Learning in Herding Mean Field Games: Single-Loop Algorithm with Finite-Time Convergence Analysis

Sihan Zeng **Sujay Bhatt** **Alec Koppel** **Sumitra Ganesh**
JPMorgan AI Research, United States

Abstract

We consider discrete-time stationary mean field games (MFG) with unknown dynamics and design algorithms for finding the equilibrium with finite-time complexity guarantees. Prior solutions to the problem assume either the contraction of a mean field optimality-consistency operator or strict weak monotonicity, which may be overly restrictive. In this work, we introduce a new class of solvable MFGs, named the "fully herding class", which expands the known solvable class of MFGs and for the first time includes problems with multiple equilibria. We propose a direct policy optimization method, Accelerated Single-loop Actor Critic Algorithm for Mean Field Games (ASAC-MFG), that provably finds a global equilibrium for MFGs within this class, under suitable access to a single trajectory of Markovian samples. Different from the prior methods, ASAC-MFG is single-loop and single-sample-path. We establish the finite-time and finite-sample convergence of ASAC-MFG to a mean field equilibrium via new techniques that we develop for multi-time-scale stochastic approximation. We support the theoretical results with illustrative numerical simulations.

When the mean field does not affect the transition and reward, an MFG reduces to a Markov decision process (MDP) and `ASAC-MFG` becomes an actor-critic algorithm for finding the optimal policy in average-reward MDPs, with a sample complexity matching the state-of-the-art. Previous works derive the complexity assuming a contraction on the Bellman operator, which is invalid for average-reward MDPs. We match the rate while removing the untenable assumption through an improved Lyapunov function.

1 INTRODUCTION

The mean field game (MFG) framework, introduced in Huang et al. [2006], Lasry and Lions [2007], provides an infinite-population approximation to the N-agent Markov game with a large number of homogeneous agents. It addresses the increasing difficulty in solving Markov games as N scales up and finds practical applications in many domains, including resource allocation [Li et al., 2020, Mao et al., 2022], wireless communication [Narasimha et al., 2019, Jiang et al., 2019], and the management of power grids [Alasseur et al., 2020, Zhang et al., 2021b].

A mean field equilibrium (MFE) describes the notion of solution in an MFG, and is a pair consisting of a policy and a mean field: the policy performs optimally in a Markov decision process (MDP) determined by the mean field, whereas the mean field is the induced stationary distribution of the states when every agent in the infinite population adopts the policy. In the discrete-time setting without explicit knowledge of the environment dynamics, reinforcement learning (RL) provides an important tool for finding an MFE using samples of state transitions and rewards. Currently two classes of MFGs are known to be solvable by RL with finite-time convergence guarantees, which we now review.

Existing classes of solvable MFGs

• *Contractive MFG*: In the context of MFGs, the optimality operator maps a mean field distribution to the optimal policy in the MDP determined by the mean field, whereas the consistency operator returns the induced mean field for a given policy. A series of recent works [Guo et al., 2019, Xie et al., 2021, Anahtarci et al., 2023, Mao et al., 2022, Zaman et al., 2023, Yardim et al., 2023] make the assumption that the composition of the optimality and consistency operator is a contractive mapping, or enforce the contraction through an entropy regularization. Fixed-point iteration methods (such as the extension of Q-learning to MFGs) are developed and analyzed by exploiting the contraction. A related concurrent work [Zhang et al., 2024a] relies on a "sufficiently Lipschitz MDP" assumption which plays a similar role of ensuring that the MFE is the unique fixed point of a contractive mapping. However, as pointed

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

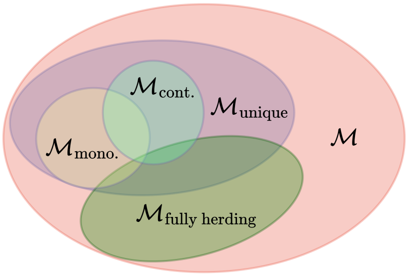
$\mathcal{M}_{\text{cont.}}$
$\mathcal{M}_{\text{unique}}$
$\mathcal{M}$
$\mathcal{M}_{\text{mono.}}$
$\mathcal{M}_{\text{fully herding}}$

Figure 1: $\mathcal{M}$ denotes the class of all MFGs. $\mathcal{M}_{\text{cont.}}$ and $\mathcal{M}_{\text{mono.}}$ are the MFG classes satisfying contraction and strict weak monotonicity, and they are subsets of $\mathcal{M}_{\text{unique}}$ which is the class of MFGs having a unique equilibrium. The proposed algorithm, `ASAC-MFG`, solves MFGs in $\mathcal{M}_{\text{fully herding}}$ [cf. Def. 2].

out in Yardim et al. [2024], the contraction assumption usually only holds when an impractically large regularization is added. Since the policy at the regularized equilibrium quickly approaches a uniform distribution as the regularization weight increases, solving such a regularized problem is usually *uninformative* about the original game. It is also worth noting that contractive MFGs can only have a unique equilibrium, while multiple equilibria usually exist in MFGs used to model practical problems [Nutz et al., 2020, Dianetti et al., 2024].

- *Strictly Weak Monotone MFG*: Another line of work relies on the assumption that the MFG is strictly weak monotone [Perrin et al., 2020, Perolat et al., 2021, Geist et al., 2021, Zhang et al., 2024b]. This condition intuitively means that the agent is "*discouraged from taking similar state-action pairs as the rest of the population*" and can be interpreted as an "aversion to crowded areas" [Perolat et al., 2021]. Strictly weakly monotone MFGs are again limited in that they need to have a unique equilibrium.

Figure. 1 gives an overview of the class of solvable MFGs. The aim of this paper is to expand the class of solvable MFGs and to design an easily implementable, provably convergent, and efficient algorithm for solving MFGs. We summarize our main contributions below.

1.1 Main Contributions

1. We introduce the κ-herding class of MFGs and the fully herding class which is a special case with $\kappa = 0$, and show that the fully herding MFGs are perfectly solvable. It is known from Yardim et al. [2024] that solving general MFGs (even with Lipschitz transition kernel and reward function) is a PPAD-complete problem, conjectured to be computationally intractable [Daskalakis et al., 2009]. Notably, we show that the fully herding class contains MFG instances admitting more than one equilibrium. Such MFGs do not satisfy either contraction or strict weak monotonicity and are not previously known to be solvable. In this sense, our work complements and expands on the finding of Yardim et al. [2024] and enlarges the class of solvable MFGs. As pointed out in Guo et al. [2024], Cui and Koeppl [2021], MFGs with multiple equilibria are very common but significantly more challenging to solve than those with a unique equilibrium.

2. We propose a single-loop, single-sample-path policy optimization algorithm `ASAC-MFG` for finding the equilibrium for MFGs in the herding class, and explicitly characterize its finite-time and finite-sample complexity. For MFGs in the fully herding class ($\kappa = 0$), `ASAC-MFG` converges to a global MFE with a rate of $\widetilde{\mathcal{O}}(k^{-1/4})$; for general herding MFGs with $\kappa > 0$, it converges to a $\sqrt{\kappa}-$approximate equilibrium at the same rate. As our algorithm draws exactly one sample in each iteration, the finite-time complexity translates to a finite-sample complexity of the same order. We note that any Lipschitz MFG can be shown to belong to the LL_V-herding class due to a simple Lipschitz continuity bound, where L and L_V are the Lipschitz constants introduced later. As a result, for MFGs not in the fully herding class, `ASAC-MFG` is still stable and can provably find a (despite non-ideal) $\mathcal{O}(\sqrt{LL_V})$ approximate solution. In comparison, the existing fixed-point iteration algorithms based on the contractive MFG assumption may in theory exhibit arbitrarily unstable behavior when the assumption fails.

3. The single-loop and single-sample-path structure makes our algorithm easily implementable. While single-loop and single-sample-path algorithms are widely used to solve RL and games in practice due to simplicity, their theoretical understanding is not as complete as their nested-loop counterparts. Specifically for MFGs, there does not currently exist a finite-time convergent single-loop and single-sample-path algorithm (see Table 1). Our work fills in the important gap. Note that our ability to make the algorithm single-loop and single-sample-path is not due to the herding condition. The analysis of `ASAC-MFG` is a technical innovation made by adapting the advances in accelerated two-time-scale stochastic approximation [Zeng and Doan, 2024] and generalizing them to a three-time-scale setting. We discuss the innovation in detail in Section 3.

4. We can regard a Markov decision process (MDP) as a degenerate MFG in which the transition kernel and reward are independent of the mean field. Recognizing this connection, we note that a simplified version of the proposed method becomes an actor-critic algorithm in an average-reward MDP and is guaranteed to converge to a stationary point of the policy optimization objective with rate $\widetilde{\mathcal{O}}(1/\sqrt{k})$. This matches the state-of-the-art complexity of the actor-critic algorithm [Chen and Zhao, 2024]. Existing works derive the complexity assuming a contraction

on the Bellman operator, which is invalid for average-reward MDPs. We maintain the rate but remove this assumption through introducing an alternative Lyapunov function.

1.2 Related Work

The classic works on MFGs study the continuous-time setting where the equilibrium point simultaneously satisfies a Hamilton–Jacobi–Bellman equation on the optimality of the policy and a Fokker–Planck equation that describes the dynamics of the mean field, and have proposed optimal control techniques that provably find the solution [Huang et al., 2006, 2007, Lasry and Lions, 2007]. In discrete time, MFGs can be considered a generalization of MDPs and are widely solved using RL. Among the latest representative works, Yang et al. [2018], Carmona et al. [2021], Perolat et al. [2021] build upon policy optimization and Anahtarcı et al. [2020], Angiuli et al. [2022, 2023] consider valued-based methods. The algorithms proposed in these works, however, either do not come with convergence analysis or are only shown to converge asymptotically.

Finite-time convergent algorithms are recently developed under the contraction assumption or strict weak monotonicity, as discussed earlier in the section. Table 1 highlights our contribution in terms of assumptions, algorithm structure, and sample complexity. Most papers listed introduce a large regularization to the MFG and establish the convergence to a regularized equilibrium. Notably, if Xie et al. [2021], Mao et al. [2022], Zaman et al. [2023], Yardim et al. [2023] could choose the regularization weight freely (note that they actually could not since the contraction condition only holds when the weight is sufficiently large as previously discussed), they can solve the original unregularized game by making the weight small enough. Solving the original game requires doubled complexities, i.e., become $\widetilde{\mathcal{O}}(\epsilon^{-10})$, $\widetilde{\mathcal{O}}(\epsilon^{-8})$, or $\widetilde{\mathcal{O}}(\epsilon^{-4})$ to the original solution.

Among works based on the strict weak monotonicity, Zhang et al. [2024b] proposes a mirror descent algorithm with a complexity matching that of `ASAC-MFG`, but is less convenient to implement due to its nested-loop structure and the requirement to pre-generate and store offline samples. Perrin et al. [2020] proposes a continuous-time algorithm and Geist et al. [2021] studies a deterministic gradient algorithm not based on samples, making their complexities not directly comparable.

Finally, we note the separate line of works [Guo et al., 2024, Mandal et al., 2023] that reformulate the MFG policy optimization problem as a constrained program with convex constraints and a bounded objective. The simple projected gradient descent algorithm provably solves the constrained program, leading to a solution of the MFG. However, a finite-time convergence guarantee is not established, unless again a sufficiently large regularization is added.

2 FORMULATION

We study MFGs in the stationary and infinite-horizon average-reward setting. An MFG is characterized by $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the *finite* state and action spaces. From the perspective of a single representative agent, the state transition depends not only on its own action but also on the aggregate behavior of all other agents. The aggregate behavior is described by the mean field $\mu \in \Delta_{\mathcal{S}}$[1], which measures the percentage of population in each state. The transition kernel of an MFG is represented by $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}} \to \Delta_{\mathcal{S}}$, where $\mathcal{P}^{\mu}(s' \mid s, a)$ denotes the probability that the state of the representative agent transitions from s to s' when it takes action a and mean field is μ. The mean field also affects the reward function $r : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}} \to [0, 1]$ – the agent receives reward $r(s, a, \mu)$ when it takes action a in state s under mean field μ. The agent *does not* observe the mean field, and takes actions according to policy $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$, which can be represented as a table $\Delta_{\mathcal{A}}^{\mathcal{S}} \subset \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$.

Given a policy π and mean field μ, the sequentially generated states form a Markov chain with transition matrix $P^{\pi,\mu} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$, with $P^{\pi,\mu}_{s',s} = \sum_{a \in \mathcal{A}} \mathcal{P}^{\mu}(s' \mid s, a)\pi(a \mid s)$. We denote by $\nu^{\pi,\mu} \in \Delta_{\mathcal{S}}$ the stationary distribution of the Markov chain, which is the right singular vector of $P^{\pi,\mu}$ associated with singular value 1, i.e., $\nu^{\pi,\mu} = P^{\pi,\mu}\nu^{\pi,\mu}$. When the mean field is μ and the agent generates actions according to π, the agent can expect to collect the cumulative reward $J(\pi, \mu)$

$$\begin{aligned} J(\pi, \mu) &\triangleq \lim_{T\to\infty} \tfrac{1}{T}\mathbb{E}_{\pi,\mathcal{P}^{\mu}}[\textstyle\sum_{t=0}^{T-1} r(s_t, a_t, \mu) \mid s_0] \\ &= \mathbb{E}_{s\sim\nu^{\pi,\mu},\, a\sim\pi(\cdot|s)}[r(s, a, \mu)]. \end{aligned} \tag{1}$$

If the mean field were fixed to a given μ, the goal of the agent would be to find a policy π that maximizes $J(\pi, \mu)$. However, when every agent in the infinite population follows the same policy as the representative agent, the mean field evolves as a function of π. We use $\mu^{\star} : \Delta_{\mathcal{A}}^{\mathcal{S}} \to \Delta_{\mathcal{S}}$ to denote the mapping from a policy to the induced mean field, which is the stationary distribution of states when the infinite number of players in the game all adopt policy π, i.e, $\mu^{\star}(\pi) = \nu^{\pi,\mu^{\star}(\pi)}$. The goal of the representative agent in an MFG is to find a policy optimal under the mean field induced by the policy. Mathematically, we want to find a pair of policy and mean field $(\bar{\pi}, \bar{\mu})$, known to always exist [Cui and Koeppl, 2021], as the solution to the system

$$\begin{cases} J(\bar{\pi}, \bar{\mu}) \geq J(\pi, \bar{\mu}), \quad \forall \pi & (2) \\ \bar{\mu} = \mu^{\star}(\bar{\pi}). & (3) \end{cases}$$

We assume that the induced mean field $\mu^{\star}(\pi)$ is unique for any π. This does not imply that the MFE $(\bar{\pi}, \bar{\mu})$ is unique.

[1] We use $\Delta_{\mathcal{S}}$ and $\Delta_{\mathcal{A}}$ to denote the probability simplex over the state and action spaces.

	Assumption	*Single Sample Path*	*Single Loop*	*Sample Complexity*
Guo et al. [2019]	*Contraction*	*No*	*No*	*Regularization Dependent*
Xie et al. [2021]	*Contraction*	*Yes**	*Yes**	$\widetilde{\mathcal{O}}(\epsilon^{-5})$, *regularized solution*
Mao et al. [2022]	*Contraction*	*No*	*No*	$\widetilde{\mathcal{O}}(\epsilon^{-5})$, *regularized solution*
Zaman et al. [2023]	*Contraction*	*Yes*	*No*	$\widetilde{\mathcal{O}}(\epsilon^{-4})$, *regularized solution*
Yardim et al. [2023]	*Contraction*	*Yes*	*No*	$\widetilde{\mathcal{O}}(\epsilon^{-2})$, *regularized solution*
Zhang et al. [2024b]	*Strict Weak Monotonicity*	*No*	*No*	$\widetilde{\mathcal{O}}(\epsilon^{-4})$, *original solution*
This Work	***Fully Herding Class*** ($\kappa = 0$)	***Yes***	***Yes***	$\widetilde{\mathcal{O}}(\epsilon^{-4})$, ***original solution***
This Work	***Partially Herding Class***($\kappa > 0$)	***Yes***	***Yes***	$\widetilde{\mathcal{O}}(\epsilon^{-4})$, $\sqrt{\kappa}$***-optimal solution***

Table 1: Assumption, structure, and complexity of existing algorithms with finite-sample analysis. * The algorithm in Xie et al. [2021] is single-loop and single-sample-path under an oracle that returns the stationary distribution of states for any π, μ. Mao et al. [2022] also relies on such an oracle. Our work, in comparison, is oracle-free.

