed point definition of V_h, Lemma B.1, and the definition of $\Pi_{\Delta h}$ in Eq. 1,

$$\begin{aligned}
&|V_h(s_1|\pi,\mu) - V_h(s_2|\pi,\mu)| \\
&\qquad \leq |\bar{R}(s_1,\pi(s_1),\mu) - \bar{R}(s_2,\pi(s_2),\mu)| + |h(\pi(s_1)) - h(\pi(s_2))| \\
&\qquad\quad + \gamma \sum_{s'} \left(\bar{P}(s'|s_1,\pi(s_1),\mu) - \bar{P}(s'|s_2,\pi(s_2),\mu)\right) V_h(s|\pi,\mu), \\
&\qquad \leq L_s + L_a + \Delta h \\
&\qquad\quad + \frac{\gamma \sup_{s,s'} |V_h(s|\pi,\mu) - V_h(s'|\pi,\mu)|}{2} \|P(s'|s,\pi(s_1),\mu) - P(s'|s_2,\pi(s_2),\mu)\|_1, \\
&\qquad \leq L_s + L_a + \Delta h + \frac{\gamma \min\{2, K_s + K_a\}}{2} \sup_{s,s'} |V_h(s|\pi,\mu) - V_h(s'|\pi,\mu)|,
\end{aligned}$$

which yields the lemma by taking the supremum of the left hand side over s_1, s_2. □

Lemma C.8 (Lipschitz continuity of V_h)**.** *Assume that $\Delta h > 0$ arbitrary. For any $\pi, \pi' \in \Pi_{\Delta h}$ and $\mu, \mu' \in \Delta_{\mathcal{S}}$,*

$$\|V_h(\cdot|\pi,\mu) - V_h(\cdot|\pi',\mu')\|_\infty \leq L_{V,\pi}\|\pi - \pi'\|_1 + L_{V,\mu}\|\mu - \mu'\|_1,$$

for $L_{V,\pi} = \frac{4L_a + \gamma K_a L_{V,s}}{4(1-\gamma)}$, $L_{V,\mu} = \frac{2L_\mu + \gamma K_\mu L_{V,s}}{2(1-\gamma)}$ and $L_{V,s}$ is as defined in Lemma C.7.

Proof. Similar to previous computations,

$$\begin{aligned}
|V_h(s|\pi,\mu) - V_h(s|\pi',\mu')| \leq &|\bar{R}(s,\pi(s),\mu) - \bar{R}(s,\pi'(s),\mu')| \\
&+ \gamma \left|\sum_{s'} \left(P(s'|s,\pi(s),\mu) V_h(s'|\pi,\mu) - P(s'|s,\pi'(s),\mu') V_h(s'|\pi',\mu')\right)\right| \\
\leq &L_a\|\pi - \pi'\|_1 + L_\mu\|\mu - \mu'\|_1 + \gamma \frac{L_{V,s}}{2}(K_\mu\|\mu - \mu'\|_1 + \frac{K_a}{2}\|\pi - \pi'\|_1) \\
&+ \gamma \sup_s |V_h(s|\pi,\mu) - V_h(s|\pi',\mu')|,
\end{aligned}$$

where the last line follows from an application of the triangle inequality and the previous lemma. □

The key result is that Γ_q is Lipschitz on a subset of policies given by $\Pi_{\Delta h}$ (Eq. (1)).

Lemma C.9 (Lipschitz continuity of Γ_q)**.** *Let $\Delta h > 0$ be arbitrary. There exists $L_{q,\pi}, L_{q,\mu}$ depending on Δh such that for all $\pi, \pi' \in \Pi_{\Delta h}$ and $\mu, \mu' \in \Delta_{\mathcal{S}}$,*

$$\|\Gamma_q(\pi, \mu) - \Gamma_q(\pi', \mu')\|_\infty \leq L_{q,\pi}\|\pi - \pi'\|_1 + L_{q,\mu}\|\mu - \mu'\|_1.$$

Proof. The result follows from the definition of q_h, in terms of V_h since the Lipschitz continuity of V_h has been shown in Lemma C.8. Specifically, we have

$$L_{q,\mu} = L_\mu + \gamma L_{V,\mu} + \gamma \frac{L_{V,s}K_\mu}{2}, \quad L_{q,\pi} = \gamma L_{V,\pi} + \gamma L_{V,s}K_a.$$

□

C.6. Lipschitz Continuity of Population Update

Proof of Lemma 3.2. The proof relies on Lemma B.2.

$$\begin{aligned}\|\Gamma_{pop}(\mu, \pi) - \Gamma_{pop}(\mu', \pi')\|_1 &= \left\| \sum_s \mu(s)\bar{P}(\cdot|s, \pi(s), \mu) - \sum_s \mu'(s)\bar{P}(\cdot|s, \pi'(s), \mu') \right\|_1 \\ &\leq \underbrace{\left\| \sum_s \mu(s)\bar{P}(\cdot|s, \pi(s), \mu) - \sum_s \mu(s)\bar{P}(\cdot|s, \pi'(s), \mu') \right\|_1}_{A} \\ &\quad + \underbrace{\left\| \sum_s \mu(s)\bar{P}(\cdot|s, \pi'(s), \mu') - \sum_s \mu'(s)\bar{P}(\cdot|s, \pi'(s), \mu') \right\|_1}_{B}.\end{aligned}$$

The first term can be bounded by using the Jensen's inequality:

$$A \leq \sum_s \mu(s) \left\|\bar{P}(\cdot|s, \pi(s), \mu) - \bar{P}(\cdot|s, \pi'(s), \mu')\right\|_1 \leq \frac{K_a}{2} \left\|\pi - \pi'\right\|_1 + K_\mu\|\mu - \mu'\|_1.$$

For the second term, using Lemma B.2, we obtain

$$B \leq \|\mu - \mu'\|_1 \frac{\sup_{s,s' \in \mathcal{S}} \|\bar{P}(\cdot|s, \pi'(s), \mu') - \bar{P}(\cdot|s', \pi'(s'), \mu')\|_1}{2}.$$

To bound the supremum, we use Lipschitz continuity of $\bar{P}$, to obtain for $s, s' \in \mathcal{S}$,

$$\begin{aligned}\|\bar{P}(\cdot|s, \pi'(s), \mu') - \bar{P}(\cdot|s', \pi'(s'), \mu')\|_1 &\leq K_s d(s, s') + K_a\|\pi'(s) - \pi'(s')\|_1 \\ &\leq (K_s + 2K_a)d(s, s'),\end{aligned}$$

from which the lemma follows. □

Finally, we characterize the Lipschitz continuity of Γ_{pop}^∞ as mentioned in the main text.

Lemma C.10 (Lipschitz continuity of Γ_{pop}^∞)**.** *Assume that Assumption 2 holds, that is, $L_{pop,\mu} < 1$. The mapping $\Gamma_{pop}^\infty : \Pi \to \Delta_{\mathcal{S}}$ is then Lipschitz with constant $L_{pop,\infty} := \frac{K_a}{2(1 - L_{pop,\mu})}$.*

Proof. Let $\pi, \pi' \in \Pi_L$, then by definition

$$\begin{aligned}\|\Gamma_{pop}^\infty(\pi) - \Gamma_{pop}^\infty(\pi')\|_1 &= \|\Gamma_{pop}(\Gamma_{pop}^\infty(\pi), \pi) - \Gamma_{pop}(\Gamma_{pop}^\infty(\pi'), \pi')\|_1 \\ &\leq L_{pop,\mu}\|\Gamma_{pop}^\infty(\pi) - \Gamma_{pop}^\infty(\pi')\|_1 + \frac{K_a}{2}\|\pi - \pi'\|_1,\end{aligned}$$

hence the result follows. □

C.7. Fixed Points and Continuity of Γ_η

Lemma C.9 proves the Lipschitz continuity of Γ_q only on a restricted subclass of policies. However, the next result shows that the MFG-NE policy can be contained in a subset $\Pi_{\Delta h}$ for some well-defined $\Delta h > 0$.

Lemma C.11. *Let $\mu \in \Delta_{\mathcal{S}}$ arbitrary, and $\pi^* \in \Pi$ the optimal response such that for all $s \in \mathcal{S}$, $V_h(s|\pi^*, \mu) = \max_{\pi \in \Pi} V_h(s|\pi, \mu)$. Then, $\pi^* \in \Pi_{L_h}$ where $L_h := L_a + \gamma \frac{L_s K_a}{2 - \gamma K_s}$.*

Proof. Firstly, using the fixed point definition of the optimal value function, we have for all $s_1, s_2 \in \mathcal{S}$:

$$\begin{aligned}
&|V_h^*(s_1|\mu) - V_h^*(s_2|\mu)| \\
&\quad \leq \Bigg| \sup_{u \in \Delta_{\mathcal{A}}} \left(\bar{R}(s_1, u, \mu) + h(u) + \gamma \sum_s \bar{P}(s|s_1, u, \mu) V_h^*(s|\mu) \right) \\
&\quad\quad - \sup_{u \in \Delta_{\mathcal{A}}} \left(\bar{R}(s_2, u, \mu) + h(u) + \gamma \sum_s \bar{P}(s|s_2, u, \mu) V_h^*(s|\mu) \right) \Bigg| \\
&\quad \leq \sup_{u \in \Delta_{\mathcal{A}}} \left| \bar{R}(s_1, u, \mu) - \bar{R}(s_2, u, \mu) + \gamma \sum_s (\bar{P}(s|s_1, u, \mu) - \bar{P}(s|s_2, u, \mu)) V_h^*(s|\mu) \right| \\
&\quad \leq L_s + \frac{\gamma K_s}{2} \sup_{s, s'} |V_h^*(s|\mu) - V_h^*(s'|\mu)|,
\end{aligned}$$

hence $|V_h^*(s_1|\mu) - V_h^*(s_2|\mu)| \leq \frac{L_s}{1 - \gamma K_s/2}$. Since $q_h^*(s, a|\mu) = R(s, a, \mu) + \gamma \sum_{s'} P(s'|s, a, \mu) V_h^*(s'|\mu)$, we can also conclude that for any $s \in \mathcal{S}$ that

$$\sup_{a, a' \in \mathcal{A}} |q_h^*(s, a|\mu) - q_h^*(s, a'|\mu)| \leq L_a + \gamma \frac{K_a}{2} \frac{L_s}{1 - \gamma K_s/2}.$$

Finally, by optimality conditions of the policy π^*, for any $s \in \mathcal{S}$ we have

$$\max_{u \in \Delta_{\mathcal{A}}} \langle q_h^*(s, \cdot|\mu), u \rangle + h(u) = \langle q_h^*(s, \cdot|\mu), \pi^*(s) \rangle + h(\pi^*(s)) \geq \langle q_h^*(s, \cdot|\mu), u_{max} \rangle + h_{max}.$$

This implies that $h_{max} - h(\pi^*(s)) \leq \langle q_h^*(s, \cdot|\mu), \pi(s) - u_{max} \rangle$. □

We note that the constant L_h obtained above is comparable to the constant K_{H_1} of Anahtarci et al. (2022). Hence the results above generalize the known Lipschitz continuity of optimal Q-values to general Q-values with respect to a policy π without introducing stringent assumptions.