Definition 1 (**ϵ−MFE**) *The pair of policy and mean field (π, μ) is an ϵ-Mean Field Equilibrium (MFE) if*

$$J(\pi', \mu) - J(\pi, \mu) \leq \epsilon, \forall \pi', \text{ and } \|\mu - \mu^\star(\pi)\| \leq \epsilon. \quad (4)$$

If the pair (π, μ) satisfies (4) with $\epsilon = 0$, it is obviously an exact MFE as a solution to (2)-(3). The definition characterizes what it mathematically means when we say (π, μ) is an approximate solution to the MFG. We also make use of the differential value function $V^{\pi,\mu} \in \mathbb{R}^{|\mathcal{S}|}$ to quantify the relative value of each initial state

$$V^{\pi,\mu}(s) \triangleq \mathbb{E}_{\pi,\mathcal{P}^\mu}\left[\textstyle\sum_{t=0}^{\infty}(r(s_t, a_t, \mu) - J(\pi, \mu)) \mid s_0 = s\right].$$

2.1 Herding MFGs

In this section, we identify a novel class of solvable MFGs named the "fully herding MFGs" and show that it contains instances that do not satisfy the contraction or strict weak monotonicity conditions.

Definition 2 (Herding MFG) *Given $\kappa \geq 0$, an MFG M is in the κ-herding class, denoted as $M \in \mathcal{M}_\kappa$, if there exists a constant $0 < \rho < \infty$ such that M satisfies the following "herding condition" for all π, π'*

$$\begin{aligned} &J(\pi, \mu^\star(\pi)) - J(\pi', \mu^\star(\pi')) \\ &\leq \rho\Big(J(\pi, \mu^\star(\pi)) - J(\pi', \mu^\star(\pi))\Big) + \kappa\|\pi - \pi'\|. \end{aligned} \quad (5)$$

The MFG M is further said to be a **fully herding MFG** if $M \in \mathcal{M}_{\text{fully herding}} \triangleq \mathcal{M}_0$ and a partially herding MFG if $M \in \mathcal{M}_\kappa$ for some $0 < \kappa < \infty$.

Conceptually, in MFGs with a small or zero κ, the representative agent receives a higher reward by "following the crowd" or displaying a "herding" behavior. The constant κ quantifies the difficulty of an MFG. We will show that the proposed algorithm `ASAC-MFG` provably finds a global equilibrium for fully herding MFGs ($\kappa = 0$), whereas for MFGs in the partial herding class we can at least provably find an $\mathcal{O}(\sqrt{\kappa})$-MFE (in the sense of Definition 1). Notably, all Lipschitz MFGs satisfy the herding condition with $\rho = 1$ and $\kappa = LL_V$ in the worst case, where L, L_V are Lipschitz constants introduced later. This means that the proposed method exhibits stable behavior and converges to an approximate (though non-ideal) solution in any MFG, whereas the existing algorithms lose stability guarantees when contraction/monotonicity fails to hold. Next, we introduce a few examples to clarify the herding condition.

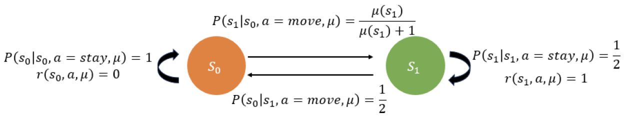
$P(s_1|s_0, a = move, \mu) = \frac{\mu(s_1)}{\mu(s_1) + 1}$
$P(s_0|s_0, a = stay, \mu) = 1$
$r(s_0, a, \mu) = 0$
S_0
S_1
$P(s_1|s_1, a = stay, \mu) = \frac{1}{2}$
$r(s_1, a, \mu) = 1$
$P(s_0|s_1, a = move, \mu) = \frac{1}{2}$

Figure 2: Illustration of Example 2

Example 1 *Consider an MFG which satisfies $\forall \mu, \mu', s, a, s'$*

$$r(s, a, \mu) = q\big(\mu(s)\big)^p \text{ and } \mathcal{P}(s' \mid s, a, \mu) = \mathcal{P}(s' \mid s, a, \mu'),$$

for some scalars $p, q > 0$, i.e., the transition probability kernel is independent of the mean field. The MFG belongs to the fully herding class ($\kappa = 0, \rho = p+1$) and is perfectly solvable by `ASAC-MFG`.

Consider an MFG instance satisfying the conditions in Example 1 with $|\mathcal{S}| = |\mathcal{A}| = 2$, $q = 1$, $p = 1$, where the transition kernel is such that in either state $s \in \{s_1, s_2\}$, the action a_1 (resp. a_2) leads the next state to s_1 (resp. s_2) with probability $p = 3/4$. There exist an infinite number of equilibria in this MFG. They occur at policies $\bar{\pi}_1$, $\bar{\pi}_2$ such that $\forall s$

$$\bar{\pi}_1(a \mid s) = \begin{cases} 1, & \text{if } a = a_1 \\ 0, & \text{if } a = a_0 \end{cases} \quad \bar{\pi}_2(a \mid s) = \begin{cases} 0, & \text{if } a = a_1 \\ 1, & \text{if } a = a_0 \end{cases}$$

with the induced mean field $\bar{\mu}_1 = [3/4, 1/4]^\top$, $\bar{\mu}_2 = [1/4, 3/4]^\top$, and at all policies that induce the mean field $[1/2, 1/2]^\top$ (such as $\bar{\pi}_3(a \mid s) = 1/2$ for all s, a). The MFE is not unique, implying that MFGs in Example 1 generally do not satisfy the contraction or strict weak monotonicity assumptions. The derivation of the mean field equilibria and the proof that MFGs in Example 1 are in the fully herding class can be found in Appendix F. We show another example of fully herding MFG in which the transition kernel is dependent on the mean field.

Example 2 *Consider an MFG with two states, s_0 and s_1, and two actions "move" and "stay". From s_0, we can move to s_1 with probability $\frac{\mu_{s_1}}{1+\mu(s_1)}$ or stay in s_0 with probability $\frac{1}{1+\mu(s_1)}$ by taking action "move". Taking the action "stay" in state s_0 makes us stay in the state with probability 1. When in state s_1, we transition to s_0 or s_1 for the next state each with probability $1/2$ under any action. In state s_1, we collect a reward of 1 regardless of action and mean field. The reward in any other situation is 0. This MFG has an infinite number of equilibria – the optimal action to take in state s_0 is always "move", but any policy can be taken in state s_1. This MFG satisfies* (5) *with $\kappa = 0, \rho = 2$ but not the contraction or strict weak monotonicity condition.*

Note that Examples 1 and 2 are instances within the set $\mathcal{M}_{\text{fully herding}} \backslash \mathcal{M}_{\text{unique}}$ in Figure 1 and not known to be provably solvable in the existing literature.

The herding class connects to practical problems in real life. An important example is crowd motion with a reward function that models "attraction to the mean". Specifically, the reward of the representative player is high for the states that match the mean field, which may take the form of $r(s, a, \mu) = \mu(s)$ and fall under Example 1. (See Section 3.1.1 of Dayanıklı and Laurière [2024]. Their reward for the "attraction to the mean" setting can be regarded as a variant of $r(s, a, \mu) = \mu(s)$ where the states are discretized continuous numbers.) Global carbon emission can be understood as a specific example of crowd motion with an "attraction to the mean" reward, in which the representative agent is a country that needs to determine its emission level. The agent has a tendency of following the emission level of the population (in this case, all other countries), since if it produces emissions lower than the average, "it behaves as an opportunity cost of not producing more" [Dayanikli and Lauriere, 2024]; on the other hand, if it produces emission higher than the average, "it behaves as a reputation cost for polluting more" [Dayanikli and Lauriere, 2024].

3 ALGORITHM

Our algorithm solves MFGs from the perspective of direct policy optimization. As we do not directly deal with the mean field optimality-consistency operator, we bypass the need to assume that it is contractive. We see from (2) that if the optimal policy under $\bar{\mu}$ were unique and we knew $\bar{\mu}$, we could easily find $\bar{\pi}$ through policy optimization with the mean field fixed to $\bar{\mu}$. On the other hand, if we knew the equilibrium policy $\bar{\pi}$, we could obtain $\bar{\mu}$ by finding $\mu^\star(\bar{\pi})$. However, we do not know either $\bar{\pi}$ or $\bar{\mu}$ in reality and therefore consider the approach of simultaneous learning. Specifically, we maintain a parameter θ that encodes the policy π_θ, and an iterate $\hat{\mu}$ to estimate the mean field induced by the current policy, and improve θ and $\hat{\mu}$ with respect to each other by iteratively taking the steps

$$\theta_{k+1} = \theta_k + \alpha_k \nabla_\theta J(\pi_{\theta_k}, \hat{\mu}_k), \quad \hat{\mu}_{k+1} = \mu^\star(\pi_{\theta_k}) \quad (9)$$

where k indexes the iteration and α_k is a step size. By the policy gradient theorem [Sutton et al., 1999], a closed-form expression for $\nabla_\theta J(\pi_\theta, \mu)$ is

$$\nabla_\theta J(\pi_\theta, \mu) = \mathbb{E}_{s \sim \nu^{\pi_\theta, \mu}, a \sim \pi_\theta(\cdot|s), s' \sim \mathcal{P}^\mu(\cdot|s,a)} \Big[(r(s,a,\mu) + V^{\pi_\theta,\mu}(s') - V^{\pi_\theta,\mu}(s)) \nabla_\theta \log \pi_\theta(a \mid s) \Big].$$

Without loss of generality, we use a softmax parameterization, i.e., the parameter $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ represents the policy as

$$\pi_\theta(a \mid s) = \frac{\exp(\theta(s,a))}{\sum_{a' \in \mathcal{A}} \exp(\theta(s,a'))}.$$

In large and/or unknown environments in the real life, performing (9) poses challenges. The updates require the

Algorithm 1 Accelerated Single-loop Actor Critic Algorithm for Mean Field Games (`ASAC-MFG`)

```
Initialize: policy parameter $\theta_0$, value function estimate $\hat{V}_0, \hat{J}_0$, mean field estimate $\hat{\mu}_0 \in \Delta_{\mathcal{S}}$, gradient/operator estimates $f_0 = 0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, g_0^V = 0 \in \mathbb{R}^{|\mathcal{S}|}, g_0^J = 0 \in \mathbb{R}, h_0 = 0 \in \mathbb{R}^{|\mathcal{S}|}$, initial state $s_0$
for iteration $k = 0, 1, 2, ...$ do
    Take action $a_k \sim \pi_{\theta_k}(\cdot \mid s_k)$. Observe $r(s_k, a_k, \hat{\mu}_k)$ and $s_{k+1} \sim \mathcal{P}^{\hat{\mu}_k}(\cdot \mid s_k, a_k)$
    Policy (actor) update:
        $\theta_{k+1} = \theta_k + \alpha_k f_k.$    (6)
    Mean field update:
        $\hat{\mu}_{k+1} = \Pi_{\Delta_{\mathcal{S}}}(\hat{\mu}_k + \xi_k h_k).$    (7)
    Value function (critic) update:
        $\hat{V}_{k+1} = \Pi_{B_V}(\hat{V}_k + \beta_k g_k^V),$
        $\hat{J}_{k+1} = \Pi_{[0,1]}(\hat{J}_k + \beta_k g_k^J).$    (8)
    Gradient/Operator estimate update:
        $f_{k+1} = (1-\lambda_k) f_k + \lambda_k \nabla \log \pi_{\theta_k}(a_k \mid s_k)$
            $\times (r(s_k, a_k, \hat{\mu}_k) + \hat{V}_k(s_{k+1}) - \hat{V}_k(s_k)),$
        $g_{k+1}^V = (1-\lambda_k) g_k^V + \lambda_k e_{s_k}$
            $\times (r(s_k, a_k, \hat{\mu}_k) - \hat{J}_k + \hat{V}_k(s_{k+1}) - \hat{V}_k(s_k)),$
        $g_{k+1}^J = (1-\lambda_k) g_k^J + \lambda_k c_J (r(s_k, a_k, \hat{\mu}_k) - \hat{J}_k),$
        $h_{k+1} = (1-\lambda_k) h_k + \lambda_k (e_{s_k} - \hat{\mu}_k).$
end for
```

knowledge of $\mu^\star(\pi_{\theta_k})$ and value function $V^{\pi_{\theta_k},\mu^\star(\pi_{\theta_k})}$. Neither of these quantities can be exactly calculated without the exact knowledge of the transition model, which is usually unavailable. We propose learning $\mu^\star(\pi_{\theta_k})$ and $V^{\pi_{\theta_k},\mu^\star(\pi_{\theta_k})}$ simultaneously, with the policy and mean field iterates updated using a continuous path of samples from the MFG. We recognize that for any θ

$$\mu^\star(\pi_\theta) = \lim_{T\to\infty} \tfrac{1}{T}\mathbb{E}_{\pi_\theta,\mathcal{P}^{\mu^\star(\pi_\theta)}}[\textstyle\sum_{t=0}^{T-1} e_{s_t} \mid s_0], \quad (10)$$

where $e_s \in \mathbb{R}^{|\mathcal{S}|}$ is the indicator vector whose entry s' is 1 if $s' = s$ and 0 otherwise. To solve Eq. (10) with multi-time-scale stochastic approximation, we carry out

$$\hat{\mu}_{k+1} = \hat{\mu}_k + \xi_k(e_{s_k} - \hat{\mu}_k) \quad (11)$$

iteratively for some step size $\xi_k \gg \alpha_k$. Due to the difference in time scales (step size), $\hat{\mu}_k$ becomes an increasingly accurate estimate of $\mu^\star(\pi_{\theta_k})$ as the iterations proceed.

We know that $V^{\pi_{\theta_k},\hat{\mu}_k}$ satisfies the Bellman equation

$$\begin{aligned} V^{\pi_{\theta_k},\hat{\mu}_k}(s) = \textstyle\sum_a \pi_{\theta_k}(a \mid s) r(s,a,\hat{\mu}_k) - J(\pi_{\theta_k},\hat{\mu}_k)& \\ + \textstyle\sum_{s'} P^{\pi_{\theta_k},\hat{\mu}_k}_{s',s} V^{\pi_{\theta_k},\hat{\mu}_k}(s'),\ \forall s.& \quad (12)\end{aligned}$$

We introduce an auxiliary variable $\hat{V} \in \mathbb{R}^{|\mathcal{S}|}$ to estimate $V^{\pi_{\theta_k},\hat{\mu}_k}$, again by stochastic approximation. The following update solves (12) under proper choices of β_k

$$\begin{aligned}\hat{V}_{k+1}(s_k) &= \hat{V}_k(s_k) \quad (13)\\ &+ \beta_k\big(r(s_k,a_k,\hat{\mu}_k) - \hat{J}_k + \hat{V}_k(s_{k+1}) - \hat{V}_k(s_k)\big),\end{aligned}$$

where the unknown $J(\pi_\theta, \mu^\star(\pi_\theta))$ is replaced with an estimate that itself is iteratively refined

$$\hat{J}_{k+1} = \hat{J}_k + \beta_k(r(s_k,a_k,\hat{\mu}_k) - \hat{J}_k). \quad (14)$$

Combining Eqs. (11), (13), and (14) with the θ update in (9) results in a single-loop single-sample-path algorithm where in the slowest time scale we ascend the policy parameter θ_k along the gradient direction and the faster time scales are used to compute the quantities necessary for the gradient evaluation. While such an algorithm can be shown to converge to an MFE, the convergence does not take the best possible rate due to the coupling between iterates: θ_k, $\hat{\mu}_k$, $\hat{V}_k$, and $\hat{J}_k$ directly affect each other's update, causing noise in any variable to immediately propagate to the others. Zeng and Doan [2024] details the degradation in the algorithm complexity resulting from such coupling effect when *two* variables are simultaneously updated. They further introduce a way of recovering the optimal complexity by modifying the algorithm with a denoising step. We adopt this technique and extend it to handle the three-time-scale coupling (α_k, β_k, ξ_k) in our updates, which is more challenging to tackle than two timescales. The modification to the algorithm is simple – we first estimate smoothed and denoised versions of the gradients before using them to update the policy, mean field, and value function iterates. We present the full details in Algorithm 1, in which the smoothed gradient estimates are f_k, g_k^V, g_k^J, and h_k updated recursively according to line 8.

In (7), $\Pi_{\Delta_\mathcal{S}} : \mathbb{R}^{|\mathcal{S}|} \to \Delta_\mathcal{S}$ denotes the projection to the simplex over $\mathcal{S}$. In (8), $\Pi_{B_V} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ denotes the projection to the ℓ_2-norm ball with radius B_V, and $\Pi_{[0,1]} : \mathbb{R} \to \mathbb{R}$ is the projection of a scalar to the range $[0,1]$. The projection operators guarantee the stability of the critic iterates in (8) and are a frequently used tool in the analysis of actor-critic algorithms in the literature [Wu et al., 2020, Chen and Zhao, 2024, Panda and Bhatnagar, 2024].

4 ASSUMPTIONS & FINITE-TIME ANALYSIS

This section introduces the main technical assumptions made in this paper and presents the finite-time convergence of Algorithm 1 to an MFE.

Assumption 1 (Uniform Geometric Ergodicity) *For any π, μ, the Markov chain $\{s_k\}$ generated by $P^{\pi,\mu}$ according to $s_{k+1} \sim P^{\pi,\mu}(\cdot \mid s_k)$ is irreducible and aperiodic. In addition, there exist $C_0 \geq 1$ and $C_1 \in (0,1)$ such that*

$$\sup_s d_{TV}\big(\mathbb{P}(s_k = \cdot \mid s_0 = s), \nu^{\pi,\mu}(\cdot)\big) \leq C_0 C_1^k,\ \forall k \geq 0,$$

where d_{TV} denotes the total variation (TV) distance (see definition in Eq. (24) *in the appendix).*

Assumption 1 requires that the k_{th} sample of the Markov chain exponentially approaches the stationary distribution as k goes up. In other words, the Markov chain generated under $P^{\pi,\mu}$ is geometrically ergodic for any π, μ. This assumption is common in papers that study the complexity of sample-based single-loop RL algorithms [Wu et al., 2020, Zeng et al., 2022, Chen and Zhao, 2024].

Assumption 2 (Estimability of Induced Mean Field) *There exists a constant $\delta \in (0,1)$ such that*

$$\|\nu^{\pi,\mu_1} - \nu^{\pi,\mu_2}\| \leq \delta\|\mu_1 - \mu_2\|, \quad \forall \pi, \mu_1, \mu_2.$$

Assumption 2 is standard (for example, see Eq.(8) in Guo et al. [2019], Assumption 3 in Xie et al. [2021], Assumption 3 in Mao et al. [2022], on the contractive Γ_2 operator) and can be viewed as an estimability condition on the induced mean field, whose validity depends only on the transition kernel $\mathcal{P}$. The assumption says that for any π the stationary distribution $\nu^{\pi,\mu}$ is contractive in μ, and implies the uniqueness of $\mu^\star(\pi)$. It guarantees that to estimate the induced mean field of a policy π, we can start from any initial μ_0, iteratively update it according to $\mu_k = \nu^{\pi,\mu_{k-1}}$, and have $\mu_k \to \mu^\star(\pi)$ as the iterations proceed. Note that this assumption is not to be confused with the contraction condition on the mean field optimality-consistency operator, which requires the existence of a $\delta \in (0,1)$ such that

$$\|\nu^{\pi^\star(\mu_1),\mu_1} - \nu^{\pi^\star(\mu_2),\mu_2}\| \leq \delta\|\mu_1 - \mu_2\|. \tag{15}$$

(15) is a much stronger assumption with validity depending on both reward and transition kernel. While (15) is popular in prior works (Table 1), we do not need the assumption.

Assumption 3 (Lipschitz Continuity and Boundedness) *Given two distributions d_1, d_2 over $\mathcal{S}$, policies π_1, π_2, and mean fields μ_1, μ_2, we draw samples according to $s \sim d_1, s' \sim P^{\pi_1,\mu_1}(\cdot \mid s)$ and $\hat{s} \sim d_2, \hat{s}' \sim P^{\pi_2,\mu_2}(\cdot \mid \hat{s})$. We assume that there exists a constant $L > 0$ such that*

$$|r(s,a,\mu_1) - r(s,a,\mu_2)| \leq L|\mu_1 - \mu_2\|, \tag{16}$$

$$d_{TV}(\mathbb{P}(s' = \cdot), \mathbb{P}(\hat{s}' = \cdot)) \leq d_{TV}(d_1, d_2) + L(\|\pi_1 - \pi_2\| + \|\mu_1 - \mu_2\|), \tag{17}$$

$$d_{TV}(\nu^{\pi_1,\mu_1}, \nu^{\pi_2,\mu_2}) \leq L(\|\pi_1 - \pi_2\| + \|\mu_1 - \mu_2\|), \tag{18}$$

$$\|\mu^\star(\pi_1) - \mu^\star(\pi_2)\| \leq L\|\pi_1 - \pi_2\|. \tag{19}$$

In addition, there is a constant $B_V > 0$ such that $\|V^{\pi,\mu}\| \leq B_V$, for all π, μ.

Eq. (16) states that the reward function is Lipschitz in the mean field. Eq. (17) amounts to a regularity condition on the transition probability matrix $P^{\pi,\mu}$ as a function of π and μ and can be shown to hold if the transition kernel $\mathcal{P}^\mu$ is Lipschitz in μ (using an argument similar to Wu et al. [2020][Lemma B.2]). Eqs. (18) and (19) impose the Lipschitz continuity of the stationary distribution and induced mean field, which also can be shown to hold under Assumption 2 if the transition kernel is Lipschitz (see Zou et al. [2019][Lemma 3]). In this work, we directly assume Eqs. (17)-(19) for simplicity. All conditions in Assumption 3 are common in the literature of MFGs and RL [Wu et al., 2020, Yardim et al., 2023, Anahtarci et al., 2023].

Assumption 4 (Fisher Non-Degenerate Policy) *Let $\mathbb{F}(\theta)$ denote the Fisher information matrix under parameter θ:*

$$\mathbb{F}(\theta) = \mathbb{E}_{s\sim\mu^\star(\pi_\theta), a\sim\pi_\theta(\cdot|s)}[\nabla_\theta \log \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s)^\top].$$

There is a constant $\sigma > 0$ such that $\mathbb{F}(\theta) - \sigma I_{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ is positive definite $\forall\theta$.

This assumption on Fisher non-degenerate policy implies a "gradient domination" condition – for any μ, every stationary point of the cumulative return $J(\pi_\theta, \mu)$ is globally optimal. This is again a standard assumption in the existing literature on policy optimization [Zhang et al., 2020, Liu et al., 2020, Fatkhullin et al., 2023, Ganesh et al., 2024]. It is worth noting that we do not need Assumption 4 to establish the main theoretical result (Theorem 1) where the policy convergence is measured by its distance to a first-order stationary point. The assumption is only used to translate a stationary point to a globally optimal policy in Corollary 1 under the average-reward formulation.

Remark 1 *Our algorithm and analysis can be extended to the discounted-reward setting, in which case Assumption 4 can be removed, as the gradient domination condition can be shown to hold under sufficient exploration (see Lemma 8 of Mei et al. [2020]). The same sample complexity can be established for* `ASAC-MFG` *under the discounted-reward formulation. The only major difference is the necessity of sampling from the discounted visitation measure, which requires using two sample trajectory (one for visitation measure, one for mean field).*

4.1 Main Results

Each variable in Algorithm 1 has a target to chase. The target of θ_k is a policy parameter optimal under its induced mean field, whereas $\hat{\mu}_k$ and $\hat{V}_k, \hat{J}_k$ aim to converge to the mean field induced by π_{θ_k} and the value functions under $\pi_{\theta_k}, \hat{\mu}_k$. We quantify the gap between these variables and their targets by the convergence metrics below, and will shortly show that they all decay to zero.

$$\varepsilon_k^\pi \triangleq \|\nabla_\theta J(\pi_{\theta_k}, \mu)|_{\mu=\mu^\star(\pi_{\theta_k})}\|^2,\ \varepsilon_k^\mu \triangleq \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\|^2,$$
$$\varepsilon_k^V \triangleq \|\Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k},\hat{\mu}_k})\|^2,\ \varepsilon_k^J \triangleq (\hat{J}_k - J(\pi_{\theta_k}, \hat{\mu}_k))^2.$$

We would like $\hat{V}_k$ to converge to $V^{\pi_{\theta_k},\hat{\mu}_k}$ which solves the Bellman equation (12). However, the solution is not unique. If $V \in \mathbb{R}^{|\mathcal{S}|}$ solves (12), so does $V + c\mathbf{1}_{|\mathcal{S}|}$ for any scalar c. We denote by $\mathcal{E}$ the subspace spanned by $\mathbf{1}_{|\mathcal{S}|}$ in $\mathbb{R}^{|\mathcal{S}|}$ and by $\mathcal{E}_\perp$ its orthogonal complement, i.e., for any $V \in \mathcal{E}_\perp$ we have $V^\top \mathbf{1}_{|\mathcal{S}|} = 0$. To make the convergence of the value function well-defined, we consider the metric ε_k^V above where $\Pi_{\mathcal{E}_\perp}$ is the orthogonal projection to $\mathcal{E}_\perp$. It is easy to see $\Pi_{\mathcal{E}_\perp} = I_{|\mathcal{S}|\times|\mathcal{S}|} - \mathbf{1}_{|\mathcal{S}|}\mathbf{1}_{|\mathcal{S}|}^\top/|\mathcal{S}|$.

Theorem 1 *Consider the iterates generated by Algorithm 1 on a κ-herding MFG, with the step sizes satisfying*

$$\lambda_k = \frac{\lambda_0}{\sqrt{k+1}},\ \alpha_k = \frac{\alpha_0}{\sqrt{k+1}},\ \beta_k = \frac{\beta_0}{\sqrt{k+1}},\ \xi_k = \frac{\xi_0}{\sqrt{k+1}},$$

where the constants $\lambda_0, \alpha_0, \beta_0, \xi_0$ are specified later in Appendix B.2. Under Assumptions 1-2, we have for all $k \geq \tau_k$

$$\min_{t<k} \mathbb{E}[\varepsilon_t^\pi + \varepsilon_t^\mu + \varepsilon_t^V + \varepsilon_t^J] \leq \mathcal{O}\left(\frac{\log^3(k+1)}{\sqrt{k+1}} + \kappa\right),$$

where τ_k denotes the mixing time, which is an affine function of $\log(k+1)$ defined in Appendix A.1.

Theorem 1 states that all main variables of Algorithm 1 converge to their learning targets with a rate of $\widetilde{\mathcal{O}}(k^{-1/2})$ up to an error linear in κ, under a single trajectory of Markovian samples. Since Algorithm 1 draws one sample in each iteration, this translates to a finite-sample complexity of the same order. We defer the detailed proof of the theorem to Appendix B but point out that the convergence rate is derived through a careful multi-time-scale analysis.

The step sizes have the same dependency on k, but need to observe $\alpha_0 \leq \xi_0 \leq \beta_0 \leq \lambda_0$.

Our ultimate goal is to find an ϵ-MFE in the sense of Definition 1. This requires us to connect the convergence of ε_k^π (gradient norm convergence) to the optimality gap below

$$\max_\pi J(\pi, \mu^\star(\pi_{\theta_k})) - J(\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})). \qquad (20)$$

Under Assumption 4 a "gradient domination" condition holds, which upper bounds (20) by $\sqrt{\varepsilon_k^\pi}$. We take advantage of the condition to derive the following corollary.

Corollary 1 *Consider the iterates generated by Algorithm 1 on a κ-herding MFG under the step sizes in Theorem 1. Under Assumptions 1-4, we have for all $k \geq \tau_k$*

$$\begin{aligned}\min_{t<k} \mathbb{E}\left[\max_\pi J(\pi, \mu^\star(\pi_{\theta_t})) - J(\pi_{\theta_t}, \mu^\star(\pi_{\theta_t}))\right] \\ \leq \widetilde{\mathcal{O}}((k+1)^{-1/4}) + \mathcal{O}(\sqrt{\kappa}),\\ \min_{t<k} \mathbb{E}[\|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\|] \leq \widetilde{\mathcal{O}}((k+1)^{-1/4}) + \mathcal{O}(\sqrt{\kappa}).\end{aligned}$$

Corollary 1 guarantees that within at most $\widetilde{\mathcal{O}}(\epsilon^{-4})$ iterations Algorithm 1 finds an ϵ-MFE in the sense of Definition 1 for MFGs in the fully herding class and an $(\epsilon + \mathcal{O}(\sqrt{\kappa}))$-MFE for general herding MFGs.

5 ACTOR-CRITIC ALGORITHM FOR MARKOV DECISION PROCESSES

An average-reward MDP can be regarded as a degenerate average-reward MFG in which the mean field has no impact on the transition kernel or reward, i.e., $\mathcal{P}^\mu(s' \mid s, a) = \mathcal{P}(s' \mid s, a)$ and $r(s, a, \mu) = r(s, a)$. Observing this connection, we recognize that Algorithm 1 (with the mean field update (7) removed; details presented in Appendix G and Algorithm 2) reduces to an online single-loop actor-critic algorithm that optimizes the following objective

$$J_{\text{MDP}}(\pi) \triangleq \lim_{T\to\infty} \tfrac{1}{T}\mathbb{E}_{\pi,\mathcal{P}}[\textstyle\sum_{t=0}^{T-1} r(s_t, a_t) \mid s_0].$$

There exist a series of works on this subject [Wu et al., 2020, Olshevsky and Gharesifard, 2023, Chen and Zhao, 2024], with the best-known complexity $\widetilde{\mathcal{O}}(1/\sqrt{k})$ established in Chen and Zhao [2024]. However, these prior works base their analyses on the unrealistic assumption that there exists a constant $\gamma \in (0, 1)$ such that given a policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$, the following inequality holds for all $V \in \mathbb{R}^{|\mathcal{S}|}$

$$V^\top \mathbb{E}_{s\sim\nu^\pi, s'\sim P^\pi_{:,s}}[e_s(e_{s'} - e_s)^\top]V \leq -\gamma\|V\|^2. \qquad (21)$$

The assumption contradicts the common knowledge that the value function in average-reward MDP is non-unique as we discussed in Sec.4.1, and therefore never holds true. Fortunately, under Assumption 1, it can be shown that the inequality (21) holds for all $V \in \mathcal{E}_\perp$ (as opposed to $V \in \mathbb{R}^{|\mathcal{S}|}$). This result is stated in Lemma 5 and the proof has been established in [Tsitsiklis and Van Roy, 1999, Zhang et al., 2021a]. This fact allows us to remove the assumption (21) in our analysis by treating the convergence of the value function in the space of $\mathcal{E}_\perp$. Specifically, while the prior works consider the Lyapunov function

$$\mathcal{L}_k = \mathbb{E}[\|\nabla_\theta J_{\text{MDP}}(\pi_{\theta_k})\|^2 + \|\hat{V}_k - V_{\text{MDP}}^{\pi_{\theta_k}}\|^2],$$

with $V_{\text{MDP}}^\pi(s) \triangleq \mathbb{E}_{\pi_\theta,\mathcal{P}}[\sum_{t=0}^\infty (r(s_t, a_t) - J_{\text{MDP}}(\pi)) \mid s_0 = s]$, we instead use $\|\Pi_{\mathcal{E}_\perp}(\hat{V}_k - V_{\text{MDP}}^{\pi_{\theta_k}})\|^2$ to replace $\|\hat{V}_k - V_{\text{MDP}}^{\pi_{\theta_k}}\|^2$. Note that we do not modify the algorithm to perform the projection to $\mathcal{E}_\perp$ but enhance the analysis only.