Proof of Lemma 3.6. If π^*, μ^* satisfy the MFG-NE conditions, the two equalities follow from MDP optimality and Lemma C.11. Conversely, assume $\pi^* = \Gamma_\eta(\pi^*)$ and $\mu^* = \Gamma_{pop}^\infty(\pi^*)$. Note that this implies $\pi^* = \Gamma_\eta^{md}(q_h(\cdot|\pi^*, \mu^*), \pi^*)$. By optimality conditions on MDPs, it follows that π^* is optimal with respect to the MDP induced by μ^*. □

Lemma C.12 (Lipschitz continuity of Γ_η)**.** *For any $\eta > 0$, the operator $\Gamma_\eta : \Pi \to \Pi$ is Lipschitz with constant L_{Γ_η} on $(\Pi, \|\cdot\|_1)$, where*

$$L_{\Gamma_\eta} := \frac{L_{\Gamma,q} \eta |\mathcal{A}|}{1 + \eta \rho |\mathcal{A}|} + \frac{1}{|\mathcal{A}|^{-1} + \eta \rho} < \frac{L_{\Gamma,q}}{\rho} + \frac{1}{\eta \rho},$$

with the constant $L_{\Gamma,q}$ defined as $L_{\Gamma,q} := L_{pop,\infty} L_{q,\mu} + L_{q,\pi}$.

Proof. The result follows from combining previously established Lipschitz continuity results of Γ_η^{md} (Lemma C.3) and Γ_q (Lemma C.9) with the definition of Γ_η. □

Proof of Theorem 3.8. π^* is a fixed point of Γ_η and Γ_η is a contraction by Lemma 3.7. □

C.8. Continuity and Best Response

In this self-contained section, we present a short proof that a large class of best response operators can not be continuous and non-trivial in μ if $\mathcal{S}, \mathcal{A}$ are finite, despite what is typically assumed in stationary MFG literature. Firstly, we define the unregularized optimal action values $Q^*(s,a|\mu)$ for each $\mu \in \Delta_{\mathcal{S}}$ as

$$Q^*(s,a|\mu) := \max_{\pi\in\Pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, \mu) \middle|_{\substack{s_0=s \\ a_0=a,}} {}_{\substack{s_{t+1}\sim P(\cdot|s_t,a_t,\mu) \\ a_{t+1}\sim\pi(s_{t+1})}}, \forall t \geq 0 \right].$$

For any $s \in \mathcal{S}$, we also define the set-valued optimal action map $\mathcal{A}_s^* : \Delta_{\mathcal{S}} \to 2^{\mathcal{A}}$ as

$$\mathcal{A}_s^*(\mu) := \{a \in \mathcal{A} : Q^*(s,a|\mu) \geq Q^*(s,a'|\mu) \text{ for all } a' \in \mathcal{A}\} \in 2^{\mathcal{A}},$$

for each $\mu \in \Delta_{\mathcal{S}}$. We call a map $\Gamma_{br} : \Delta_{\mathcal{S}} \to \Pi$ a "best-response operator" (BR) if

$$\operatorname{supp} \Gamma_{br}(\mu)(s) \subset \mathcal{A}_s^*(\mu), \forall s \in \mathcal{S}, \mu \in \Delta_{\mathcal{S}}.$$

We also denote by $\Pi^*(\mu)$ the set of optimal policies for population distribution μ, hence a valid BR operator must satisfy $\Gamma_{br}(\mu) \in \Pi^*(\mu)$ for all μ. Γ_{br} is not unique in general, since it can assign non-zero action probabilities arbitrarily on $\mathcal{A}_s^*(\mu)$. The question for this section is if there could be a Γ_{br} that assigns probabilities to optimal actions so that it is continuous on $(\Delta_{\mathcal{S}}, \|\cdot\|_1)$ (or on any equivalent norm). We provide a negative answer for a fairly general subclass of operators between $\Delta_{\mathcal{S}}$ and Π.

Definition C.13 ("Optimal action stable" policy map)**.** Let $\Gamma : \Delta_{\mathcal{S}} \to \Pi$ be an arbitrary mapping. Let $S := |\mathcal{S}|, \mathcal{S} = \{s_1, \cdots, s_S\}$. Γ is called *optimal action stable (OAS)* if for any given subsets of actions $A_1, \ldots, A_S \subset \mathcal{A}$, on the sets of the form

$$\omega(A_1, \cdots, A_S) := \bigcap_{1 \leq i \leq S} \left(\mathcal{A}_{s_i}^*\right)^{-1}(A_i) \subset \Delta_{\mathcal{S}}$$

that are non-empty, Γ takes a single value. Equivalently, for any $A_1, \cdots, A_S \subset \mathcal{A}$, if $\omega(A_1, \cdots, A_S) \neq \emptyset$, the forward image of $\omega(A_1, \cdots, A_S)$ is a single policy: $\Gamma\big(\omega(A_1, ..., A_S)\big) = \{\pi_{A_1, \cdots, A_S}\}$.

While being a technical condition, OAS is satisfied by practically all best response maps proposed by single-agent RL literature. We clarify with examples.

Example C.14 (Pure best response). Fix an ordering $a_1, a_2, \cdots, a_K$ on $\mathcal{A}$. Define $\Gamma_{det} : \Delta_{\mathcal{S}} \to \Pi$ so that

$$\Gamma_{det}(\mu)(a_k|s) = \begin{cases} 1, \text{ if } a_1, a_2, \cdots, a_{k-1} \notin \mathcal{A}_s^*(\mu), a_k \in \mathcal{A}_s^*(\mu) \\ 0, \text{ otherwise} \end{cases}.$$

Γ_{det} assigns probability 1 to the first optimal action in the ordering. Γ_{det} is OAS since the the action probabilities are completely determined by which actions are optimal.

Example C.15 (Uniform map to optimal actions). Define $\Gamma_{unif} : \Delta_{\mathcal{S}} \to \Pi$ so that $\Gamma_{unif}(\mu)(a|s) = \frac{1}{|\mathcal{A}_s^*(\mu)|}$ if $a \in \mathcal{A}_s^*(\mu)$, otherwise 0. That is, Γ_{unif} assigns equal probability to all optimal actions at every state s. Γ_{unif} is OAS, and is the limiting operator of the Boltzman policy of (Guo et al., 2019).

Example C.16 (Limit of regularized BR). Take h strongly concave as before, and consider the regularization function $\tau h(\cdot)$ for $\tau > 0$. Consider $\Gamma_\tau^h(\mu) := \arg\max_\pi V_{\tau h}(\cdot|\pi, \mu)$. While for fixed $\tau > 0$ the Γ_τ^{BR} is not a valid (unregularized) best response operator, one can take the limit $\tau \to 0$ and show that

$$\Gamma_{\tau\to 0}^h(\mu) := \lim_{\tau \to 0} \Gamma_\tau^h(\mu) = \arg\max_{\pi\in\Pi^*(\mu)} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t h(\pi(s_t)) \middle|_{\substack{s_0=\cdot \\ a_t\sim\pi(s_t) \\ s_{t+1}\sim P(\cdot|s_t,a_t,\mu)}}, \forall t \geq 0 \right].$$

which implies $\Gamma_{\tau\to 0}^h$ is a best response operator. By above, it is also OAS, since it is the unique optimal policy that maximizes the regularizer term on the restricted "sub"-MDP where at each state s the only available actions are $\mathcal{A}_s^*(\mu)$. That is, $\Gamma_{\tau\to 0}^h(\mu)$ depends only on the optimal action sets $\mathcal{A}_s^*(\mu)$.

Finally, with this definition in place, we present the main result of this subsection.

Lemma C.17 (Continuous, OAS Γ are constant)**.** *Let $\Gamma : \Delta_{\mathcal{S}} \to \Pi$ be a continuous, OAS map such that $\Gamma_{br}(\mu) \in \Pi^*(\mu)$ for all $\mu \in \Delta_{\mathcal{S}}$. Then, Γ is constant on Ω, i.e., for some $\pi_0 \in \Pi$, $\Gamma(\mu) = \pi_0, \forall \mu \in \Delta_{\mathcal{S}}$.*

Proof. $\Delta_{\mathcal{S}}$ is a connected set with the topology induced by $\|\cdot\|_1$, hence by continuity of Γ, $\Gamma(\Delta_{\mathcal{S}})$ must be a connected set in Π. There are only finitely many subsets of actions $A_1, A_2, \cdots, A_N \subset \mathcal{A}$, hence there are only finitely many subsets of Ω of the form $\omega(A_1, A_2, \cdots, A_N)$ and these form a partition of the domain Ω. Hence by OAS, the image $\operatorname{Im}\Gamma$ is a discrete set. A discrete connected set must be a singleton. □

The above lemma shows that blanket continuity assumptions of unregularized best response might be too strong in MFG, reducing the MFG problem to the learning of a constant policy. The OAS assumption of Lemma C.17 hints that simply treating best response as a single-agent RL problem will lead to operators that are not continuous (or operators that have exploding Lipschitz constants as smoothing is decreased to approximate the discontinuous best response, as in the case of $\Gamma^h_{\tau\to 0}$).

C.9. Regularization and Bias

In this subsection, we define the unregularized Nash equilibrium and quantify the relationship between the unregularized and regularized Nash equilibrium.

Lemma C.18 (Unregularized Value and MFG-NE)**.** *We define the expected unregularized mean-field reward for a population-policy pair $(\pi, \mu) \in \Pi \times \Delta_{\mathcal{S}}$ as*

$$V(\pi, \mu) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, \mu) \middle| \substack{s_0 \sim \mu \\ a_t \sim \pi(s_t) \\ s_{t+1} \sim P(\cdot|s_t, a_t, \mu)}\right].$$

A pair $(\pi^*, \mu^*) \in \Pi \times \Delta_{\mathcal{S}}$ is called an unregularized MFG-NE if the following hold:

$$\text{Stability: } \mu^*(s) = \sum_{s', a'} \mu^*(s')\pi^*(a'|s')P(s|s', a', \mu^*),$$
$$\text{Optimality: } V(\pi^*, \mu^*) = \max_{\pi} V(\pi, \mu^*).$$

If the optimality condition is only satisfied with $V(\pi^*_\delta, \mu^*_\delta) \geq \max_\pi V(\pi, \mu^*_\delta) - \delta$, we call $(\pi^*_\delta, \mu^*_\delta)$ an unregularized δ-MFG-NE.

In general, the unregularized MFG-NE will not be unique. As expected, the regularized MFG-NE pair (π^*, μ^*) also forms an unregularized δ-MFG-NE. We quantify the bias in a straightforward manner in the following lemma.

Lemma C.19 (Regularization Bias on NE)**.** *Let $h : \Delta_{\mathcal{A}} \to [0, h_{max}]$ for some $h_{max} > 0$ and let $(\pi^*_\delta, \mu^*_\delta)$ be a regularized δ-MFG-NE with the regularizer h. Then, $(\pi^*_\delta, \mu^*_\delta)$ is an unregularized $(\delta + \frac{h_{max}}{1-\gamma})$-MFG-NE.*

Proof. The stability conditions for regularized and unregularized MFG-NE are identical. For the optimality condition, we note that for any pair $(\pi, \mu) \in \Pi \times \Delta_{\mathcal{S}}$, $|V(\pi, \mu) - V_h(\pi, \mu)| \leq \frac{h_{max}}{1-\gamma}$. It follows that for any $\mu \in \Delta_{\mathcal{S}}$, $|\max_\pi V(\pi, \mu) - \max_\pi V_h(\pi, \mu)| \leq \frac{h_{max}}{1-\gamma}$. □

For instance, using the scaled entropy regularizer $\tau h_{ent}(u) = -\tau \sum_a u(a) \log u(a)$ for $\tau > 0$, the bias will be bounded by $\frac{\tau \log |\mathcal{A}|}{1-\gamma}$.