Corollary 2 *Consider the policy π_{θ_k} generated by Algorithm 2 with properly selected step sizes. Under Assumptions 1 and 3 (mapped to the context of single-agent MDP), we have for all $k \geq \tau_k$*

$$\min_{t<k} \mathbb{E}\left[\|\nabla_\theta J_{MDP}(\pi_{\theta_k})\|^2\right] \leq \mathcal{O}\left(\log^3(k+1)/\sqrt{k+1}\right).$$

Corollary 2 guarantees the best-iterate convergence of Algorithm 2 to a stationary point of J_{MDP}, with a finite-time complexity of $\widetilde{\mathcal{O}}(k^{-1/2})$. This matches the state-of-the-art bound in Chen and Zhao [2024], without making the restrictive assumption (21). More details on the problem formulation and algorithm in the context of MDP can be found in Appendix G. The proof is presented in Appendix C.2.

6 NUMERICAL SIMULATIONS

We verify the performance of the proposed algorithm through simulations on 1) small-scale synthetic MFGs, 2) the beach bar problem [Perrin et al., 2020].

First, we consider three environments of dimension $|\mathcal{S}| = |\mathcal{A}| = 10$, all with randomly generated transition kernels. Environment 1 is taken from Example 1, with the transition kernel independent of μ and the reward $r(s, a, \mu) = \mu(s)$ in expectation, which we know belongs to fully herding class and thus optimally solvable by ASAC-MFG. Environment 2 is generated the same way as Environment 1 except that the reward has a flipped sign, i.e., $r(s, a, \mu) = -\mu(s)$ in expectation, and does not satisfy (5) with $\kappa = 0$. Environment 3 has the same reward as Environment 1 and a random mean-field-dependent transition kernel.[2] With high probability the environment also does not satisfy (5) with $\kappa = 0$.

Because the equilibria are unknown, we measure the policy convergence by $\|\nabla_\theta J(\pi_{\theta_k}, \hat{\mu}_k)\|$ and the mean field convergence by $\|\hat{\mu}_k - \nu^{\pi_k, \hat{\mu}_k}\|$ as a proxy for $\|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\|$.

We compare ASAC-MFG with the algorithm proposed in Zaman et al. [2020] as the information oracles are similar, enabling a fair comparison. We consider two variations

[2] More details of the experimental are in Appendix H.

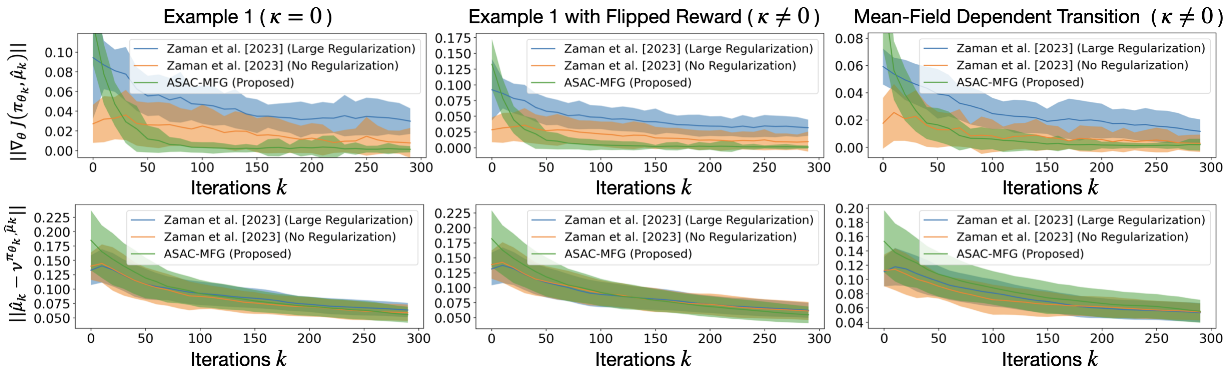

Figure 3: Algorithm performance in synthetic mean field games. First row shows sub-optimality gap of policy under latest mean field estimate. Second row shows convergence of mean field estimate to mean field induced by latest policy iterate. First column: Environment 1. Second column: Environment 2. Third column: Environment 3.

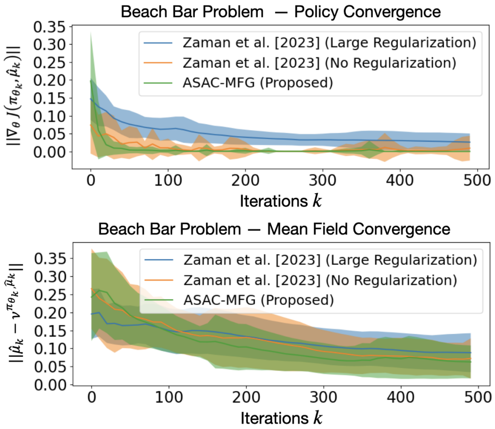

Figure 4: Algorithm performance in the beach bar problem.

of their algorithm: 1) with regularization large enough that the contraction assumption holds on the regularized game, and 2) with no regularization, which breaks the assumption.

As shown in Figure 3, all algorithms have their mean field iterates converge to the mean field induced by the latest policy iterate, while the convergence of the policy varies. ASAC−MFG and Zaman et al. [2023] with no regularization enjoy convergence to a global MFE. However, ASAC−MFG converges at a faster rate, which we believe can be attributed to the efficiency of single-loop updates. ASAC−MFG is also superior in that the convergence path has a smaller variance. The blue curve shows that while the algorithm in Zaman et al. [2023] with sufficiently large regularization may converge to a solution of the regularized problem, the persistent bias caused by the regularization prevents it from finding a solution to the original game. The theoretical result in Sec.4.1 guarantees the convergence of ASAC−MFG up to an error proportional to $\sqrt{\kappa}$, which is non-zero for Environments 2 and 3. However, we observe that in the simulations ASAC−MFG consistently converges to an equilibrium across all environments. This suggests that the herding condition with $\kappa = 0$ may only be a sufficient condition for the solvability of an MFG, a subject worth further investigating in the future.

We also apply the proposed algorithm to the beach bar problem, which is a common test case in the MFG literature. We take the formulation from Perrin et al. [2020] with $|\mathcal{S}| = 5$. The problem cannot be verified to satisfy the exact herding condition, but Figure 4 shows that ASAC−MFG still converges to an MFE, with a rate faster than that of the algorithm from Zaman et al. [2023]. The observation consistently matches that from Figure 3.

7 CONCLUSION

We made several important contributions to the literature on MFGs that are worth re-emphasizing: (i) We proposed a fast policy optimization algorithm for solving MFGs. Being the first of its kind, the algorithm is single-loop and uses a single trajectory of samples. (ii) We identified a class of MFGs – satisfying a novel herding condition – that can be optimally solved by the proposed algorithm. This expands the class of solvable MFGs, including MFGs with more than one equilibrium. It is worth noting that the proposed algorithm is general purpose, in that it can be applied to MFGs not in the proposed class to obtain stable iterates that provably converge to an approximate equilibrium (Corollary 1). (iii) The current analysis for actor-critic algorithms in average-reward MDPs requires an assumption that cannot possibly hold. By recognizing that an MFG reduces to a standard MDP with transition kernel and reward independent of the mean field, we showed that our main result leads to a finite-sample analysis of an actor-critic algorithm for average-reward MDPs that matches the state-of-the-art complexity without the untenable assumption.

While the mean-field considered in this paper is a distribution over the states, we expect the algorithm and the analysis to go through with minimal modifications when the mean-field is a more general distribution over states and actions. An important future work is to characterize the explicit connections between the different known solvable classes of MFGs (cf. 1).

Disclaimer

This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates (”JP Morgan”) and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

References

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

Clémence Alasseur, Imen Ben Taher, and Anis Matoussi. An extended mean field game for storage in smart grids. *Journal of Optimization Theory and Applications*, 184: 644–670, 2020.

Berkay Anahtarcı, Can Deha Karıksız, and Naci Saldi. Value iteration algorithm for mean-field games. *Systems & Control Letters*, 143:104744, 2020.

Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Q-learning in regularized mean-field games. *Dynamic Games and Applications*, 13(1):89–117, 2023.

Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Laurière. Unified reinforcement q-learning for mean field game and control problems. *Mathematics of Control, Signals, and Systems*, 34(2):217–271, 2022.

Andrea Angiuli, Jean-Pierre Fouque, Mathieu Laurière, and Mengrui Zhang. Convergence of multi-scale reinforcement q-learning algorithms for mean field game and control problems. *arXiv preprint arXiv:2312.06659*, 2023.

René Carmona, Kenza Hamidouche, Mathieu Laurière, and Zongjun Tan. Linear-quadratic zero-sum mean-field type games: Optimality conditions and policy optimization. *Journal of Dynamics and Games*, 8(4):403–443, 2021.

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

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

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

Gökçe Dayanıklı and Mathieu Laurière. Machine learning methods for large population games with applications in operations research. In *Tutorials in Operations Research: Smarter Decisions for a Better World*, pages 50–89. INFORMS, 2024.

Gökçe Dayanikli and Mathieu Lauriere. Multi-population mean field games with multiple major players: Application to carbon emission regulations. In *2024 American Control Conference (ACC)*, pages 5075–5081. IEEE, 2024.

Jodi Dianetti, Salvatore Federico, Giorgio Ferrari, and Giuseppe Floccari. Multiple equilibria in mean-field game models for large oligopolies with strategic complementarities. *arXiv preprint arXiv:2401.17034*, 2024.

Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for fisher-non-degenerate policies. In *International Conference on Machine Learning*, pages 9827–9869. PMLR, 2023.

Swetha Ganesh, Washim Uddin Mondal, and Vaneet Aggarwal. Variance-reduced policy gradient approaches for infinite horizon average reward markov decision processes. *arXiv preprint arXiv:2404.02108*, 2024.

Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Olivier Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: The mean-field game viewpoint. *arXiv preprint arXiv:2106.03787*, 2021.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. *Advances in neural information processing systems*, 32, 2019.

Xin Guo, Anran Hu, and Junzi Zhang. Mf-omo: An optimization formulation of mean-field games. *SIAM Journal on Control and Optimization*, 62(1):243–270, 2024.

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

Minyi Huang, Peter E Caines, and Roland P Malhame. Large-population cost-coupled lqg problems with nonuniform agents: Individual-mass behavior and decentralized ε-nash equilibria. *IEEE transactions on automatic control*, 52(9):1560–1571, 2007.

Yanxiang Jiang, Yabai Hu, Mehdi Bennis, Fu-Chun Zheng, and Xiaohu You. A mean field game-based distributed edge caching in fog radio access networks. *IEEE Transactions on Communications*, 68(3):1567–1580, 2019.

Navdeep Kumar, Yashaswini Murthy, Itai Shufaro, Kfir Y Levy, R Srikant, and Shie Mannor. On the global convergence of policy gradient in average reward markov decision processes. *arXiv preprint arXiv:2403.06806*, 2024.

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

Lixin Li, Qianqian Cheng, Xiao Tang, Tong Bai, Wei Chen, Zhiguo Ding, and Zhu Han. Resource allocation for noma-mec systems in ultra-dense networks: A learning aided mean-field game approach. *IEEE Transactions on Wireless Communications*, 20(3):1487–1500, 2020.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636, 2020.

Debmalya Mandal, Stelios Triantafyllou, and Goran Radanovic. Performative reinforcement learning. In *International Conference on Machine Learning*, pages 23642–23680. PMLR, 2023.

Weichao Mao, Haoran Qiu, Chen Wang, Hubertus Franke, Zbigniew Kalbarczyk, Ravishankar Iyer, and Tamer Basar. A mean-field game approach to cloud resource management with function approximation. *Advances in Neural Information Processing Systems*, 35:36243–36258, 2022.

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

Dheeraj Narasimha, Srinivas Shakkottai, and Lei Ying. A mean field game analysis of distributed mac in ultra-dense multichannel wireless networks. In *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pages 1–10, 2019.

Marcel Nutz, Jaime San Martin, and Xiaowei Tan. Convergence to the mean field game limit: A case study. *The Annals of Applied Probability*, 30(1), 2020.

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

Prashansa Panda and Shalabh Bhatnagar. Critic-actor for average reward mdps with function approximation: A finite-time analysis. *arXiv preprint arXiv:2402.01371*, 2024.

Julien Perolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. Scaling up mean field games with online mirror descent. *arXiv preprint arXiv:2103.00623*, 2021.

Sarah Perrin, Julien Pérolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in neural information processing systems*, 33:13199–13213, 2020.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

John N Tsitsiklis and Benjamin Van Roy. Average cost temporal-difference learning. *Automatica*, 35(11):1799–1808, 1999.

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

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

Jiachen Yang, Xiaojing Ye, Rakshit Trivedi, Huan Xu, and Hongyuan Zha. Learning deep mean field games for modeling large population behavior. In *International Conference on Learning Representations*, 2018.

Batuhan Yardim, Semih Cayci, Matthieu Geist, and Niao He. Policy mirror ascent for efficient and independent learning in mean field games. In *International Conference on Machine Learning*, pages 39722–39754. PMLR, 2023.

Batuhan Yardim, Artur Goldman, and Niao He. When is mean-field reinforcement learning tractable and relevant? *arXiv preprint arXiv:2402.05757*, 2024.

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

Muhammad Aneeq uz Zaman, Alec Koppel, Sujay Bhatt, and Tamer Basar. Oracle-free reinforcement learning in mean-field games along a single sample path. In *International Conference on Artificial Intelligence and Statistics*, pages 10178–10206. PMLR, 2023.

Sihan Zeng and Thinh Doan. Fast two-time-scale stochastic gradient method with applications in reinforcement learning. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 5166–5212. PMLR, 2024.

Sihan Zeng, Thinh T Doan, and Justin Romberg. Finite-time complexity of online primal-dual natural actor-critic algorithm for constrained markov decision processes. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 4028–4033. IEEE, 2022.

Sihan Zeng, Thinh T Doan, and Justin Romberg. A two-time-scale stochastic optimization framework with applications in control and reinforcement learning. *SIAM Journal on Optimization*, 34(1):946–976, 2024.

Chenyu Zhang, Xu Chen, and Xuan Di. Stochastic semi-gradient descent for learning mean field games with population-aware function approximation. *arXiv preprint arXiv:2408.08192*, 2024a.

Fengzhuo Zhang, Vincent Tan, Zhaoran Wang, and Zhuoran Yang. Learning regularized monotone graphon mean-field games. *Advances in Neural Information Processing Systems*, 36, 2024b.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Basar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6):3586–3612, 2020.

Sheng Zhang, Zhe Zhang, and Siva Theja Maguluri. Finite sample analysis of average-reward td learning and q-learning. *Advances in Neural Information Processing Systems*, 34:1230–1242, 2021a.

Yaoyu Zhang, Jian Sun, and Chenye Wu. Vehicle-to-grid coordination via mean field game. *IEEE Control Systems Letters*, 6:2084–2089, 2021b.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for sarsa with linear function approximation. *Advances in neural information processing systems*, 32, 2019.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Not Applicable. No training involved for the problem we study.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes. We plot both mean and standard deviation in all figures.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Not Applicable. Experiments are all small-scale and no more than 30 minutes to run on a standard computer.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable.
   (b) The license information of the assets, if applicable. Not Applicable.
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.
   (d) Information about consent from data providers/curators. Not Applicable.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

Contents

A Notations and Frequently Used Identities

We introduce a few shorthand notations frequently used in the analysis. First, we define

$$\begin{aligned}
F(\theta,V,\mu,s,a,s') &\triangleq (r(s,a,\mu)+V(s')-V(s))\nabla_\theta \log\pi_\theta(a\mid s),\\
G^V(V,J,\mu,s,a,s') &\triangleq (r(s,a,\mu)-J+V(s')-V(s))e_s,\\
G^J(J,\mu,s,a) &\triangleq c_J(r(s,a,\mu)-J),\\
G(V,J,\mu,s,a,s') &\triangleq \left[\begin{array}{c} G^V(V,J,\mu,s,a,s') \\ G^J(J,\mu,s,a)\end{array}\right] = \left[\begin{array}{c} (r(s,a,\mu)-J+V(s')-V(s))e_s \\ c_J(r(s,a,\mu)-J)\end{array}\right],\\
H(\mu,s) &\triangleq e_s-\mu.
\end{aligned} \tag{22}$$

Then, the update of f_k, g_k^V, g_k^J, and h_k in Algorithm 1 can be expressed as

$$\begin{aligned}
f_{k+1} &= (1-\lambda_k)f_k+\lambda_k F(\theta_k,\hat{V}_k,\hat{\mu}_k,s_k,a_k,s_{k+1}),\\
g_{k+1}^V &= (1-\lambda_k)g_k^V+\lambda_k G^V(\hat{V}_k,\hat{J}_k,\hat{\mu}_k,s_k,a_k,s_{k+1}),\\
g_{k+1}^J &= (1-\lambda_k)g_k^J+\lambda_k G^J(\hat{J}_k,\hat{\mu}_k,s_k,a_k),\\
h_{k+1} &= (1-\lambda_k)h_k+\lambda_k H(\hat{\mu}_k,s_k).
\end{aligned}$$

Denote $g_k=[(g_k^V)^\top, g_k^J]^\top$. The update of g_k is

$$g_{k+1} = \left[\begin{array}{c} g_{k+1}^V \\ g_{k+1}^J\end{array}\right] = (1-\lambda_k)g_k+\lambda_k G(\hat{V}_k,\hat{J}_k,\hat{\mu}_k,s_k,a_k,s_{k+1}).$$

We also define

$$
\begin{aligned}
\bar{F}(\theta, V, \mu) &\triangleq \mathbb{E}_{s\sim\nu^{\pi_\theta,\,\mu}, a\sim\pi_\theta(\cdot|s), s'\sim\mathcal{P}^{\mu}(\cdot|s,a)}[F(\theta, V, \mu, s, a, s')],\\
\bar{G}^V(\theta, V, J, \mu) &\triangleq \mathbb{E}_{s\sim\nu^{\pi_\theta,\,\mu}, a\sim\pi_\theta(\cdot|s), s'\sim\mathcal{P}^{\mu}(\cdot|s,a)}[G^V(V, J, \mu, s, a, s')],\\
\bar{G}^J(\theta, J, \mu) &\triangleq \mathbb{E}_{s\sim\nu^{\pi_\theta,\,\mu}, a\sim\pi_\theta(\cdot|s)}[G^J(J, \mu, s, a)],\\
\bar{G}(\theta, V, J, \mu) &\triangleq \mathbb{E}_{s\sim\nu^{\pi_\theta,\,\mu}, a\sim\pi_\theta(\cdot|s), s'\sim\mathcal{P}^{\mu}(\cdot|s,a)}[G(V, J, \mu, s, a, s')] = \left[\begin{array}{c} \bar{G}^V(\theta, V, J, \mu) \\ \bar{G}^J(\theta, J, \mu) \end{array}\right],\\
\bar{H}(\theta, \mu) &\triangleq \mathbb{E}_{s\sim\nu^{\pi_\theta,\,\mu}}[H(\mu, s)] = \mathbb{E}_{s\sim\nu^{\pi_\theta,\,\mu}}[e_s - \mu].
\end{aligned}
\tag{23}
$$

We measure the convergence of auxiliary variables f_k, g_k^V, g_k^J, and h_k by

$$
\begin{aligned}
\Delta f_k \triangleq f_k - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k), \quad \Delta g_k^V \triangleq g_k^V - \bar{G}^V(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k),\\
\Delta g_k^J \triangleq g_k^J - \bar{G}^J(\theta_k, \hat{J}_k, \hat{\mu}_k), \quad \Delta h_k \triangleq h_k - \bar{H}(\theta_k, \hat{\mu}_k),
\end{aligned}
$$

and denote

$$
\Delta g_k = \left[\begin{array}{c} \Delta g_k^V \\ \Delta g_k^J \end{array}\right] = g_k - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k).
$$

We use $\ell(\pi)$ to denote the cumulative reward collected by policy π under the induced mean field $\mu^\star(\pi)$

$$
\ell(\pi) \triangleq J(\pi, \mu^\star(\pi)).
$$

This is well-defined since $\mu^\star(\pi)$ is unique.

We denote by $\mathcal{F}_k = \{s_0, a_0, s_1, a_1 \cdots, s_k, a_k\}$ denote the filtration (set of all randomness information) up to iteration k. Given two probability distributions ϕ_1 and ϕ_2 over space $\mathcal{X}$, their TV distance is defined as

$$
d_{\text{TV}}(\phi_1, \phi_2) = \frac{1}{2}\sup_{\psi:\mathcal{X}\to[-1,1]} \left|\int \psi d\phi_1 - \int \psi d\phi_2\right|. \tag{24}
$$

Under Assumptions 1 and 3, it can be shown using an argument similar to Lemma B.1 of Wu et al. [2020] that there exists a constant L_{TV} depending only on $|\mathcal{A}|$, L, C_0, and C_1 such that for all $\pi_1, \pi_2, \mu_1, \mu_2$

$$
d_{TV}(\nu^{\pi_1,\,\mu_1} \otimes \pi_1 \otimes \mathcal{P}^{\mu_1}, \nu^{\pi_2,\,\mu_2} \otimes \pi_2 \otimes \mathcal{P}^{\mu_2}) \leq L_{TV}(\|\pi_1 - \pi_2\| + \|\mu_1 - \mu_2\|). \tag{25}
$$

Without loss of generality, we assume $L \geq 1$, a condition that we will sometimes use to simplify and combine terms.

A.1 Mixing Time

An immediate consequence of Assumption 1 is that the Markov chain under any policy and mean field has a geometric mixing time.

Definition 3 *Consider a Markov chain $\{\hat{s}_k\}$ generated according to $\hat{s}_k \sim P^{\pi,\,\mu}(\cdot \mid \hat{s}_{k-1})$, for which $\nu^{\pi,\,\mu}$ is the stationary distribution. For any $c > 0$, the c-mixing time of the Markov chain is*

$$
\tau^{\pi,\,\mu}(c) \triangleq \min\left\{k \in \mathbb{N} : \sup_s d_{TV}\big(\mathbb{P}(\hat{s}_k = \cdot \mid \hat{s}_0 = s),\, \nu^{\pi,\,\mu}(\cdot)\big) \leq c\right\}.
$$

The mixing time measures time for the samples of the Markov chain to approach its stationary distribution in TV distance. We define $\tau_k \triangleq \sup_{\pi,\,\mu} \tau^{\pi,\,\mu}(\alpha_k)$ as the time when the TV distance drops below α_k, where α_k is a step size for the policy parameter update in Algorithm 1. Under Assumption 1, it is obvious that there exists a constant C as a function of C_0, C_1 such that

$$
\tau_k \leq C\log(1/\alpha_k) = C\log(\frac{(k+1)^{1/2}}{\alpha_0}) = \frac{C}{2}\log(k+1) - C\log(\alpha_0).
$$

A.2 Supporting Lemmas

The value function $V^{\pi_\theta,\,\mu}$ is Lipschitz in both θ and μ, as shown in the lemma below.

Lemma 1 *Under Assumption 3, there exist a bounded constant $L_V \geq 1$ such that for any policy parameter θ_1, θ_2 and mean field μ_1, μ_2, we have*

$$\begin{aligned}
\|\Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_1},\,\mu_1} - V^{\pi_{\theta_2},\,\mu_2})\| &\leq L_V\left(\|\theta_1 - \theta_2\| + \|\mu_1 - \mu_2\|\right),\\
\|J(\pi_{\theta_1},\,\mu_1) - J(\pi_{\theta_2},\,\mu_2)\| &\leq L_V\left(\|\theta_1 - \theta_2\| + \|\mu_1 - \mu_2\|\right),\\
\|\nabla_\theta J(\pi_{\theta_1},\,\mu_1) - \nabla_\theta J(\pi_{\theta_2},\,\mu_2)\| &\leq L_V\left(\|\theta_1 - \theta_2\| + \|\mu_1 - \mu_2\|\right),\\
\|\nabla_\mu J(\pi_{\theta_1},\,\mu_1) - \nabla_\mu J(\pi_{\theta_2},\,\mu_2)\| &\leq L_V\left(\|\theta_1 - \theta_2\| + \|\mu_1 - \mu_2\|\right).
\end{aligned}$$

We establish the boundedness of the operators F, G, and H.

Lemma 2 *For any $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $V \in \mathbb{R}^{|\mathcal{S}|}$ with norm bounded by B_V, $J \in [0,1]$, $\mu \in \Delta_{\mathcal{S}}$, and s, a, s', we have*

$$\|F(\theta, V, \mu, s, a, s')\| \leq B_F, \|G(V, J, \mu, s, a, s')\| \leq B_G, \|H(\mu, s)\| \leq B_H,$$

where $B_F = B_V + 1$, $B_G = 2(B_V + c_J + 2)$, $B_H = 2$.

Since f_k, g_k^V, g_k^J, and h_k are simply convex combination with the operators F, G^V, G^J, and H, Lemma 2 implies for all k

$$\|f_k\| \leq B_F, \quad \|g_k^V\| \leq B_G, \quad |g_k^J| \leq B_G, \quad \|h_k\| \leq B_H.$$

We also establish the Lipschitz continuity of a few important operators.

Lemma 3 *We have for any $\theta_1, \theta_2 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\mu_1, \mu_2 \in \Delta_{\mathcal{S}}$, $V_1, V_2 \in \mathbb{R}^{|\mathcal{S}|}$, and $J_1, J_2 \in \mathbb{R}$*

$$\begin{aligned}
&\|\bar{F}(\theta_1, V_1, \mu_1) - \bar{F}(\theta_2, V_2, \mu_2)\| \leq L_F\Big(\|\theta_1 - \theta_2\| + \|\Pi_{\mathcal{E}_\perp}(V_1 - V_2)\| + \|\mu_1 - \mu_2\|\Big)\\
&\|\bar{G}(\theta_1, V_1, J_1, \mu_1) - \bar{G}(\theta_2, V_2, J_2, \mu_2)\|\\
&\qquad\qquad\qquad\leq L_G\Big(\|\theta_1 - \theta_2\| + \|\Pi_{\mathcal{E}_\perp}(V_1 - V_2)\| + |J_1 - J_2| + \|\mu_1 - \mu_2\|\Big),\\
&\|\bar{H}(\theta_1, \mu_1) - \bar{H}(\theta_2, \mu_2)\| \leq L_H\Big(\|\theta_1 - \theta_2\| + \|\mu_1 - \mu_2\|\Big),
\end{aligned}$$

where the constants are $L_F = 10B_V + L + 2B_F L_{TV} + 5$, $L_G = 2B_G L_{TV} + (L+1)(c_J + 1) + 2$, and $L_H = L + 1$.

As a result of Lemma 3, we can establish the following bounds on the energy of the auxiliary variables f_k, g_k, and h_k.

Lemma 4 *We have for any $k \geq 0$*

$$\begin{aligned}
\|f_k\| &\leq \|\Delta f_k\| + L_F\sqrt{\varepsilon_k^V} + L_F(L_V + 1)\sqrt{\varepsilon_k^\mu} + \sqrt{\varepsilon_k^\pi},\\
\|g_k\| &\leq \|\Delta g_k\| + L_G\sqrt{\varepsilon_k^V} + L_G\sqrt{\varepsilon_k^J},\\
\|h_k\| &\leq \|\Delta h_k\| + L_H\sqrt{\epsilon_k^\mu}.
\end{aligned}$$

Also as a consequence of Assumption 1, the following lemma holds which states that the Bellman backup operator of the value function is almost everywhere contractive (except along the direction of the all-one vector). This lemma is adapted from Zhang et al. [2021a][Lemma 2] and Tsitsiklis and Van Roy [1999][Lemma 7].

Lemma 5 *Recall the definition of $\mathcal{E}_\perp$ in Sec.4.1. There exists a constant $\gamma \in (0,1)$ such that for any θ, μ and $V \in \mathcal{E}_\perp$*

$$V^\top \mathbb{E}_{s\sim\nu^{\pi_\theta,\,\mu}, a\sim\pi_\theta(\cdot|s), s'\sim\mathcal{P}^\mu(\cdot|s,a)}[e_s(e_{s'} - e_s)^\top]V \leq -\gamma\|V\|^2.$$

B Proof of Main Theorem

B.1 Intermediate Results

The proof of Theorem 1 relies critically on the iteration-wise convergence of policy iterate θ_k, mean field iterate $\hat{\mu}_k$, value function estimate $\hat{V}_k$, $\hat{J}_k$, and auxiliary variables f_k, h_k, and g_k, which we bound individually in the propositions below.

B.1.1 Convergence of Policy Iterate

Proposition 1 *Under Assumptions 1-3, we have*

$$\begin{aligned}\ell(\pi_{\theta_k}) - \ell(\pi_{\theta_{k+1}}) \leq& -\frac{\rho\alpha_k}{2}\varepsilon_k^{\pi} + \rho\alpha_k\|\Delta f_k\|^2 \\ &+ \rho L_F^2\alpha_k(\varepsilon_k^V + \varepsilon_k^{\mu}) + \frac{\rho L_V B_F^2\alpha_k^2}{2} + B_F\alpha_k\kappa.\end{aligned}$$

Proposition 2 *Under Assumptions 1-3, we have for all $k \geq \tau_k$*

$$\begin{aligned}&\mathbb{E}[\|\Delta f_{k+1}\|^2] \\ &\leq (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] + (-\frac{\lambda_k}{2} + \lambda_k^2 + \frac{48L_F^2\alpha_k^2}{\lambda_k})\mathbb{E}[\|\Delta f_k\|^2] \\ &\quad + \frac{36L_F^2\beta_k^2}{\lambda_k}\mathbb{E}[\|\Delta g_k\|^2] + \frac{24L_F^2L_H^2\xi_k^2}{\lambda_k}\mathbb{E}[\|\Delta h_k\|^2] + \frac{48L_F^2\alpha_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^{\pi}] + \frac{216L_F^4L_V^2\xi_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^{\mu}] \\ &\quad + \frac{96L_F^4L_G^2\beta_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^V] + \frac{48L_F^2L_G^2\beta_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^J] + (28L + 2|\mathcal{A}|)L_F L_{TV}B_F^3B_GB_H^2\tau_k^2\lambda_k\lambda_{k-\tau_k}.\end{aligned}$$

The proofs of Propositions 1 and 2 can be found in Sec.D.1 and D.2.

B.1.2 Convergence of Mean Field Estimate

Proposition 3 *Under Assumptions 1-2, we have for all k*

$$\varepsilon_{k+1}^{\mu} \leq (1 - \frac{(1-\delta)\xi_k}{8})\varepsilon_k^{\mu} + \frac{8\xi_k}{1-\delta}\|\Delta h_k\|^2 + \frac{32L^2\alpha_k^2}{(1-\delta)\xi_k}\left(\|\Delta f_k\|^2 + L_F^2\varepsilon_k^V + \varepsilon_k^{\pi}\right) + 9L^2B_F^2B_H^2\xi_k^2.$$

Proposition 4 *Under Assumptions 1-3, we have for all $k \geq \tau_k$*

$$\begin{aligned}&\mathbb{E}[\|\Delta h_{k+1}\|^2] \\ &\leq (1-\lambda_k)\mathbb{E}[\|\Delta h_k\|^2] + (-\frac{\lambda_k}{2} + \lambda_k^2 + \frac{16L_H^2\xi_k^2}{\lambda_k})\mathbb{E}[\|\Delta h_k\|^2] + \frac{32L_H^2\alpha_k^2}{\lambda_k}\mathbb{E}[\|\Delta f_k\|^2] \\ &\quad + \frac{32L_H^2L_F^2\alpha_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^V] + \frac{144L_F^2L_V^2L_H^4\xi_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^{\mu}] + \frac{32L_H^2\alpha_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^{\pi}] + 24LB_FB_H^2\tau_k^2\lambda_k\lambda_{k-\tau_k}.\end{aligned}$$

The proofs of Propositions 3 and 4 can be found in Sec.D.3 and D.4.