D. Sample Based Learning

D.1. Conditions on h for Persistence of Excitation

The learning algorithms presented in this paper assume that persistence of excitation holds throughout training. We show that for many choices of h, the persistence of excitation (PE) assumption (Assumption 3) is automatically satisfied.

Lemma D.1 (PE conditions on h)**.** *Assume that h is strongly concave with modulus ρ, differentiable in $\Delta^\circ_{\mathcal{A}}$, $u_{max} \in \Delta^\circ_{\mathcal{A}}$, and define $U_\delta := \{u \in \Delta^\circ_{\mathcal{A}} : d(u, \partial\Delta_{\mathcal{A}}) < \delta\}$. Further assume that*

$$\lim_{\delta\to 0} \inf_{u\in U_\delta} \nabla h(u)^\top (u_{max} - u) > Q_{max} + \frac{4}{\eta}.$$

Then, there exists $p_{inf} > 0$ such that for all $q \in \mathcal{Q}, 0 \le q(\cdot,\cdot) \le Q_{max}, \pi \in \Pi$, it holds that $\Gamma^\eta_{md}(q,\pi)(a|s) > p_{inf}$ for all $s \in \mathcal{S}, a \in \mathcal{A}$.

Proof. $\mathcal{U}_{L_h}$ is a closed convex set (see Definition 3.5 and Eq. (1)). If $\mathcal{U}_{L_h} \cap \partial\Delta_{\mathcal{A}} = \varnothing$, we are done, since the image of Γ^η_{md} is a compact set and $\Gamma^\eta_{md}(s)$ is contained in $\Delta^\circ_{\mathcal{A}}$ for all $s \in \mathcal{S}$. So we assume $\mathcal{U}_{L_h} \cap \partial\Delta_{\mathcal{A}} \neq \varnothing$.

Denote $\mathcal{Q}' := \{q \in \mathcal{Q} : 0 \le q(\cdot,\cdot) \le Q_{max}\}$. Let $\bar{u} \in \mathcal{U}_{L_h} \cap \partial\Delta_{\mathcal{A}}$ and $q \in \mathcal{Q}', \pi \in \Pi$ arbitrary. Define the functions $f : \mathcal{U}_{L_h} \to \mathbb{R}$ and $g : [0,1] \to \mathbb{R}$ as

$$f(u) := \langle u, q(s,\cdot)\rangle + h(u) - \frac{1}{2\eta}\|u - \pi(s)\|_2^2, \qquad g(t) := f(\bar{u} + t(u_{max} - \bar{u})).$$

Here $\bar{u} + t(u_{max} - \bar{u}) \in \mathcal{U}_{L_h}$ for all $t \in [0,1]$ by convexity of $\mathcal{U}_{L_h}$ and the fact that $u_{max} \in \mathcal{U}_{L_h}$. We will show that $f(\bar{u})$ can not be a maximum in $\mathcal{U}_{L_h}$ by proving $g(0)$ has a direction of increase. Since g is differentiable in $(0,1)$ and continuous in $[0,1]$, for any $\varepsilon > 0$ by the mean value theorem there exists $\bar{\varepsilon} \in (0,\varepsilon)$ such that $g'(\bar{\varepsilon}) = \frac{g(\varepsilon)-g(0)}{\varepsilon}$. It is sufficient to show that for small enough ε, $g'(\bar{\varepsilon}) > 0$ for any $\bar{\varepsilon} \in (0,\varepsilon)$. Computing the derivatives, we obtain

$$\begin{aligned} g'(\varepsilon) =& \nabla f(\bar{u} + \varepsilon(u_{max} - \bar{u}))^\top (u_{max} - \bar{u}) \\ =& q(s,\cdot)^\top (u_{max} - \bar{u}) + \nabla h(\bar{u} + \varepsilon(u_{max} - \bar{u}))^\top (u_{max} - \bar{u}) \\ &- \frac{1}{\eta}(\bar{u} + \varepsilon(u_{max} - \bar{u}) - \pi(s))^\top (u_{max} - \bar{u}). \end{aligned}$$

Note that the magnitudes of the first and last terms can be bounded by

$$\left|q(s,\cdot)^\top (u_{max} - \bar{u})\right| \le Q_{max}, \qquad \left|\frac{1}{\eta}(\bar{u} + \varepsilon(u_{max} - \bar{u}) - \pi(s))^\top (u_{max} - \bar{u})\right| \le \frac{4}{\eta}.$$

Hence, choosing $\varepsilon > 0$ small enough so that $\nabla h(\bar{u} + \bar{\varepsilon}(u_{max} - \bar{u}))^\top (u_{max} - \bar{u}) > Q_{max} + \frac{4}{\eta}$ for all $\bar{\varepsilon} \in (0,\varepsilon)$ shows that $g(\varepsilon) > g(0)$, implying $\bar{u}$ can not be a maximizer of f. This implies $\Gamma^\eta_{md}(q,\pi)(a|s) > 0$ for any $s, a \in \mathcal{S} \times \mathcal{A}$.

Since the domain set $\mathcal{Q}' \times \Pi$ is compact and Γ^η_{md} is continuous (see Lemma C.3), the image set $\Gamma^\eta_{md}(\mathcal{Q}', \Pi)(a|s)$ is compact for all $s, a \in \mathcal{S} \times \mathcal{A}$ and there exists $p_{inf} > 0$ such that $\Gamma^\eta_{md}(q,\pi)(a|s) \ge p_{inf}$ for all $q, \pi \in \mathcal{Q}' \times \Pi$ and $s, a \in \mathcal{S} \times \mathcal{A}$. □

The lemma above does not require differentiability of h at the boundary $\partial\Delta_{\mathcal{A}}$. For instance, for the entropy regularizer $h_{ent}(u) = -\sum_a u(a) \log u(a)$ the assumption of Lemma D.1 is satisfied for any learning rate $\eta > 0$, since $\lim_{\delta\to 0} \inf_{u\in U_\delta} \nabla h(u)^\top (u_{max} - u) = \infty$.

D.2. Generalized Monotone Variational Inequalities under Biased Markovian Noise

In this section, we generalize the results from Kotsalis et al. (2022) for conditional TD learning to incorporate non-vanishing bias in estimates of the operator with possibly non-Markovian sequences. We refer the reader to their work for the full presentation of their ideas, while we provide a self-contained proof incorporating bias.

Theorem D.2 (CTD under bias)**.** *Let $F : \mathcal{X} \to \mathbb{R}^D$ be an operator on some bounded set $\mathcal{X} \subset \mathbb{R}^d$. Let $\{\xi_t : t \in \mathbb{Z}_+\}$ be a random process on space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\xi_t \in \Xi \subset \mathbb{R}^n$ with probability 1. Let $\mathcal{F}_t = \sigma(\xi_1, \ldots, \xi_t)$ and $\widetilde{F} : \mathcal{X} \times \Xi \to \mathbb{R}^D$. Let $\|\cdot\|$ be a norm on $\mathbb{R}^D$ with dual $\|\cdot\|_*$, and V be a Bregman divergence satisfying $V(x,y) \ge \frac{1}{2}\|x-y\|^2$ for all $x, y \in \mathcal{X}$. We assume the following hold.*

A1. There exists (a unique) point $x^* \in X$ such that $\langle F(x^*), x - x^*\rangle \ge 0, \quad \forall x \in X$.

A2. F is "generalized" strongly monotone, hence for some $\mu > 0$,

$$\langle F(x), x - x^*\rangle \ge 2\mu V(x, x^*), \quad \forall x \in X. \tag{5}$$

A3. F is Lipschitz with constant L, so that $\|F(x_1) - F(x_2)\|_* \leq L\|x_1 - x_2\|, \forall x_1, x_2 \in \mathcal{X}$.

A4. There exists $\sigma, \varsigma > 0$ such that for all iterates $\{x_t\}$ and $t, \tau \in \mathbb{Z}_+$,

$$\mathbb{E}\left[\left\|\widetilde{F}(x_t, \xi_{t+\tau}) - \mathbb{E}\left[\widetilde{F}(x_t, \xi_{t+\tau}) \mid \mathcal{F}_{t-1}\right]\right\|_*^2 \mid \mathcal{F}_{t-1}\right] \leq \frac{\sigma^2}{2} + \frac{\varsigma^2}{2}\|x_t - x^*\|^2. \tag{6}$$

A5. There exists $\delta > 0, C > 0, \rho \in (0, 1)$ such that $\forall t, \tau \in \mathbb{Z}_+$ almost surely

$$\left\|F(x) - \mathbb{E}\left[\widetilde{F}(x, \xi_{t+\tau}) \mid \mathcal{F}_{t-1}\right]\right\|_* \leq C\rho^\tau\|x - x^*\| + \delta, \forall x \in \mathcal{X}. \tag{7}$$

Then, for $\tau > \frac{\log 1/\mu + \log 20C}{\log 1/\rho}$, $t_0 = \max\left\{\frac{8L^2}{\mu^2}, \frac{16\varsigma^2}{\mu^2}\right\}$, $\beta_k = \frac{2}{\mu(t_0+k-1)}$, the iterations given by arbitrary $x_1 \in \mathcal{X}$ and

$$x_{k+1} = \arg\min_{x\in\mathcal{X}} \beta_k\left\langle\widetilde{F}(x_k, \xi_{k\tau}), x\right\rangle + V(x_k, x), \forall k \in \mathbb{Z}_+, \tag{8}$$

satisfy

$$\mathbb{E}[V(x_{k+1}, x^*)] \leq \frac{2(t_0+1)(t_0+2)V(x_1, x^*)}{(k+t_0)(k+t_0+1)} + \frac{6k(\sigma^2 + 4\delta^2)}{\mu^2(k+t_0)(k+t_0+1)} + \frac{100\delta^2}{\mu^2}.$$

Proof. Firstly, we note that by the triangle inequality and application of Young's inequality, for any sequence $\{x_t\}_t$ we have that

$$\begin{aligned}\|F(x_t) - \widetilde{F}(x_t, \xi_{t+\tau})\|_*^2 \leq& 2\|F(x_t) - \mathbb{E}[\widetilde{F}(x_t, \xi_{t+\tau}) \mid \mathcal{F}_{t-1}]\|_*^2 \\ &+ 2\|\mathbb{E}[\widetilde{F}(x_t, \xi_{t+\tau}) \mid \mathcal{F}_{t-1}] - \widetilde{F}(x_t, \xi_{t+\tau})\|_*^2.\end{aligned}$$

By taking expectations, applying Equations 7 and 6 and Youngs inequality, we can conclude (similar to Lemma 2.1 of Kotsalis et al. (2022) apart from the bias term) that

$$\mathbb{E}[\|F(x_t) - \widetilde{F}(x_t, \xi_{t+\tau})\|_*^2] \leq \sigma^2 + 4\delta^2 + (4C^2\rho^{2\tau} + \varsigma^2)\mathbb{E}[\|x_t - x^*\|^2]. \tag{9}$$

Now for the sequence of updates $\{x_k\}$ defined by Equation 8, as in Kotsalis et al. (2022), we define

$$\Delta F_k := F(x_k) - F(x_{k-1}) \text{ and } \delta_k^\tau := \widetilde{F}(x_k, \xi_{k\tau}) - F(x_k).$$

Reiterating Proposition 3.7 of (Kotsalis et al., 2022), by optimality conditions of Equation 8 it holds again that for all $x \in \mathcal{X}$,

$$\beta_k\langle F(x_{k+1}), x_{k+1} - x\rangle + \left(1 - 2\beta_k^2L^2\right)V(x_{k+1}, x) + \beta_k\langle\delta_k^\tau, x_k - x\rangle \leq V(x_k, x) + \beta_k^2\|\delta_k^\tau\|_*^2.$$