B.1.3 Convergence of Valuation Function Estimate

Proposition 5 *Under Assumptions 1-3,*

$$\begin{aligned}\varepsilon_{k+1}^V + \varepsilon_{k+1}^J \leq& (1 - \frac{\gamma\beta_k}{4})(\varepsilon_k^V + \varepsilon_k^J) + \frac{128L_V^2\alpha_k^2}{\gamma\beta_k}\|\Delta f_k\|^2 + \frac{8\beta_k}{\gamma}\|\Delta g_k\|^2 + \frac{64L_V^2\xi_k^2}{\gamma\beta_k}\|\Delta h_k\|^2 \\ &+ \frac{128L_V^2\alpha_k^2}{\gamma\beta_k}(L_F^2\varepsilon_k^V + \varepsilon_k^{\pi}) + \frac{192L_V^2\xi_k^2}{\gamma\beta_k}\varepsilon_k^{\mu} + 28L_V^2B_F^2B_G^2B_H^2\beta_k^2.\end{aligned}$$

Proposition 6 *Under Assumptions 1-3, we have for all $k \geq \tau_k$*

$$\begin{aligned}\mathbb{E}[\|\Delta g_{k+1}\|^2] \leq& (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] + (-\frac{\lambda_k}{2} + \lambda_k^2 + \frac{72|\mathcal{S}|L_G^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta g_k\|^2] + \frac{48L_G^2\alpha_k^2}{\lambda_k}\mathbb{E}[\|\Delta f_k\|^2] \\ &+ \frac{24L_G^2\xi_k^2}{\lambda_k}\mathbb{E}[\|\Delta h_k\|^2] + \frac{48L_G^2\alpha_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^{\pi}] + \frac{216L_F^2L_G^2L_H^2L_V^2\xi_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^{\mu}] \\ &+ \frac{120|\mathcal{S}|L_F^2L_G^4\beta_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^V + \varepsilon_k^J] + (30L + 2|\mathcal{A}|)L_F L_{TV}B_FB_G^2B_H\tau_k^2\lambda_k\lambda_{k-\tau_k}.\end{aligned}$$

The proofs of Propositions 5 and 6 can be found in Sec.D.5 and D.6.

B.2 Proof of Theorem 1

The exact requirements on $\lambda_0, \alpha_0, \beta_0, \xi_0$ include $c_J \geq 1/\gamma$, $\alpha_0 \leq \xi_0 \leq \beta_0 \leq \lambda_0$, and

$$\begin{aligned}\alpha_0 &\leq \min\left\{\frac{1}{192(L_F^2+L_G^2+L_H^2+L_V^2+L^2/(1-\delta)+\rho)}\lambda_0,\, C_\beta\beta_0,\, C_\xi\xi_0\right\},\\ \xi_0 &\leq \min\left\{\frac{\lambda_0}{64(L_H^2L_F^2+L_G^2+L_V^2/\gamma+1/(1-\delta))},\, \frac{(1-\delta)\gamma\beta_0}{6912(L_F^4L_V^2+L_F^2L_G^2L_H^2L_V^2+L_F^2L_H^4L_V^2+L_V^2)}\right\},\\ \beta_0 &\leq \min\left\{\frac{\lambda_0}{72|\mathcal{S}|L_G^2+36L_F^2+8/\gamma},\, \frac{\gamma}{4L_G^2},\frac{1-\delta}{2L_H^2}\right\},\quad \lambda_0\leq\frac{1}{4},\end{aligned} \tag{26}$$

where $C_\xi = \min\{\frac{(1-\delta)}{32\rho L_F^2}, \frac{1-\delta}{4\rho}, \frac{L_H}{2L_FL_V}, \frac{1-\delta}{16LL_FL_V}\}$ and

$$\begin{aligned}C_\beta = \min\Big\{&\frac{\gamma}{4}, \frac{\rho\gamma}{512(L_F^2+L_G^2+L_H^2+L_V^2+L^2/(1-\delta)},\\ &\sqrt{\frac{\gamma}{3456|\mathcal{S}|(L_F^4L_G^4+L_F^2L_H^2+\rho L_F^2+L_V^2/\gamma+L^2L_F^2(1-\delta))}}, \frac{\gamma}{2\rho}\Big\}.\end{aligned}$$

We note that such parameters can always chosen with no conflict in any MFG.

We consider the potential function

$$\mathcal{L}_k = \mathbb{E}[\|\Delta f_k\|^2 + \|\Delta g_k\|^2 + \|\Delta h_k\|^2 - \ell(\pi_{\theta_k}) + \varepsilon_k^V + \varepsilon_k^J + \varepsilon_k^\mu].$$

Collecting the bounds from Propositions 1-6, we have for all $k \geq \tau_k$

$$\begin{aligned}&\mathcal{L}_{k+1}\\ &= \mathbb{E}[\|\Delta f_{k+1}\|^2 + \|\Delta g_{k+1}\|^2 + \|\Delta h_{k+1}\|^2 - \ell(\pi_{\theta_{k+1}}) + \varepsilon_{k+1}^V + \varepsilon_{k+1}^J + \varepsilon_{k+1}^\mu]\\ &\leq (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] + (-\frac{\lambda_k}{2} + \lambda_k^2 + \frac{48L_F^2\alpha_k^2}{\lambda_k})\mathbb{E}[\|\Delta f_k\|^2]\\ &\quad + \frac{36L_F^2\beta_k^2}{\lambda_k}\mathbb{E}[\|\Delta g_k\|^2] + \frac{24L_F^2L_H^2\xi_k^2}{\lambda_k}\mathbb{E}[\|\Delta h_k\|^2] + \frac{48L_F^2\alpha_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^\pi] + \frac{216L_F^4L_V^2\xi_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^\mu]\\ &\quad + \frac{96L_F^4L_G^2\beta_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^V] + \frac{48L_F^2L_G^2\beta_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^J] + (28L+2|\mathcal{A}|)L_FL_{TV}B_F^3B_GB_H^2\tau_k^2\lambda_k\lambda_{k-\tau_k}\\ &\quad + (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] + (-\frac{\lambda_k}{2} + \lambda_k^2 + \frac{72|\mathcal{S}|L_G^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta g_k\|^2] + \frac{48L_G^2\alpha_k^2}{\lambda_k}\mathbb{E}[\|\Delta f_k\|^2]\\ &\quad + \frac{24L_G^2\xi_k^2}{\lambda_k}\mathbb{E}[\|\Delta h_k\|^2] + \frac{48L_G^2\alpha_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^\pi] + \frac{216L_F^2L_G^2L_H^2L_V^2\xi_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^\mu]\\ &\quad + \frac{120|\mathcal{S}|L_F^2L_G^4\beta_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^V + \varepsilon_k^J] + (30L+2|\mathcal{A}|)L_FL_{TV}B_FB_G^2B_H\tau_k^2\lambda_k\lambda_{k-\tau_k}\\ &\quad + (1-\lambda_k)\mathbb{E}[\|\Delta h_k\|^2] + (-\frac{\lambda_k}{2} + \lambda_k^2 + \frac{16L_H^2\xi_k^2}{\lambda_k})\mathbb{E}[\|\Delta h_k\|^2] + \frac{32L_H^2\alpha_k^2}{\lambda_k}\mathbb{E}[\|\Delta f_k\|^2]\\ &\quad + \frac{32L_H^2L_F^2\alpha_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^V] + \frac{144L_F^2L_V^2L_H^4\xi_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^\mu] + \frac{32L_H^2\alpha_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^\pi] + 24LB_FB_H^2\tau_k^2\lambda_k\lambda_{k-\tau_k}\\ &\quad - \mathbb{E}[\ell(\pi_{\theta_k})] - \frac{\rho\alpha_k}{2}\mathbb{E}[\varepsilon_k^\pi] + \rho\alpha_k\mathbb{E}[\|\Delta f_k\|^2]\\ &\quad + \rho L_F^2\alpha_k\mathbb{E}[\varepsilon_k^V + \varepsilon_k^\mu] + \frac{\rho L_VB_F^2\alpha_k^2}{2} + B_F\alpha_k\kappa\\ &\quad + (1-\frac{\gamma\beta_k}{4})\mathbb{E}[\varepsilon_k^V + \varepsilon_k^J] + \frac{128L_V^2\alpha_k^2}{\gamma\beta_k}\mathbb{E}[\|\Delta f_k\|^2] + \frac{8\beta_k}{\gamma}\mathbb{E}[\|\Delta g_k\|^2] + \frac{64L_V^2\xi_k^2}{\gamma\beta_k}\mathbb{E}[\|\Delta h_k\|^2]\\ &\quad + \frac{128L_V^2\alpha_k^2}{\gamma\beta_k}(L_F^2\mathbb{E}[\varepsilon_k^V] + \mathbb{E}[\varepsilon_k^\pi]) + \frac{192L_V^2\xi_k^2}{\gamma\beta_k}\mathbb{E}[\varepsilon_k^\mu] + 28L_V^2B_F^2B_G^2B_H^2\beta_k^2\\ &\quad +(1-\frac{(1-\delta)\xi_k}{8})\mathbb{E}[\varepsilon_k^\mu] + \frac{8\xi_k}{1-\delta}\mathbb{E}[\|\Delta h_k\|^2] + \frac{32L^2\alpha_k^2}{(1-\delta)\xi_k}\mathbb{E}[\|\Delta f_k\|^2 + L_F^2\varepsilon_k^V + \varepsilon_k^\pi] + 9L^2B_F^2B_H^2\xi_k^2\end{aligned}$$

$$
\begin{aligned}
&\leq (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2 + \|\Delta g_k\|^2 + \|\Delta h_k\|^2] - \mathbb{E}[\ell(\pi_{\theta_k})] - \frac{\rho\alpha_k}{4}\mathbb{E}[\varepsilon_k^\pi] \\
&\quad + (1-\frac{\gamma\beta_k}{8})\mathbb{E}[\varepsilon_k^V + \varepsilon_k^J] + (1-\frac{(1-\delta)\xi_k}{16})\mathbb{E}[\varepsilon_k^\mu] + B_F\alpha_k\kappa \\
&\quad + (28L+2|\mathcal{A}|)L_F L_{TV} B_F^3 B_G B_H^2 \tau_k^2 \lambda_k \lambda_{k-\tau_k} + (30L+2|\mathcal{A}|)L_F L_{TV} B_F B_G^2 B_H \tau_k^2 \lambda_k \lambda_{k-\tau_k} \\
&\quad + 24LB_F B_H^2 \tau_k^2 \lambda_k \lambda_{k-\tau_k} + \frac{\rho L_V B_F^2 \alpha_k^2}{2} + 28L_V^2 B_F^2 B_G^2 B_H^2 \beta_k^2 + 9L^2 B_F^2 B_H^2 \xi_k^2 \\
&\quad + \underbrace{(-\frac{\lambda_k}{2} + \lambda_k^2 + \frac{48L_F^2\alpha_k^2}{\lambda_k} + \frac{48L_G^2\alpha_k^2}{\lambda_k} + \frac{32L_H^2\alpha_k^2}{\lambda_k} + \rho\alpha_k + \frac{128L_V^2\alpha_k^2}{\gamma\beta_k} + \frac{32L^2\alpha_k^2}{(1-\delta)\lambda_k})}_{A_1}\mathbb{E}[\|\Delta f_k\|^2] \\
&\quad + \underbrace{(-\frac{\lambda_k}{2} + \lambda_k^2 + \frac{72|\mathcal{S}|L_G^2\beta_k^2}{\lambda_k} + \frac{36L_F^2\beta_k^2}{\lambda_k} + \frac{8\beta_k}{\gamma})}_{A_2}\mathbb{E}[\|\Delta g_k\|^2] \\
&\quad + \underbrace{(-\frac{\lambda_k}{2} + \lambda_k^2 + \frac{16L_H^2\xi_k^2}{\lambda_k} + \frac{24L_F^2 L_H^2\xi_k^2}{\lambda_k} + \frac{24L_G^2\xi_k^2}{\lambda_k} + \frac{64L_V^2\xi_k^2}{\gamma\lambda_k} + \frac{8\xi_k}{1-\delta})}_{A_3}\mathbb{E}[\|\Delta h_k\|^2] \\
&\quad + \underbrace{(-\frac{\rho\alpha_k}{4} + \frac{48L_F^2\alpha_k^2}{\lambda_k} + \frac{48L_G^2\alpha_k^2}{\lambda_k} + \frac{32L_H^2\alpha_k^2}{\lambda_k} + \frac{128L_V^2\alpha_k^2}{\gamma\beta_k} + \frac{32L^2\alpha_k^2}{(1-\delta)\lambda_k})}_{A_4}\mathbb{E}[\varepsilon_k^\pi] \\
&\quad + \underbrace{(-\frac{\gamma\beta_k}{8} + \frac{96L_F^4 L_G^2\beta_k^2}{\lambda_k} + \frac{120|\mathcal{S}|L_F^2 L_G^4\beta_k^2}{\lambda_k} + \frac{32L_F^2 L_H^2\alpha_k^2}{\lambda_k} + \rho L_F^2\alpha_k + \frac{128L_V^2\alpha_k^2}{\gamma\beta_k} + \frac{32L^2 L_F^2\alpha_k^2}{(1-\delta)\lambda_k})}_{A_5}\mathbb{E}[\varepsilon_k^V + \varepsilon_k^J] \\
&\quad + \underbrace{(-\frac{(1-\delta)\xi_k}{16} + \frac{216L_F^4 L_V^2\xi_k^2}{\lambda_k} + \frac{216L_F^2 L_G^2 L_H^2 L_V^2\xi_k^2}{\lambda_k} + \frac{144L_F^2 L_H^4 L_V^2\xi_k^2}{\lambda_k} + \rho L_F^2\alpha_k + \frac{192L_V^2\xi_k^2}{\gamma\beta_k})}_{A_6}\mathbb{E}[\varepsilon_k^\mu].
\end{aligned}
\tag{27}
$$

We show that the terms A_1-A_6 are all non-positive under the step size conditions in (26). First, under the step size condition $\alpha_k \leq \frac{\gamma}{4}\beta_k$, $\lambda_k \leq 1/4$, and $\alpha_k \leq (192(L_F^2 + L_G^2 + L_H^2 + L_V^2 + L^2/(1-\delta) + \rho))^{-1}\lambda_k$

$$
\begin{aligned}
A_1 &= -\frac{\lambda_k}{2} + \lambda_k^2 + \frac{48L_F^2\alpha_k^2}{\lambda_k} + \frac{48L_G^2\alpha_k^2}{\lambda_k} + \frac{32L_H^2\alpha_k^2}{\lambda_k} + \rho\alpha_k + \frac{128L_V^2\alpha_k^2}{\gamma\beta_k} + \frac{32L^2\alpha_k^2}{(1-\delta)\lambda_k} \\
&\leq -\frac{\lambda_k}{4} + \frac{48(L_F^2 + L_G^2 + L_H^2 + L^2/(1-\delta))\alpha_k^2}{\lambda_k} + \rho\alpha_k + 32L_V^2\alpha_k \\
&\leq -\frac{\lambda_k}{4} + 48(L_F^2 + L_G^2 + L_H^2 + L_V^2 + L^2/(1-\delta) + \rho)\alpha_k \\
&\leq 0.
\end{aligned}
\tag{28}
$$