Fixing $x = x^*$ and using the strong generalized monotonicity condition of (5), we obtain

$$\begin{aligned}(1 + 2\mu\beta_k - 2\beta_k^2L^2)\mathbb{E}[V(x_{k+1}, x^*)] \leq& \mathbb{E}[V(x_k, x^*)] + \beta_k^2\mathbb{E}[\|\delta_t^\tau\|_*^2] + \beta_k\mathbb{E}[|\langle\delta_k^\tau, x_k - x^*\rangle|] \\ \leq& \mathbb{E}[V(x_k, x^*)] + \beta_k^2\mathbb{E}[\|\delta_t^\tau\|_*^2] + \beta_k\mathbb{E}[\|\delta_k^\tau\|_*\|x_k - x^*\|] \\ \leq& \mathbb{E}[V(x_k, x^*)] + \beta_k^2(\sigma^2 + 4\delta^2) + \delta\beta_k\mathbb{E}[\|x_k - x^*\|] \\ &+ (4C^2\rho^{2\tau}\beta_k^2 + \varsigma^2\beta_k^2 + C\beta_k\rho^\tau)\,\mathbb{E}[\|x_k - x^*\|^2],\end{aligned}$$

where the last inequality holds by (9) and (7). Hence by the property of divergence V:

$$\begin{aligned}(1 + 2\mu\beta_k - 2\beta_k^2L^2)\mathbb{E}[V(x_{k+1}, x^*)] \leq& \mathbb{E}[V(x_k, x^*)](1 + 8C^2\rho^{2\tau}\beta_k^2 + 2\varsigma^2\beta_k^2 + 2C\beta_k\rho^\tau) \\ &+ \beta_k^2(\sigma^2 + 4\delta^2) + \beta_k\delta\mathbb{E}[\|x_k - x^*\|].\end{aligned}$$

Applying Young's inequality on the term $\beta_k\delta\mathbb{E}[\|x_k - x^*\|]$, we obtain:

$$\begin{aligned}(1 + 2\mu\beta_k - 2\beta_k^2L^2)\mathbb{E}[V(x_{k+1}, x^*)] \leq& \mathbb{E}[V(x_k, x^*)](1 + 8C^2\rho^{2\tau}\beta_k^2 + 2\varsigma^2\beta_k^2 + 2C\beta_k\rho^\tau) \\ &+ \beta_k^2(\sigma^2 + 4\delta^2) + \beta_k\frac{10\delta^2}{\mu} + \beta_k\frac{\mu}{10}\mathbb{E}[\|x_k - x^*\|^2].\end{aligned}$$

Hence, we conclude

$$\begin{aligned}(1+2\mu\beta_k-2\beta_k^2L^2)\mathbb{E}[V(x_{k+1},x^*)] \leq &\mathbb{E}[V(x_k,x^*)](1+8C^2\rho^{2\tau}\beta_k^2+2\varsigma^2\beta_k^2+2C\beta_k\rho^\tau+\frac{\beta_k\mu}{10})\\ &+\beta_k^2(\sigma^2+4\delta^2)+\beta_k\frac{10\delta^2}{\mu}.\end{aligned} \tag{10}$$

Finally, setting $\theta_k := (k+t_0)(k+t_0+1)$, we note that with the assumptions on τ, t_0,

$$\theta_k(1+8C^2\rho^{2\tau}\beta_k^2+2\varsigma^2\beta_k^2+2C\beta_k\rho^\tau+\frac{\beta_k\mu}{10}) \leq \theta_{k-1}(1+2\mu\beta_{k-1}-2\beta_{k-1}^2L^2).$$

Multiplying both sides of (10) by θ_k and summing over $k=1,\dots,K$, we obtain

$$\begin{aligned}\theta_K(1+2\mu\beta_K-2\beta_K^2L^2)\,\mathbb{E}[V(x_{K+1},x^*)] \leq &\theta_1(1+8C^2\rho^{2\tau}\beta_1^2+2\varsigma^2\beta_1^2+2C\beta_1\rho^\tau)V(x_1,x)\\ &+\sum_{k=1}^K\theta_k\beta_k^2(\sigma^2+4\delta^2)+\sum_{k=1}^K\theta_k\beta_k\frac{10\delta^2}{\mu}.\end{aligned}$$

The result follows by computing the sums and $\sum_{k=1}^K\theta_k\beta_k \leq \frac{10}{\mu}(K+t_0)(K+t_0+1)$. □

D.3. Stochastic Population Update Bounds

We establish two useful results that (1) bound the expected deviation of the empirical population distribution from the mean field and (2) characterize mixing in terms of the state visitation probabilities of each agent. Firstly, we show that the empirical population distribution $\widehat{\mu}_t$ approximates the mean field $\Gamma_{pop}^\infty(\pi)$ in expectation up to a bias scaling with $\mathcal{O}(\frac{1}{\sqrt{N}})$.

Lemma D.3 (Empirical population bound)**.** *Assume that at any time $t \geq 0$, each agent i follows a given (arbitrary) policy $\pi^i \in \Pi$ so that,*

$$a_t^i \sim \pi^i(s_t^i), \quad s_{t+1}^i \sim P(\cdot|s_t^i, a_t^i, \widehat{\mu}_t), \quad \forall t \geq 0, i=1,\dots,N.$$

Let $\bar{\pi} \in \Pi$ be an arbitrary policy and $\Delta_{\bar{\pi}} := \frac{1}{N}\sum_i \|\bar{\pi}-\pi^i\|_1$. For any $\tau, t \geq 0$, it holds that

$$\mathbb{E}\left[\left\|\widehat{\mu}_{t+\tau}-\Gamma_{pop}^\tau(\widehat{\mu}_t,\bar{\pi})\right\|_1\middle|\mathcal{F}_t\right] \leq \frac{1-L_{pop,\mu}^\tau}{1-L_{pop,\mu}}\left(\frac{\sqrt{2|\mathcal{S}|}}{\sqrt{N}}+\frac{\Delta_{\bar{\pi}}K_a}{2}\right).$$

Proof. For the bias term when $\tau=1$, we compute:

$$\mathbb{E}\left[\widehat{\mu}_{t+1}|\mathcal{F}_t\right] = \mathbb{E}\left[\frac{1}{N}\sum_{s'\in\mathcal{S}}\sum_{i=1}^N \mathbb{1}(s_{t+1}^i=s')\mathbf{e}_{s'}\middle|\mathcal{F}_t\right] = \sum_{s'\in\mathcal{S}}\mathbf{e}_{s'}\sum_{i=1}^N\frac{1}{N}\bar{P}(s'|s_t^i,\pi^i(s_t^i),\widehat{\mu}_t),$$

where we use the $\mathcal{F}_t$-measurability of $\widehat{\mu}_t$ and s_t^i. We then obtain

$$\begin{aligned}\|\Gamma_{pop}(\pi,\widehat{\mu}_t)-\mathbb{E}\left[\widehat{\mu}_{t+1}|\mathcal{F}_t\right]\|_1 = &\left\|\sum_{s'\in\mathcal{S}}\mathbf{e}_{s'}\sum_{i=1}^N\frac{1}{N}(\bar{P}(s'|s_t^i,\pi^i(s_t^i),\widehat{\mu}_t)-\bar{P}(s'|s_t^i,\bar{\pi}(s_t^i),\widehat{\mu}_t))\right\|_1\\ \leq &\frac{1}{N}\sum_{i=1}^N\left\|(\bar{P}(\cdot|s_t^i,\pi^i(s_t^i),\widehat{\mu}_t)-\bar{P}(\cdot|s_t^i,\bar{\pi}(\cdot|s_t^i),\widehat{\mu}_t))\right\|_1 \leq \frac{K_a\Delta_{\bar{\pi}}}{2}.\end{aligned}$$

We also compute the variance at timestep $t+1$. For $\tau=1$, by independence, we can decompose the ℓ_2-variance:

$$\mathbb{E}[\|\widehat{\mu}_{t+1}-\mathbb{E}[\widehat{\mu}_{t+1}|\mathcal{F}_t]\|_2^2|\mathcal{F}_t] = \frac{1}{N^2}\sum_{i=1}^N\mathbb{E}[\|\mathbf{e}_{s_{t+1}^i}-\mathbb{E}[\mathbf{e}_{s_{t+1}^i}|\mathcal{F}_t]\|_2^2|\mathcal{F}_t] \leq \frac{2}{N},$$

where we use the fact that $\|\mathbf{e}_{s^i_{t+1}} - \mathbb{E}[\mathbf{e}_{s^i_{t+1}}|\mathcal{F}_t]\|_2^2 \leq 2$. In particular, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{\mu}_{t+1} - \mathbb{E}\left[\widehat{\mu}_{t+1}|\mathcal{F}_t\right]\|_1|\mathcal{F}_t\right] &= \sqrt{\mathbb{E}\left[\|\widehat{\mu}_{t+1} - \mathbb{E}\left[\widehat{\mu}_{t+1}|\mathcal{F}_t\right]\|_1|\mathcal{F}_t\right]^2} \\
&\leq \sqrt{|\mathcal{S}|\,\mathbb{E}\left[\|\widehat{\mu}_{t+1} - \mathbb{E}\left[\widehat{\mu}_{t+1}|\mathcal{F}_t\right]\|_2^2|\mathcal{F}_t\right]} \\
&\leq \frac{\sqrt{2|\mathcal{S}|}}{\sqrt{N}}
\end{aligned}$$

using Jensen's inequality and the fact that $\|v\|_1 \leq \sqrt{d}\|v\|_2$ for any $v \in \mathbb{R}^d$. Hence we have

$$\begin{aligned}
\mathbb{E}[\|\widehat{\mu}_{t+1} - \Gamma_{pop}(\widehat{\mu}_t, \bar{\pi})\|_1|\mathcal{F}_t] \leq& \mathbb{E}[\|\widehat{\mu}_{t+1} - \mathbb{E}\left[\widehat{\mu}_{t+1}|\mathcal{F}_t\right]\|_1|\mathcal{F}_t] + \mathbb{E}[\|\,\mathbb{E}\left[\widehat{\mu}_{t+1}|\mathcal{F}_t\right] - \Gamma_{pop}(\widehat{\mu}_t, \bar{\pi})\|_1|\mathcal{F}_t] \\
\leq& \frac{\sqrt{2|\mathcal{S}|}}{\sqrt{N}} + \frac{K_a}{2}\Delta_{\bar{\pi}}.
\end{aligned}$$

For $\tau > 1$, we inductively generalize the result. By law of iterated expectations (and the fact that $\mathcal{F}_t \subset \mathcal{F}_{t+\tau}$), we can derive

$$\begin{aligned}
\mathbb{E}\left[\left\|\widehat{\mu}_{t+\tau+1} - \Gamma_{pop}^{\tau+1}(\widehat{\mu}_t, \pi)\right\|_1\middle|\mathcal{F}_t\right] \leq& \mathbb{E}\left[\|\widehat{\mu}_{t+\tau+1} - \Gamma_{pop}(\widehat{\mu}_{t+\tau}, \bar{\pi})\|_1|\mathcal{F}_t\right] \\
&+ \mathbb{E}\left[\left\|\Gamma_{pop}(\widehat{\mu}_{t+\tau}, \pi) - \Gamma_{pop}^{\tau+1}(\widehat{\mu}_t, \pi)\right\|_1\middle|\mathcal{F}_t\right] \\
\leq& \mathbb{E}\left[\mathbb{E}\left[\|\widehat{\mu}_{t+\tau+1} - \Gamma_{pop}(\widehat{\mu}_{t+\tau}, \pi)\|_1|\mathcal{F}_{t+\tau}\right]|\mathcal{F}_t\right] \\
&+ \mathbb{E}\left[L_{pop,\mu}\left\|\widehat{\mu}_{t+\tau} - \Gamma_{pop}^{\tau}(\widehat{\mu}_t, \pi)\right\|_1\middle|\mathcal{F}_t\right] \\
\leq& \frac{\sqrt{2|\mathcal{S}|}}{\sqrt{N}} + \frac{\Delta_{\bar{\pi}}K_a}{2} + L_{pop,\mu}\,\mathbb{E}\left[\left\|\widehat{\mu}_{t+\tau} - \Gamma_{pop}^{\tau}(\widehat{\mu}_t, \pi)\right\|_1\middle|\mathcal{F}_t\right],
\end{aligned}$$

where the last inequality follows from what has been proven above for $\tau = 1$ and the Lipschitz continuity of the operator $\Gamma_{pop}(\cdot, \pi)$.

Hence, we inductively obtain the bound for any $\tau > 0$,

$$\mathbb{E}\left[\left\|\widehat{\mu}_{t+\tau} - \Gamma_{pop}^{\tau}(\widehat{\mu}_t, \pi)\right\|_1\middle|\mathcal{F}_t\right] \leq \frac{1 - L_{pop,\mu}^{\tau}}{1 - L_{pop,\mu}}\left(\frac{\sqrt{2|\mathcal{S}|}}{\sqrt{N}} + \frac{\Delta_{\bar{\pi}}K_a}{2}\right).$$

□

The dependence on $\Delta_{\bar{\pi}}$ above indicates that if the policies of agents deviate from each other, an additional bias will be incurred on the empirical population distribution. In the centralized learning case, we will have $\Delta_{\bar{\pi}} = 0$, whereas the term will be significant for the independent learning case due to the variance in agents' policy updates. As corollaries, we have the following bounds in terms of the limiting population distribution and the empirical population bound conditioned on the state of a single agent.

Corollary D.4 (Convergence to stable distribution)**.** *Under the conditions of Lemma D.3 for any $t, \tau \geq 0$, we have*

$$\mathbb{E}\left[\left\|\widehat{\mu}_{t+\tau} - \Gamma_{pop}^{\infty}(\bar{\pi})\right\|_1\middle|\mathcal{F}_t\right] \leq \frac{1}{1 - L_{pop,\mu}}\left(\frac{\sqrt{2|\mathcal{S}|}}{\sqrt{N}} + \frac{K_a\Delta_{\bar{\pi}}}{2}\right) + 2L_{pop,\mu}^{\tau}.$$

Proof. The proof follows from an application of triangle inequality and Lemma D.3. □

Corollary D.5. *Assume the conditions of Lemma D.3 and Assumption 4. Then, for any $\bar{s} \in \mathcal{S}$, agent $j \in [N]$, and for $\tau > T_{mix}$, we have*

$$\mathbb{E}\left[\left\|\widehat{\mu}_{t+\tau} - \Gamma_{pop}^{\tau}(\widehat{\mu}_t, \pi)\right\|_1\middle|s_{t+\tau}^j = \bar{s}, \mathcal{F}_t\right] \leq \frac{1 - L_{pop,\mu}^{\tau}}{(1 - L_{pop,\mu})\delta_{mix}}\left(\frac{\sqrt{2|\mathcal{S}|}}{\sqrt{N}} + \frac{\Delta_{\bar{\pi}}K_a}{2}\right).$$

Proof. The corollary follows from the law of total expectation and the fact that by mixing conditions, $\mathbb{P}[s^j_{t+\tau} = \bar{s}|\mathcal{F}_t] > \delta_{mix}$. □

As a stronger result, we prove that any agent's state visitation probabilities also converge to the mean field up to a population bias term. This result generalizes similar mixing theorems for Markov chains to the case where there is time dependence due to population dynamics and later facilitates proving learning bounds on a single sample path. The proof is loosely based on the ideas from Theorem 4.9 of (Levin & Peres, 2017), with the additional complication that the transitions are not homogeneous due to population effects.

Proposition D.6 (Mean field MC convergence)**.** *Let $\{s^i_0\}_i \in \mathcal{S}^N$ be (arbitrary) initial states of each agent. Assume that each player follows a policy $\pi^i \in \Pi$, that is,*

$$a^i_t \sim \pi^i(s^i_t), \quad s^i_{t+1} \sim P(\cdot|s^i_t, a^i_t, \widehat{\mu}_t), \quad \forall t \geq 0, i = 1, \dots, N.$$

Let $\bar{\pi} \in \Pi$ be an arbitrary policy and $\Delta_{\bar{\pi}} := \frac{1}{N}\sum_i \|\bar{\pi} - \pi^i\|_1$. For any $T \geq 0$ and for any $i = 1, \dots, N$, it holds that

$$\begin{aligned}\| \mathbb{P}[s^i_T = \cdot] - \Gamma^\infty_{pop}(\bar{\pi})\|_1 \leq & C_{mix}\rho^T_{mix} + \frac{2K_\mu T_{mix}}{\delta^2_{mix}} \frac{\sqrt{2|\mathcal{S}|}}{\sqrt{N}} \\ & + \frac{T_{mix}K_a K_\mu}{\delta^2_{mix}(1 - L_{pop,\mu})}\Delta_{\bar{\pi}} + \frac{T_{mix}K_a}{\delta_{mix}}\|\pi^i - \bar{\pi}\|_1,\end{aligned}$$

where $\rho_{mix} := \max\left\{L_{pop,\mu}, (1-\delta_{mix})^{1/T_{mix}}\right\}$ and $C_{mix} := \frac{4T_{mix}\max\{K_\mu, 1\}}{\delta_{mix}\rho^{T_{mix}}_{mix}|L_{pop,\mu} - (1-\delta_{mix})^{1/T_{mix}}|}$.

Proof. Denote $\mu_\infty = \Gamma^\infty_{pop}(\pi)$, and denote the transition matrix induced by the limiting population as $[\mathbf{P}_\infty]_{s,s'} = \bar{P}(s'|s, \bar{\pi}(s), \mu_\infty)$ and note that by definition, μ_∞ is the limiting distribution of the Markov chain induced by transition probabilities $[\mathbf{P}_\infty]_{s,s'}$. By the irreducability and aperiodicity implied by Assumption 4, it is in fact the unique stationary distribution of the stochastic matrix $\mathbf{P}_\infty$. Finally, by $\mathbf{M}_\infty$ denote the (stochastic) matrix with all rows equal to μ_∞. Then it holds that $\mathbf{X}\mathbf{M}_\infty = \mathbf{M}_\infty$ for *any* stochastic matrix $\mathbf{X}$, and $\mathbf{M}_\infty\mathbf{P}_\infty = \mathbf{M}_\infty$. For a sequence of matrices $\{\mathbf{A}_i\}_i$, while matrix multiplication is not commutative in general, in this proof we denote $\prod_{i=1}^I \mathbf{A}_i := \mathbf{A}_1\mathbf{A}_2 \dots \mathbf{A}_I$ for simplicity.

We prove the result for the first agent ($i = 1$), from which the theorem will follow by symmetry. By $\mathbf{P}_t$ denote the stochastic transition matrix at time t given by $[\mathbf{P}_t]_{s,s'} = \mathbb{P}(s^1_t = s'|s^1_{t-1} = s)$. By Assumption 4, there exists a $T_{mix} > 0$ such that for some $\delta_{mix} > 0$, the matrix defined by $\mathbf{P}^{(j)} = \prod_{t=(j-1)T_{mix}+1}^{jT_{mix}} \mathbf{P}_t$ satisfies for all j that $[\mathbf{P}^{(j)}]_{s,s'} > \delta_{mix}\mu_\infty(s') > 0$. Hence for each j, we define the stochastic matrices $\mathbf{Q}^{(j)}$ implicitly given by the equation $\mathbf{P}^{(j)} = (1-\theta)\mathbf{M}_\infty + \theta\mathbf{Q}^{(j)}$. where $\theta := 1 - \delta_{mix}$. The proof will follow in three steps.

Step 1: Induction on j. By induction, we will prove that for all $J > 0$, the following holds:

$$\prod_{j=1}^J \mathbf{P}^{(j)} = (1-\theta^J)\mathbf{M}_\infty + \theta^J \prod_{j=1}^J \mathbf{Q}^{(j)} + \sum_{l=2}^J (1-\theta^{l-1})\theta^{J-l}\mathbf{M}_\infty \left(\mathbf{P}^{(j)} - \mathbf{P}^{T_{mix}}_\infty\right) \prod_{l'=l+1}^J \mathbf{Q}^{(l')}. \tag{11}$$

The identity can be verified in a straightforward manner for $J = 1, 2$. Assuming the identity holds for $J > 1$, we show it holds for $J + 1$. Denoting $\triangle := (\prod_{j=1}^J \mathbf{P}^{(j)})\mathbf{P}^{(J+1)}$, and distributing over the equality for the inductive assumption on J, we obtain

$$\begin{aligned}\triangle =& (1-\theta^J)\mathbf{M}_\infty\mathbf{P}^{(J+1)} + \theta^J \prod_{j=1}^J \mathbf{Q}^{(j)} \left((1-\theta)\mathbf{M}_\infty + \theta\mathbf{Q}^{(J+1)}\right) \\ & + \left(\sum_{l=2}^J (1-\theta^{l-1})\theta^{J-j}\mathbf{M}_\infty \left(\mathbf{P}^{(j)} - \mathbf{P}_\infty\right) \prod_{l'=l+1}^J \mathbf{Q}^{(l')}\right) \left((1-\theta)\mathbf{M}_\infty + \theta\mathbf{Q}^{(J+1)}\right).\end{aligned}$$

We observe that $\prod_{j=1}^{J} \mathbf{Q}^{(j)}$ is a stochastic matrix and that $\mathbf{M}_\infty \mathbf{P}_\infty = \mathbf{M}_\infty$ to obtain:

$$
\begin{aligned}
=&(1-\theta^J)\mathbf{M}_\infty \left(\mathbf{P}^{(J+1)} - \mathbf{P}_\infty^{T_{mix}}\right) + (1-\theta^J)\mathbf{M}_\infty \mathbf{P}_\infty^{T_{mix}} + \theta^J(1-\theta)\mathbf{M}_\infty + \theta^{J+1}\prod_{j=1}^{J}\mathbf{Q}^{(j)} \\
&+ \left(\sum_{l=2}^{J}(1-\theta^{l-1})\theta^{J-j}\mathbf{M}_\infty\left(\mathbf{P}^{(j)} - \mathbf{P}_\infty^{T_{mix}}\right)\prod_{l'=l+1}^{J}\mathbf{Q}^{(l')}\right)\left((1-\theta)\mathbf{M}_\infty + \theta\mathbf{Q}^{(J+1)}\right) \\
=&(1-\theta^J)\mathbf{M}_\infty \left(\mathbf{P}^{(J+1)} - \mathbf{P}_\infty^{T_{mix}}\right) + (1-\theta^{J+1})\mathbf{M}_\infty + \theta^{J+1}\prod_{j=1}^{J}\mathbf{Q}^{(j)} \\
&+ \left(\sum_{j=2}^{J}(1-\theta^{l-1})\theta^{J-j}\mathbf{M}_\infty\left(\mathbf{P}^{(j)} - \mathbf{P}_\infty^{T_{mix}}\right)\prod_{l'=l+1}^{J}\mathbf{Q}^{(l')}\right)\left((1-\theta)\mathbf{M}_\infty + \theta\mathbf{Q}^{(J+1)}\right).
\end{aligned}
$$

The result follows by the identity

$$
\left(\sum_{j=2}^{J}(1-\theta^{l-1})\theta^{J-j}\mathbf{M}_\infty\left(\mathbf{P}^{(j)} - \mathbf{P}_\infty^{T_{mix}}\right)\prod_{l'=l+1}^{J}\mathbf{Q}^{(l')}\right)(1-\theta)\mathbf{M}_\infty = 0
$$

since both $\mathbf{P}^{(j)}\prod_{l'=j+1}^{J}\mathbf{Q}^{(l')}$ and $\mathbf{P}_\infty\prod_{l'=j+1}^{J}\mathbf{Q}^{(l')}$ are stochastic matrices.