Next, under the step size condition $\lambda_k \leq 1/4$ and $\beta_k \leq (72|\mathcal{S}|L_G^2 + 36L_F^2 + 8/\gamma)^{-1}\lambda_k$

$$
\begin{aligned}
A_2 &= -\frac{\lambda_k}{2} + \lambda_k^2 + \frac{72|\mathcal{S}|L_G^2\beta_k^2}{\lambda_k} + \frac{36L_F^2\beta_k^2}{\lambda_k} + \frac{8\beta_k}{\gamma} \\
&\leq -\frac{\lambda_k}{4} + (72|\mathcal{S}|L_G^2 + 36L_F^2 + 8/\gamma)\beta_k \\
&\leq 0.
\end{aligned}
\tag{29}
$$

Next, under the step size condition $\lambda_k \leq 1/4$ and $\xi_k \leq (64(L_H^2 L_F^2 + L_G^2 + L_V^2/\gamma + 1/(1-\delta)))^{-1}\lambda_k$

$$
\begin{aligned}
A_3 &= -\frac{\lambda_k}{2} + \lambda_k^2 + \frac{16L_H^2\xi_k^2}{\lambda_k} + \frac{24L_F^2 L_H^2\xi_k^2}{\lambda_k} + \frac{24L_G^2\xi_k^2}{\lambda_k} + \frac{64L_V^2\xi_k^2}{\gamma\lambda_k} + \frac{8\xi_k}{1-\delta} \\
&\leq -\frac{\lambda_k}{4} + 64(L_H^2 L_F^2 + L_G^2 + L_V^2/\gamma + 1/(1-\delta))\xi_k
\end{aligned}
$$

$$\leq 0. \tag{30}$$

Next, we have

$$\begin{aligned}
A_4 &= -\frac{\rho\alpha_k}{4} + \frac{48L_F^2\alpha_k^2}{\lambda_k} + \frac{48L_G^2\alpha_k^2}{\lambda_k} + \frac{32L_H^2\alpha_k^2}{\lambda_k} + \frac{128L_V^2\alpha_k^2}{\gamma\beta_k} + \frac{32L^2\alpha_k^2}{(1-\delta)\lambda_k} \\
&\leq -\frac{\rho\alpha_k}{4} + \frac{128}{\gamma}(L_F^2 + L_G^2 + L_H^2 + L_V^2 + L^2/(1-\delta))\frac{\alpha_k^2}{\beta_k} \\
&\leq 0,
\end{aligned} \tag{31}$$

under the step size condition

$$\alpha_k \leq \frac{\rho\gamma}{512(L_F^2 + L_G^2 + L_H^2 + L_V^2 + L^2/(1-\delta)}\beta_k.$$

Then,

$$\begin{aligned}
A_5 &= -\frac{\gamma\beta_k}{8} + \frac{96L_F^4L_G^2\beta_k^2}{\lambda_k} + \frac{120|\mathcal{S}|L_F^2L_G^4\beta_k^2}{\lambda_k} + \frac{32L_F^2L_H^2\alpha_k^2}{\lambda_k} \\
&\quad + \rho L_F^2\alpha_k + \frac{128L_V^2\alpha_k^2}{\gamma\beta_k} + \frac{32L^2L_F^2\alpha_k^2}{(1-\delta)\lambda_k} \\
&\leq -\frac{\gamma\beta_k}{8} + 432|\mathcal{S}|(L_F^4L_G^4 + L_F^2L_H^2 + \rho L_F^2 + L_V^2/\gamma + L^2L_F^2(1-\delta))\frac{\alpha_k^2}{\beta_k} \\
&\leq 0,
\end{aligned} \tag{32}$$

due to the condition

$$\alpha_k \leq \sqrt{\frac{\gamma}{3456|\mathcal{S}|(L_F^4L_G^4 + L_F^2L_H^2 + \rho L_F^2 + L_V^2/\gamma + L^2L_F^2(1-\delta))}}\beta_k.$$

Finally, as a result of $\alpha_k \leq \frac{(1-\delta)}{32\rho L_F^2}\xi_k$ and $\xi_k \leq \frac{(1-\delta)\gamma}{6912(L_F^4L_V^2+L_F^2L_G^2L_H^2L_V^2+L_F^2L_H^4L_V^2+L_V^2)}\beta_k$

$$\begin{aligned}
A_6 &= -\frac{(1-\delta)\xi_k}{16} + \frac{216L_F^4L_V^2\xi_k^2}{\lambda_k} + \frac{216L_F^2L_G^2L_H^2L_V^2\xi_k^2}{\lambda_k} \\
&\quad + \frac{144L_F^2L_H^4L_V^2\xi_k^2}{\lambda_k} + \rho L_F^2\alpha_k + \frac{192L_V^2\xi_k^2}{\gamma\beta_k} \\
&\leq -\frac{(1-\delta)\xi_k}{32} + \frac{216}{\gamma}(L_F^4L_V^2 + L_F^2L_G^2L_H^2L_V^2 + L_F^2L_H^4L_V^2 + L_V^2)\frac{\xi_k^2}{\beta_k} \\
&\leq 0.
\end{aligned} \tag{33}$$

Plugging (28)-(33) into (27), we have for all $k \geq \tau_k$

$$\begin{aligned}
&\mathcal{L}_{k+1} \\
&\leq (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2 + \|\Delta g_k\|^2 + \|\Delta h_k\|^2] - \mathbb{E}[\ell(\pi_{\theta_k})] - \frac{\rho\alpha_k}{4}\mathbb{E}[\varepsilon_k^\pi] \\
&\quad + (1 - \frac{\gamma\beta_k}{8})\mathbb{E}[\varepsilon_k^V + \varepsilon_k^J] + (1 - \frac{(1-\delta)\xi_k}{16})\mathbb{E}[\varepsilon_k^\mu] + B_F\alpha_k\kappa \\
&\quad + (28L{+}2|\mathcal{A}|)L_FL_{TV}B_F^3B_GB_H^2\tau_k^2\lambda_k\lambda_{k-\tau_k} + (30L{+}2|\mathcal{A}|)L_FL_{TV}B_FB_G^2B_H\tau_k^2\lambda_k\lambda_{k-\tau_k} \\
&\quad + 24LB_FB_H^2\tau_k^2\lambda_k\lambda_{k-\tau_k} + \frac{\rho L_VB_F^2\alpha_k^2}{2} + 28L_V^2B_F^2B_G^2B_H^2\beta_k^2 + 9L^2B_F^2B_H^2\xi_k^2 \\
&\leq \mathcal{L}_k - \min\left\{\frac{\rho\alpha_k}{4}, \frac{\gamma\beta_k}{8}, \frac{(1{-}\delta)\xi_k}{16}\right\}\mathbb{E}[\varepsilon_k^\pi + \varepsilon_k^\mu + \varepsilon_k^V + \varepsilon_k^J] + B_F\alpha_k\kappa + \mathcal{O}(\frac{\log^2(k+1)}{k+1}) \\
&\leq \mathcal{L}_k - \frac{\rho\alpha_k}{4}\mathbb{E}[\varepsilon_k^\pi + \varepsilon_k^\mu + \varepsilon_k^V + \varepsilon_k^J] + B_F\alpha_k\kappa + \mathcal{O}(\frac{\log^2(k+1)}{k+1}),
\end{aligned} \tag{34}$$

where the last inequality follows from the step size condition $\alpha_k \leq \frac{\gamma}{2\rho}\beta_k$ and $\alpha_k \leq \frac{1-\delta}{4\rho}\xi_k$.

Re-arranging the terms and summing over iterations, we have

$$\begin{aligned}\sum_{t=\tau_k}^{k-1} \alpha_t \mathbb{E}[\varepsilon_t^\pi + \varepsilon_t^\mu + \varepsilon_t^V + \varepsilon_t^J] &\leq \frac{4}{\rho}\sum_{t=\tau_k}^{k-1}(\mathcal{L}_t - \mathcal{L}_{t+1}) + B_F\kappa \sum_{t=\tau_k}^{k-1}\alpha_t + \sum_{t=\tau_k}^{k-1}\mathcal{O}(\frac{\log^2(t+1)}{t+1}) \\ &\leq \frac{4}{\rho}(\mathcal{L}_{\tau_k} + 1) + B_F\kappa\sum_{t=\tau_k}^{k-1}\alpha_t + \mathcal{O}(\log^3(k+1)),\end{aligned}$$

where the second inequality follows from $-\mathcal{L}_{k+1} \leq -\ell(\pi_{\theta_{k+1}}) \leq 1$ and the well-known relationship that

$$\sum_{t=\tau_k}^{k-1}\frac{1}{t+1} \leq \sum_{t=0}^{k-1}\frac{1}{t+1} \leq 2\log(k+1).$$

Due to $\tau_k \leq \mathcal{O}(\log(k+1))$, it is also a standard result that (for example, see Zeng et al. [2024][Lemma 3])

$$\sum_{t=\tau_k}^{k-1}\alpha_t = \sum_{t=\tau_k}^{k-1}\frac{\alpha_0}{\sqrt{t+1}} = \Theta(k+1).$$

Dividing both sides of the inequality by $\sum_{t=\tau_k}^{k-1}\alpha_t$, we get

$$\begin{aligned}\min_{t<k}\mathbb{E}[\varepsilon_t^\pi + \varepsilon_t^\mu + \varepsilon_t^V + \varepsilon_t^J] &\leq \frac{\sum_{t=\tau_k}^{k-1}\alpha_t\mathbb{E}[\varepsilon_t^\pi + \varepsilon_t^\mu + \varepsilon_t^V + \varepsilon_t^J]}{\sum_{t=\tau_k}^{k-1}\alpha_t} \\ &\leq \mathcal{O}(\frac{1}{\sqrt{k+1}})\left(\frac{4}{\rho}(\mathcal{L}_{\tau_k}+1) + \mathcal{O}(\log^3(k+1))\right) + B_F\kappa.\end{aligned}$$

Since the updates of all iterates in Algorithm 1 are bounded, $\mathcal{L}_{\tau_k} \leq \mathcal{O}(\tau_k) \leq \mathcal{O}(\log(k+1))$. As a result, we eventually have

$$\min_{\tau_k \leq t < k}\mathbb{E}[\varepsilon_t^\pi + \varepsilon_t^\mu + \varepsilon_t^V + \varepsilon_t^J] \leq \mathcal{O}\left(\frac{\log^3(k+1)}{\sqrt{k+1}}\right) + \mathcal{O}(\kappa).$$

■

C Proof of Corollaries

C.1 Proof of Corollary 1

As a result of Assumption 4, we have the following gradient domination condition, which is adapted from Lemma 19 of Ganesh et al. [2024].

Lemma 6 *Under Assumption 4, we have the following gradient domination condition for any policy parameter θ and mean field μ*

$$\max_{\bar{\pi}} J(\bar{\pi},\mu) - J(\pi_\theta,\mu) \leq \frac{1}{\sigma}\|\nabla_\theta J(\pi_\theta,\mu)\|.$$

Since $\varepsilon_t^\pi, \varepsilon_t^\mu, \varepsilon_t^V, \varepsilon_t^J$ are all non-negative, we have

$$\begin{aligned}\min_{\tau_k \leq t < k}\mathbb{E}\left[\|\nabla_\theta J(\pi_{\theta_t},\mu)\mid_{\mu=\mu^\star(\pi_{\theta_t})}\|^2\right] &\leq \mathcal{O}\left(\frac{\log^3(k+1)}{\sqrt{k+1}}\right) + \mathcal{O}(\kappa) = \widetilde{\mathcal{O}}\left(\frac{\log^3(k+1)}{\sqrt{k+1}}\right) + \mathcal{O}(\kappa), \\ \min_{\tau_k \leq t < k}\mathbb{E}[\|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\|^2] &\leq \mathcal{O}\left(\frac{\log^3(k+1)}{\sqrt{k+1}}\right) + \mathcal{O}(\kappa) = \widetilde{\mathcal{O}}\left(\frac{\log^3(k+1)}{\sqrt{k+1}}\right) + \mathcal{O}(\kappa).\end{aligned}$$

Applying Lemma 6 with $\theta = \theta_t$ and $\mu = \mu^\star(\pi_{\theta_t})$,

$$\max_\pi J(\pi, \mu^\star(\pi_{\theta_t})) - J(\pi_\theta, \mu^\star(\pi_{\theta_t})) \leq \frac{1}{\sigma}\|\nabla_\theta J(\pi_{\theta_t}, \mu)\mid_{\mu=\mu^\star(\pi_{\theta_t})}\|.$$

By Jensen's inequality,

$$\begin{aligned}
&\left(\min_{\tau_k \leq t < k} \mathbb{E}\left[\max_\pi J(\pi, \mu^\star(\pi_{\theta_t})) - J(\pi_{\theta_t}, \mu^\star(\pi_{\theta_t}))\right]\right)^2 \\
&\leq \min_{\tau_k \leq t < k} \mathbb{E}\left[\left(\max_\pi J(\pi, \mu^\star(\pi_{\theta_t})) - J(\pi_{\theta_t}, \mu^\star(\pi_{\theta_t}))\right)^2\right] \\
&\leq \frac{1}{\sigma^2}\min_{\tau_k \leq t < k} \mathbb{E}\left[\|\nabla_\theta J(\pi_{\theta_t}, \mu)\mid_{\mu=\mu^\star(\pi_{\theta_t})}\|^2\right] \\
&\leq \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{k+1}}\right) + \mathcal{O}(\kappa).
\end{aligned}$$

Taking square root on both sides of this inequality leads to the claimed result on the convergence of the policy.

Similarly, we have

$$\begin{aligned}
\min_{\tau_k \leq t < k} \mathbb{E}[\|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\|] &\leq \sqrt{\min_{\tau_k \leq t < k} \mathbb{E}[\|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\|^2]} \\
&\leq \sqrt{\widetilde{\mathcal{O}}\left(\frac{\log^3(k+1)}{\sqrt{k+1}}\right) + \mathcal{O}(\kappa)} \\
&\leq \widetilde{\mathcal{O}}\left(\frac{1}{(k+1)^{1/4}}\right) + \mathcal{O}(\sqrt{\kappa}).
\end{aligned}$$

■

C.2 Proof of Corollary 2

In the context of single-agent MDP we define

$$\varepsilon_k^\pi = \|\nabla_\theta J_{\text{MDP}}(\pi_{\theta_k})\|^2, \quad \varepsilon_k^V = \|\Pi_{\mathcal{E}_\perp}(\hat{V}_k - V_{MDP}^{\pi_{\theta_k}})\|^2, \quad \varepsilon_k^J = (\hat{J}_k - J_{\text{MDP}}(\pi_{\theta_k}))^2.$$

Theorem 1 implies that

$$\min_{\tau_k \leq t < k} \mathbb{E}[\varepsilon_t^\pi + \varepsilon_t^V + \varepsilon_t^J] \leq \mathcal{O}\left(\frac{\log^3(k+1)}{\sqrt{k+1}}\right) + \mathcal{O}(\kappa).$$

As neither the transition kernel nor the reward function depends on the mean field, (5) trivially holds with $\rho = 1, \kappa = 0$. Therefore, the claimed result holds by recognizing that ε_k^V and ε_k^J are non-negative for any k.