Step 2: Quantifying convergence. Using the definition $\mathbf{P}^{(j)} = \prod_{t=(j-1)T_{mix}+1}^{jT_{mix}}\mathbf{P}_t$ we can rewrite (11) equivalently as

$$
\left(\prod_{t=1}^{JT_{mix}}\mathbf{P}_t\right) = (1-\theta^J)\mathbf{M}_\infty + \theta^J\prod_{j=1}^{J}\mathbf{Q}^{(j)} + \sum_{j=2}^{J}(1-\theta^{l-1})\theta^{J-j}\mathbf{M}_\infty\left(\mathbf{P}^{(j)} - \mathbf{P}_\infty\right)\prod_{l'=l+1}^{J}\mathbf{Q}^{(l')}.
$$

Multiplying both sides by $\mathbf{E}^{(r)} := \prod_{t=1}^{r}\mathbf{P}_{JT_{mix}+t}$ from the right for $r < T_{mix}$ and using $\mathbf{M}_\infty\mathbf{P}_\infty^r = \mathbf{M}_\infty$, we obtain

$$
\begin{aligned}
\prod_{t=1}^{JT_{mix}+r}\mathbf{P}_t - \mathbf{M}_\infty =&\theta^J\left(\prod_{j=1}^{J}\mathbf{Q}^{(j)}\mathbf{E}^{(r)} - \mathbf{M}_\infty\mathbf{E}^{(r)}\right) + \mathbf{M}_\infty\left(\mathbf{E}^{(r)} - \mathbf{P}_\infty^r\right) \\
&+ \sum_{j=2}^{J}(1-\theta^{l-1})\theta^{J-j}\mathbf{M}_\infty\left(\mathbf{P}^{(j)} - \mathbf{P}_\infty\right)\prod_{l'=l+1}^{J}\mathbf{Q}^{(l')}\mathbf{E}^{(r)}.
\end{aligned}
$$

Let $\mathbf{u} \in \Delta_{\mathcal{S}}$ be an arbitrary column vector. Multiplying each side by $\mathbf{u}^\top$ on the left, and taking the ℓ_1 norm, we have for $\square := \left\|\mathbf{u}^\top\prod_{t=1}^{JT_{mix}+r}\mathbf{P}_t - \boldsymbol{\mu}_\infty\right\|_1$,

$$
\begin{aligned}
\square \leq& 2\theta^J + \sum_{j=2}^{J}\theta^{J-j}\left\|\bar{\mathbf{u}}^\top\left(\mathbf{P}^{(j)} - \mathbf{P}_\infty\right)\bar{\mathbf{Q}}^{(j)}\right\|_1 + \left\|\bar{\mathbf{u}}^\top\left(\mathbf{E}^{(r)} - \mathbf{P}_\infty^r\right)\right\|_1 \\
\leq& 2\theta^J + \sum_{j=2}^{J}\theta^{J-j}\sum_{t=(j-1)T_{mix}+1}^{jT_{mix}}\sup_s\|\mathbf{P}_{t,s\cdot} - \mathbf{P}_{\infty,j\cdot}\|_1 + \sum_{t=jT_{mix}+1}^{jT_{mix}+r}\sup_s\|\mathbf{P}_{t,s\cdot} - \mathbf{P}_{\infty,j\cdot}\|_1,
\end{aligned}
$$

where $\mathbf{P}_{t,s\cdot}, \mathbf{P}_{\infty,s\cdot}$ indicate s-th row vectors of matrices $\mathbf{P}_t, \mathbf{P}_\infty$, and $\bar{\mathbf{u}} \in \Delta_{\mathcal{S}}$ and $\bar{\mathbf{Q}}^{(j)}$ is a stochastic matrix, using Lemma B.3 in the last line.

Step 3: Quantifying finite population bias. Finally, we will quantify the error due to the non-vanishing $\sup_s\|\mathbf{P}_{t,s\cdot} - \mathbf{P}_{\infty,s\cdot}\|_1$ terms. We denote by $\widehat{\Delta}_{N,\mathcal{S}} \subset \Delta_{\mathcal{S}}$ the (finite) set of possible empirical state distributions with N agents. Observe that for any $s \in \mathcal{S}$,

$$\begin{aligned}
\mathbf{P}_{t,s\cdot} - \mathbf{P}_{\infty,s\cdot} &= \mathbb{P}(s^1_{t+1} = \cdot | s^1_t = s) - P(\cdot|s, \bar{\pi}(s), \mu_\infty) \\
&\overset{\star}{=} \sum_{\mu \in \widehat{\Delta}_{N,\mathcal{S}}} \left(\mathbb{P}(s^1_{t+1} = \cdot | \widehat{\mu}_t = \mu, s^1_t = s) - P(\cdot|s, \bar{\pi}(s), \mu_\infty)\right) \mathbb{P}(\widehat{\mu}_t = \mu | s^1_t = s) \\
&\overset{\triangle}{=} \sum_{\mu \in \widehat{\Delta}_{N,\mathcal{S}}} \left(P(\cdot|\mu, \pi^1(s), s) - P(\cdot|s, \bar{\pi}(s), \mu_\infty)\right) \mathbb{P}(\widehat{\mu}_t = \mu | s^1_t = s),
\end{aligned}$$

where $(\star)$ follows from the law of total probability and $(\triangle)$ follows from the definition of SAGS dynamics. Assume $t \geq T_{mix}$. We then conclude

$$\begin{aligned}
\|\mathbf{P}_{t,s\cdot} - \mathbf{P}_{\infty,s\cdot}\|_1 \leq& \sum_{\mu \in \widehat{\Delta}_{N,\mathcal{S}}} \left\|P(\cdot|\mu, \pi^1(s), s) - P(\cdot|s, \bar{\pi}(s), \mu_\infty)\right\|_1 \mathbb{P}(\widehat{\mu}_t = \mu | s^1_t = s) \\
\leq& \sum_{\mu \in \widehat{\Delta}_{N,\mathcal{S}}} \left(K_\mu \|\mu_\infty - \mu\|_1 + \frac{K_a}{2}\|\pi^1 - \bar{\pi}\|_1\right) \mathbb{P}(\widehat{\mu}_t = \mu | s^1_t = s) \\
\leq& \frac{K_a}{2}\|\pi^1 - \bar{\pi}\|_1 + \frac{K_\mu}{\delta_{mix}} \mathbb{E}[\|\mu_\infty - \widehat{\mu}_t\|_1],
\end{aligned}$$

where the last line is a consequence of Corollary D.5. Using Lemma D.3, for $t \geq T_{mix}$,

$$\|\mathbf{P}_{t,s\cdot} - \mathbf{P}_{\infty,s\cdot}\|_1 \leq \left(\frac{K_\mu K_a}{(1 - L_{pop,\mu})\delta_{mix}}\right)\frac{\Delta_{\bar{\pi}}}{2} + \frac{K_a}{2}\|\pi^1 - \bar{\pi}\|_1 + \frac{K_\mu}{\delta_{mix}}\frac{\sqrt{2|\mathcal{S}|}}{\sqrt{N}} + \frac{2K_\mu}{\delta_{mix}} L^t_{pop,\mu}.$$

Placing this in the inequality we obtained from step 2, we obtain result. □

An intuitive detail above is that the mixing constant ρ_{mix} is equal to the maximum of the Markov chain mixing constant $(1 - \delta_{mix})^{1/T_{mix}}$ and the population mixing (or contraction) constant $L_{pop,\mu}$. Hence, the bound above suggests waiting for both the Markov chain *and* the population is essential. We use these convergence results to prove finite sample bounds for TD-learning in N-player SAGS in the next section.

D.4. CTD with Population

For the purpose of having a clearer presentation and clarifying the dependence on problem parameters, we define the problem-dependent constants (where L_h such that $\pi \in \Pi_{L_h}$):

$$\begin{aligned}
&\underline{\mathbf{M}}_{td} := \frac{\log 1/\mu_F + \log 40 C_{mix}}{\log 1/\rho_{mix}}, \qquad t_0 := \frac{16(1+\gamma)^2}{\mu_F^2}, \\
&C_1^{TD} := \frac{2(1+L_h)(t_0+2)|\mathcal{S}||\mathcal{A}|}{1-\gamma}, \qquad C_2^{TD} := \frac{16(1+L_h)}{(1-\gamma)^2\delta_{mix}p_{inf}}, \\
&C_{pop,1}^{TD} := \frac{20(K_\mu + L_\mu)(1+L_h)}{(1-\gamma)^2\delta_{mix}p_{inf}}, \qquad C_{pop,2}^{TD} := \frac{10(9K_\mu + L_\mu)(1+L_h)T_{mix}\sqrt{2|\mathcal{S}|}}{(1-L_{pop,\mu})(1-\gamma)^2\delta_{mix}^3 p_{inf}}, \\
&C_{pol,1}^{TD} := \frac{5T_{mix}(1+L_h)(K_\mu + L_\mu + 8K_aK_\mu)}{(1-L_{pop,\mu})(1-\gamma)^2\delta_{mix}^3 p_{inf}}, \qquad C_{pol,2}^{TD} := \frac{40(1+L_h)K_aT_{mix}}{(1-\gamma)^2\delta_{mix}^2 p_{inf}} + \frac{20C_h + 10L_h}{(1-\gamma)\delta_{mix}p_{inf}},
\end{aligned}$$

where C_h is defined as the Lipschitz constant of h on the set $\{u \in \Delta_\mathcal{A} : u(a) \geq p_{inf}, \forall a \in \mathcal{A}\}$ with respect to the $\|\cdot\|_1$, which is guaranteed to exists since ∇h is continuous on the compact set $\{u \in \Delta_\mathcal{A} : u(a) \geq p_{inf}, \forall a \in \mathcal{A}\}$. We provide the full proof and restatement of Theorem 4.2.

Theorem D.7 (CTD learning with population)**.** *Assume Assumption 4 holds and let policies $\{\pi^i\}_i$ be given so that $\pi^i(a|s) \geq p_{inf}$ for all i. Assume Algorithm 1 is run with policies $\{\pi^i\}_i$,* arbitrary *initial agent states $\{s^i_0\}_i$, learning rates $\beta_m = \frac{2}{(1-\gamma)(t_0+m-1)}, \forall m \geq 0$ and any $M > 0$, $M_{td} > \underline{\mathbf{M}}_{td}$. If $\bar{\pi} \in \Pi$ is an arbitrary policy, $\Delta_{\bar{\pi}} := \frac{1}{N}\sum_i \|\pi^i - \bar{\pi}\|_1$ and*

$Q^* := Q_h(\cdot,\cdot|\bar{\pi},\mu_{\bar{\pi}})$, then the (random) output $\widehat{Q}_M$ of Algorithm 1 satisfies

$$\begin{aligned}\mathbb{E}[\|\widehat{Q}_M - \widehat{Q}^*\|_\infty] \leq & \frac{C_1^{TD}}{\sqrt{(M+t_0)(M+t_0+1)}} + \frac{C_2^{TD}\sqrt{M}}{\sqrt{(M+t_0)(M+t_0+1)}} + C_{pop,1}^{TD} L_{pop,\mu}^{M_{td}} \\ & + \frac{C_{pop,2}^{TD}}{\sqrt{N}} + C_{pol,1}^{TD}\Delta_{\bar{\pi}} + C_{pol,2}^{TD}\|\pi^1 - \bar{\pi}\|_1.\end{aligned}$$

Proof. We verify the assumptions of Theorem D.2. Take the divergence $V(q,q') = \frac{1}{2}\|q-q'\|_2^2$, and denote $\zeta_\alpha^m := \zeta_{mM_{td}+\alpha}, \mathcal{F}_\alpha^m := \mathcal{F}_{mM_{td}+\alpha}$. Assumptions A1-A3 have been shown, with generalized strong monotonicity modulus $\mu_F := (1-\gamma)\delta_{mix}p_{inf}$ and Lipschitz constant $L_F := (1+\gamma)$. For A4, we bound the variance of $\widetilde{F}$ by (see Lan (2022)):

$$\mathbb{E}\left[\|\widetilde{F}^{\pi^1}(\widehat{Q}_m,\zeta_\alpha^m) - \mathbb{E}[\widetilde{F}^{\pi^1}(\widehat{Q}_m,\zeta_\alpha^m)|\mathcal{F}_0^m]\|_2^2|\mathcal{F}_0^m\right] \leq 4(1+\gamma)^2\,\mathbb{E}[\|\widehat{Q}_m - Q^*\|_2^2] + \frac{4(1+L_h)^2}{(1-\gamma)^2}.$$

During CTD learning, the iterates $\widehat{Q}_m$ stay in the set $[\frac{h_{max}-L_h}{1-\gamma}, Q_{max}]$ since $\widehat{Q}_0(s,a) = Q_{max}$ at initialization.

Finally, we verify the (more challenging) mixing condition. As before, let $\mu_\pi := \Gamma_{pop}^\infty(\bar{\pi})$. For any $Q \in \mathcal{Q}$, we have

$$\begin{aligned}\mathbb{E}[\widetilde{F}^{\pi^1}(Q,\zeta_{t+\tau})|\mathcal{F}_t] =& \mathbb{E}[(Q(s_{t+\tau},a_{t+\tau}) - R(s_{t+\tau},a_{t+\tau},\mu_\pi) - h(\pi^1(s)) - \gamma Q(s_{t+\tau+1},a_{t+\tau+1}))\mathbf{e}_{s_{t+\tau},a_{t+\tau}}|\mathcal{F}_t] \\ & + \mathbb{E}[(R(s_{t+\tau},a_{t+\tau},\mu_\pi) - R(s_{t+\tau},a_{t+\tau},\widehat{\mu}_{t+\tau}))\mathbf{e}_{s_{t+\tau},a_{t+\tau}}|\mathcal{F}_t].\end{aligned}$$

By Lemma D.3, the second term can be bounded by

$$\|\,\mathbb{E}[(R(s_{t+\tau},a_{t+\tau},\mu_\pi) - R(s_{t+\tau},a_{t+\tau},\widehat{\mu}_{t+\tau}))\mathbf{e}_{s_{t+\tau},a_{t+\tau}}|\mathcal{F}_t]\|_2 \leq \frac{L_\mu}{1-L_{pop,\mu}}\left(\frac{\sqrt{2|\mathcal{S}|}}{\sqrt{N}} + \frac{\Delta_{\bar{\pi}}K_a}{2}\right) + 2L_\mu L_{pop,\mu}^\tau.$$

For the first term, abbreviating $\mathbf{v}_{s,a}^{s',a'} := \pi^1(a|s)\pi^1(a'|s')(Q(s,a) - R(s,a,\mu_\pi) - h(\pi^1(s)) - \gamma Q(s',a'))\mathbf{e}_{s,a}$ and $\bar{\mathbf{v}}_{s,a}^{s',a'} := \bar{\pi}(a|s)\bar{\pi}(a'|s')(Q(s,a) - R(s,a,\mu_\pi) - h(\bar{\pi}(s)) - \gamma Q(s',a'))\mathbf{e}_{s,a}$, and again denoting by $\widehat{\Delta}_{N,\mathcal{S}} \subset \Delta_{\mathcal{S}}$ the (finite) set of possible empirical state distributions with N agents,

$$\begin{aligned}&\mathbb{E}[(Q(s_{t+\tau},a_{t+\tau}) - R(s_{t+\tau},a_{t+\tau},\mu_\pi) - h(\pi^1(s)) - \gamma Q(s_{t+\tau+1},a_{t+\tau+1}))\mathbf{e}_{s_{t+\tau},a_{t+\tau}}|\mathcal{F}_t] \\ &= \sum_{\substack{s,s',a,a' \\ \mu\in\widehat{\Delta}_{N,\mathcal{S}}}} \mathbb{P}[s_{t+\tau} = s|\mathcal{F}_t]\,\mathbb{P}[\widehat{\mu}_{t+\tau} = \mu|s_{t+\tau} = s,\mathcal{F}_t]P(s'|s,a,\mu)\mathbf{v}_{s,a}^{s',a'} \\ &= \underbrace{\sum_{\substack{s,s',a,a' \\ \mu\in\widehat{\Delta}_{N,\mathcal{S}}}} \mathbb{P}[s_{t+\tau} = s|\mathcal{F}_t]\,\mathbb{P}[\widehat{\mu}_{t+\tau} = \mu|s_{t+\tau} = s,\mathcal{F}_t]\left(P(s'|s,a,\mu) - P(s'|s,a,\mu_\pi)\right)\mathbf{v}_{s,a}^{s',a'}}_{(\star)} \\ &\quad + \underbrace{\sum_{s,s',a,a'} \mathbb{P}[s_{t+\tau} = s|\mathcal{F}_t]P(s'|s,a,\mu_\pi)(\mathbf{v}_{s,a}^{s',a'} - \bar{\mathbf{v}}_{s,a}^{s',a'})}_{(\square)} + \underbrace{\sum_{s,s',a,a'} \mathbb{P}[s_{t+\tau} = s|\mathcal{F}_t]P(s'|s,a,\mu_\pi)\bar{\mathbf{v}}_{s,a}^{s',a'}}_{(\triangle)}.\end{aligned}$$

We analyze the three terms above separately. For $(\star)$, observe the inequality using Corollary D.5:

$$\sum_{\mu\in\widehat{\Delta}_{N,\mathcal{S}}} \mathbb{P}[\widehat{\mu}_{t+\tau} = \mu|s_{t+\tau} = s,\mathcal{F}_t]\,\|P(\cdot|s,a,\mu) - P(\cdot|s,a,\mu_\pi)\|_1 \leq \frac{K_\mu}{(1-L_{pop,\mu})\delta_{mix}}\left(\frac{\sqrt{2|\mathcal{S}|}}{\sqrt{N}} + \frac{\Delta_{\bar{\pi}}K_a}{2}\right) + 2K_\mu L_{pop,\mu}^\tau.$$

Here using Jensen's inequality and Lemma B.2,

$$\|\star\|_1 \leq \left(\frac{K_\mu}{(1-L_{pop,\mu})\delta_{mix}}\left(\frac{\sqrt{2|\mathcal{S}|}}{\sqrt{N}} + \frac{\Delta_{\bar{\pi}}K_a}{2}\right) + 2K_\mu L_{pop,\mu}^\tau\right)\frac{1+L_h}{1-\gamma}.$$

Similarly, the term ($\Box$) can be bounded by $\|\Box\|_1 \leq \|\pi^1 - \bar{\pi}\|_1(2C_h + L_h)$.

We finally analyze the term ($\triangle$) using Proposition D.6. We note that

$$(\triangle) - F^{\pi}(Q) = (\mathbf{M}_\tau - \mathbf{M}^{\pi})(\mathbf{I} - \gamma \mathbf{P}_\infty)(Q - Q^*),$$

where $\mathbf{M}_\tau := \operatorname{diag}(\{\mathbb{P}[s^1_{t+\tau} = s|\mathcal{F}_t]\bar{\pi}(a|s)\}_{s,a}) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ and taking Q, Q^* as vectors in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Utilizing Proposition D.6 to bound $\sigma_{max}(\mathbf{M}_\tau - \mathbf{M}^\pi)$, we conclude

$$\begin{aligned}\|\triangle - F^{\pi}(Q)\|_2 \leq & \Bigg(2C_{mix}\rho^T_{mix} + \frac{4K_\mu T_{mix}}{\delta^2_{mix}}\frac{\sqrt{2|\mathcal{S}|}}{\sqrt{N}} \\ & + \frac{2T_{mix}K_a K_\mu}{\delta^2_{mix}(1 - L_{pop,\mu})}\Delta_{\bar{\pi}} + \frac{2T_{mix}K_a}{\delta_{mix}}\|\pi^i - \bar{\pi}\|_1\Bigg)\|Q - Q^*\|_2.\end{aligned}$$

Since we have $\|Q - Q^*\|_2 \leq \frac{2(1+L_h)}{1-\gamma}$ and $\|\cdot\|_2 \leq \|\cdot\|_1$, the theorem follows. □

D.5. Main Results

We restate and prove Theorem 4.3 and Theorem 4.5.

Theorem D.8 (Centralized learning)**.** *Assume that $\eta > 0$ an arbitrary learning rate which satisfies $L_{\Gamma_\eta} < 1$, Assumptions 1, 2, 3 and 4 hold and π^* is the unique MFG-NE. Let $\varepsilon > 0$ be arbitrary. If the learning rates $\{\beta_m\}$ are as defined in Lemma 4.2,*

$$\begin{aligned} M_{td} > \max\left\{\frac{\log(4(1 - L_{\Gamma_\eta})^{-1}L_{md,q}C^{TD}_{pop,1}\varepsilon^{-1})}{\log(L^{-1}_{pop,\mu})}, \underline{M}_{td}\right\}, \quad K > \frac{\log 8\varepsilon^{-1}}{\log L^{-1}_{\Gamma_\eta}}, \quad \textit{and} \\ M_{pg} > \max\left\{4C^{TD}_1 L_{md,q}(1 - L_{\Gamma_\eta})^{-1}\varepsilon^{-1}, 16(C^{TD}_2)^2 L^2_{md,q}(1 - L_{\Gamma_\eta})^{-2}\varepsilon^{-2}\right\},\end{aligned}$$

then the (random) output π_K of Algorithm 2 satisfies

$$\mathbb{E}\left[\|\pi_K - \pi^*\|_1\right] \leq \varepsilon + \frac{L_{md,q}C^{TD}_{pop,2}}{(1 - L_{\Gamma_\eta})\sqrt{N}}.$$

Proof. Denote $q_k := q_h(\cdot,\cdot|\pi_k, \Gamma^\infty_{pop}(\pi_k))$. Firstly, by Theorem 4.2, for any $k \in 0, \ldots, K-1$, it holds for all combinations of states $\{\bar{s}^i\}_i \in \mathcal{S}^N$ that with probability 1,

$$\begin{aligned}\mathbb{E}[\|\widehat{q}_k - q_k\|_\infty | s^i_{kM_{td}M_{pg}} = \bar{s}^i, \pi_k] \leq & \frac{C^{TD}_1}{\sqrt{(M_{pg} + t_0)(M_{pg} + t_0 + 1)}} + \frac{C^{TD}_2\sqrt{M_{pg}}}{\sqrt{(M_{pg} + t_0)(M_{pg} + t_0 + 1)}} \\ & + C^{TD}_{pop,1}L^{M_{td}}_{pop,\mu} + \frac{C^{TD}_{pop,2}}{\sqrt{N}},\end{aligned}$$

since the policies followed by each agent are the same. Hence by iterated expectations and placing the definitions of M_{pg}, M_{td}, we obtain with probability 1,

$$\mathbb{E}[\|\widehat{q}_k - q_k\|_\infty|\pi^k] \leq \frac{3(1 - L_{\Gamma_\eta})\varepsilon}{4L_{md,q}} + \frac{C^{TD}_{pop,2}}{\sqrt{N}}.$$

Moreover, with probability 1,

$$\begin{aligned}\mathbb{E}[\|\pi_{k+1} - \pi^*\|_1|\pi_k] =& \mathbb{E}[\|\Gamma^{md}_\eta(\widehat{q}_k, \pi_k) - \pi^*\|_1|\pi_k] \\ \leq & \mathbb{E}[\|\Gamma^{md}_\eta(q_k, \pi_k) - \pi^*\|_1|\pi_k] + \mathbb{E}[\|\Gamma^{md}_\eta(q_k, \pi_k) - \Gamma^{md}_\eta(\widehat{q}_k, \pi_k)\|_1|\pi_k] \\ \leq & \mathbb{E}[\|\Gamma_\eta(\pi_k) - \pi^*\|_1|\pi_k] + \mathbb{E}[\|\Gamma^{md}_\eta(q_k, \pi_k) - \Gamma^{md}_\eta(\widehat{q}_k, \pi_k)\|_1|\pi_k] \\ \leq & L_{\Gamma_\eta}\|\pi_k - \pi^*\|_1 + L_{md,q}\,\mathbb{E}[\|\widehat{q}_k - q_k\|_\infty|\pi_k],\end{aligned}$$

which implies by the law of iterated expectations:

$$\mathbb{E}[\|\widehat{\pi}_{k+1} - \pi^*\|_1] \leq L_{\Gamma_\eta}\,\mathbb{E}[\|\widehat{\pi}_k - \pi^*\|_1] + \frac{3(1 - L_{\Gamma_\eta})\varepsilon}{4} + \frac{L_{md,q} C^{TD}_{pop,2}}{\sqrt{N}}.$$

Inductively applying the inequality for $k = 0, \ldots, K-1$ implies the statement of the theorem, noting $\|\pi_0 - \pi^*\|_1 \leq 2$. □

Finally, we restate Theorem 4.5 with explicit constants and provide the proof.

Theorem D.9 (Independent learning)**.** *Assume that $\eta > 0$ satisfies $L_{\Gamma_\eta} < 1$, Assumptions 1, 2, 3 and 4 hold and π^* is the unique MFG-NE. Let $\varepsilon > 0$ be arbitrary. Let the learning rates $\{\beta_m\}$ for CTD be as defined in Lemma 4.2, and $K > \frac{\log 8\varepsilon^{-1}}{\log L^{-1}_{\Gamma_\eta}}$. For $c_\eta := L_{md,q}(C^{TD}_{pol,1} + C^{TD}_{pol,2}) + L_{md,\pi} = \mathcal{O}(\frac{1}{\rho})$:*

1. *If $c_\eta < 1$, let $M_{td} > \max\left\{\frac{\log[4(1-L_{\Gamma_\eta})^{-1}(1-c_\eta)^{-1}(L_{md,q}+L^2_{md,q}C^{TD}_{pop,1})\varepsilon^{-1}]}{\log(L^{-1}_{pop,\mu})}, \underline{M}_{td}\right\}$ and $M_{pg} > 4(1-c_\eta)^{-1}(1-L_{\Gamma_\eta})^{-1}(L_{md,q} + L^2_{md,q}C^{TD}_{pop,1})\max\left\{C^{TD}_1\varepsilon^{-1}, (C^{TD}_2)^2(1-L_{\Gamma_\eta})^{-1}\varepsilon^{-2}\right\}$.*
2. *If $c_\eta = 1$, let $M_{td} > \max\left\{\frac{\log[4(1-L_{\Gamma_\eta})^{-1}L_{md,q}C^{TD}_{pop,1}\varepsilon^{-1}(1+C^{TD}_{pol,1}L_{md,q}K)]}{\log(L^{-1}_{pop,\mu})}, \underline{M}_{td}\right\}$ and $M_{pg} > 4\max\left\{\frac{L_{md,q}C^{TD}_1}{1-L_{\Gamma_\eta}}(1+C^{TD}_{pol,1}L_{md,q}K)\varepsilon^{-1}, \frac{4(C^{TD}_2)^2L^2_{md,q}}{(1-L_{\Gamma_\eta})^2}(1+C^{TD}_{pol,1}L_{md,q}K)^2\varepsilon^{-2}\right\}$.*
3. *If $c_\eta > 1$, let $M_{td} > \max\left\{\frac{\log[4(1-L_{\Gamma_\eta})^{-1}L_{md,q}C^{TD}_{pop,1}(1+\frac{C^{TD}_{pol,1}L_{md,q}}{c_\eta-1}c^K_\eta)\varepsilon^{-1}]}{\log(L^{-1}_{pop,\mu})}, \underline{M}_{td}\right\}$ and $M_{pg} > 4(1-L_{\Gamma_\eta})^{-2}L_{md,q}\max\left\{C^{TD}_1\varepsilon^{-1}, 4(C^{TD}_2L_{md,q})^2\varepsilon^{-2}\right\}(1+\frac{C^{TD}_{pol,1}L_{md,q}}{c_\eta-1}c^K_\eta)^2$.*

Then, the (random) output $\{\pi^i_K\}_i$ of Algorithm 3 satisfies for all agents $i = 1, \ldots, N$, $\mathbb{E}\left[\|\pi^i_K - \pi^*\|_1\right] \leq \varepsilon + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$.

Proof. The proof again follows by using previous error propagation results for CTD learning with population and the contractivity of Γ_η. As the main difference from the centralized algorithm analysis, the constant c_η characterizes if stochasticity will cause the policies of agents to diverge over PMA epochs.

By symmetry, we only prove the theorem for the first agent ($i = 1$). We will use the reference policy $\bar{\pi}_k := \pi^1_k$ at each iteration k. For all $k = 0, 1, \ldots, K$ define the random variable $\Delta_k := \sum_{i=1}^N \|\pi^i_k - \pi^1_k\|_1$. Clearly $\Delta_0 = 0$. We will prove a bound for Δ_k throughout training. Using Theorem 4.2 on the CTD iterations of agent i, we obtain

$$\begin{aligned}\mathbb{E}[\|\widehat{q}^i_k - \widehat{q}^1_k\|_\infty | \{\pi^i_k\}_i] \leq & \frac{C^{TD}_1}{\sqrt{(M_{pg}+t_0)(M_{pg}+t_0+1)}} + \frac{C^{TD}_2\sqrt{M_{pg}}}{\sqrt{(M_{pg}+t_0)(M_{pg}+t_0+1)}} \\ & + C^{TD}_{pop,1}L^{M_{td}}_{pop,\mu} + \frac{C^{TD}_{pop,2}}{\sqrt{N}} + C^{TD}_{pol,1}\Delta_k + C^{TD}_{pol,2}\|\pi^i_k - \pi^1_k\|_1.\end{aligned}$$

For simplicity, we denote the first four summands independent of k as ϵ_{TD}, yielding

$$\mathbb{E}[\|\widehat{q}^i_k - \widehat{q}^1_k\|_\infty | \{\pi^i_k\}_i] \leq \epsilon_{TD} + C^{TD}_{pol,1}\Delta_k + C^{TD}_{pol,2}\|\pi^i_k - \pi^1_k\|_1.$$

Using the iterative expectations used to prove Theorem 4.3, we have for all $i \neq 1$ with probability 1:

$$\begin{aligned}\mathbb{E}[\|\pi^i_{k+1} - \pi^1_{k+1}\|_1 | \{\pi^i_k\}_i] =& \mathbb{E}[\|\Gamma^{md}_\eta(\widehat{q}^i_k, \pi^i_k) - \Gamma^{md}_\eta(\widehat{q}^1_k, \pi^1_k)\|_1 | \{\pi^i_k\}_i] \\ \leq& L_{md,q}\,\mathbb{E}[\|\widehat{q}^i_k - \widehat{q}^1_k\|_\infty | \{\pi^i_k\}_i] + L_{md,\pi}\|\pi^i_k - \pi^1_k\|_1 \\ \leq& L_{md,q}\epsilon_{TD} + L_{md,q}C^{TD}_{pol,1}\Delta_k + (L_{md,q}C^{TD}_{pol,2} + L_{md,\pi})\|\pi^i_k - \pi^1_k\|_1.\end{aligned}$$

In this last inequality, dividing by N, summing over all $i = 1, \ldots, N$ and taking the expectation of both sides we obtain:

$$\mathbb{E}[\Delta_{k+1}] \leq L_{md,q}(C^{TD}_{pol,1} + C^{TD}_{pol,2})\,\mathbb{E}[\Delta_k] + L_{md,\pi}\,\mathbb{E}[\Delta_k] + L_{md,q}\epsilon_{TD}. \tag{12}$$

Hence we obtain inductively the bound (for all k) and using the definition of c_η and $\Delta_0 = 0$,

$$\mathbb{E}[\Delta_k] \leq \frac{(L_{md,q}(C^{TD}_{pol,1} + C^{TD}_{pol,2}) + L_{md,\pi})^{k+1} - 1}{L_{md,q}(C^{TD}_{pol,1} + C^{TD}_{pol,2}) + L_{md,\pi} - 1} L_{md,q}\epsilon_{TD} = \frac{c_\eta^{k+1} - 1}{c_\eta - 1} L_{md,q}\epsilon_{TD},$$

if $c_\eta \neq 1$, otherwise $\mathbb{E}[\Delta_k] \leq k L_{md,q}\epsilon_{TD}$. Applying Theorem 4.2 once more on agent 1:

$$\mathbb{E}[\|\widehat{\pi}^1_{k+1} - \pi^*\|_1] \leq L_{\Gamma_\eta} \mathbb{E}[\|\widehat{\pi}_k - \pi^*\|_1] + L_{md,q}\epsilon_{TD} + L_{md,q} C^{TD}_{pol,1} \mathbb{E}[\Delta_k],$$

and summing over k iteratively and denoting $c_\eta := L_{md,q}(C^{TD}_{pol,1} + C^{TD}_{pol,2}) + L_{md,\pi}$:

$$\mathbb{E}[\|\widehat{\pi}^1_K - \pi^*\|_1] \leq 2L^K_{\Gamma_\eta} + \frac{L_{md,q}\epsilon_{TD}}{1 - L_{\Gamma_\eta}} + L_{md,q} C^{TD}_{pol,1} \sum_{k=1}^{K-1} L^{K-k-1}_{\Gamma_\eta} \mathbb{E}[\Delta_k].$$

The result follows by placing the defined constants in the theorem. □