■

D Proof of Propositions

D.1 Proof of Proposition 1

By the L_V-Lipschitz continuity of the function J

$$\begin{aligned}
&J(\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})) - J(\pi_{\theta_{k+1}}, \mu^\star(\pi_{\theta_k})) \\
&\leq -\langle \nabla_\theta J(\pi_{\theta_k}, \mu)\mid_{\mu=\mu^\star(\pi_{\theta_k})}, \theta_{k+1} - \theta_k\rangle + \frac{L_V}{2}\|\theta_{k+1} - \theta_k\|^2 \\
&= -\alpha_k\langle \nabla_\theta J(\pi_{\theta_k}, \mu)\mid_{\mu=\mu^\star(\pi_{\theta_k})}, f_k\rangle + \frac{L_V\alpha_k^2}{2}\|f_k\|^2
\end{aligned}$$

$$
\begin{aligned}
&= -\alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu) |_{\mu=\mu^\star(\pi_{\theta_k})}, \Delta f_k \rangle - \alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu) |_{\mu=\mu^\star(\pi_{\theta_k})}, \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \rangle + \frac{L_V \alpha_k^2}{2} \|f_k\|^2 \\
&= -\alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})}, \Delta f_k \rangle - \alpha_k \| \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 \\
&\qquad + \alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu) |_{\mu=\mu^\star(\pi_{\theta_k})}, \bar{F}(\theta_k, V^{\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})}, \mu^\star(\pi_{\theta_k})) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \rangle + \frac{L_V \alpha_k^2}{2} \|f_k\|^2 \\
&\le -\alpha_k \| \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 - \alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})}, \Delta f_k \rangle \\
&\qquad + \alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})}, \bar{F}(\theta_k, V^{\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})}, \mu^\star(\pi_{\theta_k})) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \rangle + \frac{L_V B_F^2 \alpha_k^2}{2}, \qquad (35)
\end{aligned}
$$

where the third equation follows from $\nabla_\theta J(\pi_\theta, \mu) \mid_{\mu=\mu^\star(\pi_\theta)} = \bar{F}(\theta, V^{\pi_\theta, \mu^\star(\pi_\theta)}, \mu^\star(\pi_\theta))$ for any θ.

To bound the second term on the right hand side of (35), we use the fact that $\langle \vec{a}, \vec{b} \rangle \le \frac{c}{2} \|\vec{a}\|^2 + \frac{1}{2c} \|\vec{b}\|^2$ for any vectors $\vec{a}, \vec{b}$ and scalar $c > 0$

$$
-\alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})}, \Delta f_k \rangle \le \frac{\alpha_k}{4} \| \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 + \alpha_k \|\Delta f_k\|^2. \qquad (36)
$$

Similarly, for the third term of (35), we have

$$
\begin{aligned}
&\alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})}, \mu^\star(\pi_{\theta_k})) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \rangle \\
&\le \frac{\alpha_k}{4} \| \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 + \alpha_k \| \bar{F}(\theta_k, V^{\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})}, \mu^\star(\pi_{\theta_k})) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \|^2 \\
&\le \frac{\alpha_k}{4} \| \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 + L_F^2 \alpha_k \| \Pi_{\mathcal{E}_\perp} (\hat{V}_k - V^{\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})}) \|^2 + L_F^2 \alpha_k \| \hat{\mu}_k - \mu^\star(\pi_{\theta_k}) \|^2 \\
&= \frac{\alpha_k}{4} \| \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 + L_F^2 \alpha_k (\varepsilon_k^V + \varepsilon_k^\mu). \qquad (37)
\end{aligned}
$$

Plugging (36)-(37) into (35), we have

$$
\begin{aligned}
&J(\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})) - J(\pi_{\theta_{k+1}}, \mu^\star(\pi_{\theta_k})) \\
&\le -\alpha_k \| \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 - \alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})}, \Delta f_k \rangle \\
&\qquad + \alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})}, \bar{F}(\theta_k, V^{\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})}, \mu^\star(\pi_{\theta_k})) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \rangle + \frac{L_V B_F^2 \alpha_k^2}{2} \\
&\le -\alpha_k \| \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 + \frac{\alpha_k}{4} \| \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 + \alpha_k \|\Delta f_k\|^2 \\
&\qquad + \frac{\alpha_k}{4} \| \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 + L_F^2 \alpha_k (\varepsilon_k^V + \varepsilon_k^\mu) + \frac{L_V B_F^2 \alpha_k^2}{2} \\
&\le -\frac{\alpha_k}{2} \| \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 + \alpha_k \|\Delta f_k\|^2 + L_F^2 \alpha_k (\varepsilon_k^V + \varepsilon_k^\mu) + \frac{L_V B_F^2 \alpha_k^2}{2}. \qquad (38)
\end{aligned}
$$

By (5), we have

$$
\begin{aligned}
&J(\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})) - J(\pi_{\theta_{k+1}}, \mu^\star(\pi_{\theta_{k+1}})) \\
&\le \rho \Big(-\frac{\alpha_k}{2} \| \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 + \alpha_k \|\Delta f_k\|^2 + L_F^2 \alpha_k (\varepsilon_k^V + \varepsilon_k^\mu) + \frac{L_V B_F^2 \alpha_k^2}{2} \Big) + B_F \alpha_k \kappa \\
&\le -\frac{\rho \alpha_k}{2} \| \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 + \rho \alpha_k \|\Delta f_k\|^2 + \rho L_F^2 \alpha_k (\varepsilon_k^V + \varepsilon_k^\mu) + \frac{\rho L_V B_F^2 \alpha_k^2}{2} + B_F \alpha_k \kappa.
\end{aligned}
$$

■

D.2 Proof of Proposition 2

The proof of Proposition 2 relies on the lemma below. We defer the proof of the lemma to Sec.E.7.

Lemma 7 *We have for all $k \ge \tau_k$*

$$
\mathbb{E}[\langle \Delta f_k, F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \rangle] \le (20L + 2|\mathcal{A}|) L_F L_{TV} B_F^3 B_G B_H^2 \tau_k^2 \lambda_{k-\tau_k}.
$$

By the update rule of f_k,

$$\begin{aligned}
\Delta f_{k+1} &= f_{k+1} - \bar{F}(\theta_{k+1}, \hat{V}_{k+1}, \hat{\mu}_{k+1}) \\
&= (1-\lambda_k) f_k + \lambda_k F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{F}(\theta_{k+1}, \hat{V}_{k+1}, \hat{\mu}_{k+1}) \\
&= (1-\lambda_k) f_k + \lambda_k \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) - \bar{F}(\theta_{k+1}, \hat{V}_{k+1}, \hat{\mu}_{k+1}) \\
&\qquad + \lambda_k \Big(F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \Big) \\
&= (1-\lambda_k) \Delta f_k + \Big(\bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) - \bar{F}(\theta_{k+1}, \hat{V}_{k+1}, \hat{\mu}_{k+1}) \Big) \\
&\qquad + \lambda_k \Big(F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \Big).
\end{aligned}$$

Taking the norm, we have

$$\begin{aligned}
&\|\Delta f_{k+1}\|^2 \\
&= (1-\lambda_k)^2 \|\Delta f_k\|^2 + \|\bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) - \bar{F}(\theta_{k+1}, \hat{V}_{k+1}, \hat{\mu}_{k+1})\|^2 \\
&\quad + \lambda_k^2 \|F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k)\|^2 \\
&\quad + (1-\lambda_k) \langle \Delta f_k, \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) - \bar{F}(\theta_{k+1}, \hat{V}_{k+1}, \hat{\mu}_{k+1}) \rangle \\
&\quad + (1-\lambda_k)\lambda_k \langle \Delta f_k, F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \rangle \\
&\quad + \lambda_k \langle \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) - \bar{F}(\theta_{k+1}, \hat{V}_{k+1}, \hat{\mu}_{k+1}), F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \rangle \\
&\leq (1-\lambda_k)^2 \|\Delta f_k\|^2 + 2\|\bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) - \bar{F}(\theta_{k+1}, \hat{V}_{k+1}, \hat{\mu}_{k+1})\|^2 \\
&\quad + 2\lambda_k^2 \|F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k)\|^2 \\
&\quad + \frac{\lambda_k}{2} \|\Delta f_k\|^2 + \frac{2}{\lambda_k} \|\bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) - \bar{F}(\theta_{k+1}, \hat{V}_{k+1}, \hat{\mu}_{k+1})\|^2 \\
&\quad + (1-\lambda_k)\lambda_k \langle \Delta f_k, F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \rangle \\
&\leq (1-\lambda_k) \|\Delta f_k\|^2 + (-\frac{\lambda_k}{2} + \lambda_k^2) \|\Delta f_k\|^2 + \frac{4}{\lambda_k} \|\bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) - \bar{F}(\theta_{k+1}, \hat{V}_{k+1}, \hat{\mu}_{k+1})\|^2 \\
&\quad + 8B_F^2 \lambda_k^2 + (1-\lambda_k)\lambda_k \langle \Delta f_k, F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \rangle, \qquad (39)
\end{aligned}$$

where the final inequality follows from the step size condition $\lambda_k \leq 1$ and the boundedness of operator F which implies

$$\|\bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) - \bar{F}(\theta_{k+1}, \hat{V}_{k+1}, \hat{\mu}_{k+1})\| \leq 2B_F.$$

Taking the expectation, we can simplify (39) as

$$\begin{aligned}
&\mathbb{E}[\|\Delta f_{k+1}\|^2] \\
&\leq \mathbb{E}\Big[(1-\lambda_k) \|\Delta f_k\|^2 + (-\frac{\lambda_k}{2} + \lambda_k^2) \|\Delta f_k\|^2 + \frac{4}{\lambda_k} \|\bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) - \bar{F}(\theta_{k+1}, \hat{V}_{k+1}, \hat{\mu}_{k+1})\|^2 \\
&\quad + 8B_F^2 \lambda_k^2 + (1-\lambda_k)\lambda_k \langle \Delta f_k, F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \rangle \Big] \\
&\leq (1-\lambda_k) \mathbb{E}[\|\Delta f_k\|^2] + (-\frac{\lambda_k}{2} + \lambda_k^2) \mathbb{E}[\|\Delta f_k\|^2] + 8B_F^2 \lambda_k^2 \\
&\quad + \frac{4L_F^2}{\lambda_k} \mathbb{E}[\Big(\|\theta_k - \theta_{k+1}\| + \|\hat{V}_k - \hat{V}_{k+1}\| + \|\hat{\mu}_k - \hat{\mu}_{k+1}\| \Big)^2] \\
&\quad + (1-\lambda_k)\lambda_k \cdot (20L + 2|\mathcal{A}|) L_F L_{TV} B_F^3 B_G B_H^2 \tau_k^2 \lambda_{k-\tau_k} \\
&\leq (1-\lambda_k) \mathbb{E}[\|\Delta f_k\|^2] + (-\frac{\lambda_k}{2} + \lambda_k^2) \mathbb{E}[\|\Delta f_k\|^2] + (28L{+}2|\mathcal{A}|) L_F L_{TV} B_F^3 B_G B_H^2 \tau_k^2 \lambda_k \lambda_{k-\tau_k} \\
&\quad + \frac{4L_F^2}{\lambda_k} \mathbb{E}[(\alpha_k \|f_k\| + \beta_k \|g_k\| + \xi_k \|h_k\|)^2] \\
&\leq (1-\lambda_k) \mathbb{E}[\|\Delta f_k\|^2] + (-\frac{\lambda_k}{2} + \lambda_k^2) \mathbb{E}[\|\Delta f_k\|^2] + (28L{+}2|\mathcal{A}|) L_F L_{TV} B_F^3 B_G B_H^2 \tau_k^2 \lambda_k \lambda_{k-\tau_k}
\end{aligned}$$

$$\begin{aligned}
&+\mathbb{E}\Big[\frac{12L_F^2\alpha_k}{\lambda_k}\left(\|\Delta f_k\|+L_F\sqrt{\varepsilon_k^V}+L_F(L_V+1)\sqrt{\varepsilon_k^\mu}+\sqrt{\varepsilon_k^\pi}\right)^2\\
&+\frac{12L_F^2\beta_k}{\lambda_k}\left(\|\Delta g_k\|+L_G\sqrt{\varepsilon_k^V}+L_G\sqrt{\varepsilon_k^J}\right)^2+\frac{12L_F^2\xi_k}{\lambda_k}\left(L_H\|\Delta h_k\|+\sqrt{\varepsilon_k^\mu}\right)^2\Big],
\end{aligned}\tag{40}$$

where the second inequality plugs in the result of Lemma 7 and bounds $\|\bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k)-\bar{F}(\theta_{k+1},\hat{V}_{k+1},\hat{\mu}_{k+1})\|^2$ using the Lipschitz condition established in Lemma 3.

The sum of the last three terms can be bounded as

$$\begin{aligned}
&\frac{12L_F^2\alpha_k}{\lambda_k}\left(\|\Delta f_k\|+L_F\sqrt{\varepsilon_k^V}+L_F(L_V+1)\sqrt{\varepsilon_k^\mu}+\sqrt{\varepsilon_k^\pi}\right)^2\\
&\quad+\frac{12L_F^2\beta_k}{\lambda_k}\left(\|\Delta g_k\|+L_G\sqrt{\varepsilon_k^V}+L_G\sqrt{\varepsilon_k^J}\right)^2+\frac{12L_F^2\xi_k}{\lambda_k}\left(L_H\|\Delta h_k\|+\sqrt{\varepsilon_k^\mu}\right)^2\\
&\leq\frac{48L_F^2\alpha_k^2}{\lambda_k}\|\Delta f_k\|^2+\frac{48L_F^4\alpha_k^2}{\lambda_k}\varepsilon_k^V+\frac{192L_F^4L_V^2\alpha_k^2}{\lambda_k}\varepsilon_k^\mu+\frac{48L_F^2\alpha_k^2}{\lambda_k}\varepsilon_k^\pi\\
&\quad+\frac{36L_F^2\beta_k^2}{\lambda_k}\|\Delta g_k\|^2+\frac{48L_F^2L_G^2\beta_k^2}{\lambda_k}\varepsilon_k^V+\frac{48L_F^2L_G^2\beta_k^2}{\lambda_k}\varepsilon_k^J\\
&\quad+\frac{24L_F^2L_H^2\xi_k^2}{\lambda_k}\|\Delta h_k\|^2+\frac{24L_F^2\xi_k^2}{\lambda_k}\varepsilon_k^\mu\\
&\leq\frac{48L_F^2\alpha_k^2}{\lambda_k}\|\Delta f_k\|^2+\frac{36L_F^2\beta_k^2}{\lambda_k}\|\Delta g_k\|^2+\frac{24L_F^2L_H^2\xi_k^2}{\lambda_k}\|\Delta h_k\|^2+\frac{48L_F^2\alpha_k^2}{\lambda_k}\varepsilon_k^\pi\\
&\quad+\frac{216L_F^4L_V^2\xi_k^2}{\lambda_k}\varepsilon_k^\mu+\frac{96L_F^4L_G^2\beta_k^2}{\lambda_k}\varepsilon_k^V+\frac{48L_F^2L_G^2\beta_k^2}{\lambda_k}\varepsilon_k^J.
\end{aligned}\tag{41}$$

Combining (40) and (41), we get

$$\begin{aligned}
&\mathbb{E}[\|\Delta f_{k+1}\|^2]\\
&\leq(1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2]+(-\frac{\lambda_k}{2}+\lambda_k^2)\mathbb{E}[\|\Delta f_k\|^2]+(28L+2|\mathcal{A}|)L_FL_{TV}B_F^3B_GB_H^2\tau_k^2\lambda_k\lambda_{k-\tau_k}\\
&\quad+\mathbb{E}\Big[\frac{48L_F^2\alpha_k^2}{\lambda_k}\|\Delta f_k\|^2+\frac{36L_F^2\beta_k^2}{\lambda_k}\|\Delta g_k\|^2+\frac{24L_F^2L_H^2\xi_k^2}{\lambda_k}\|\Delta h_k\|^2+\frac{48L_F^2\alpha_k^2}{\lambda_k}\varepsilon_k^\pi\\
&\quad+\frac{216L_F^4L_V^2\xi_k^2}{\lambda_k}\varepsilon_k^\mu+\frac{96L_F^4L_G^2\beta_k^2}{\lambda_k}\varepsilon_k^V+\frac{48L_F^2L_G^2\beta_k^2}{\lambda_k}\varepsilon_k^J\Big]\\
&=(1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2]+(-\frac{\lambda_k}{2}+\lambda_k^2+\frac{48L_F^2\alpha_k^2}{\lambda_k})\mathbb{E}[\|\Delta f_k\|^2]\\
&\quad+\frac{36L_F^2\beta_k^2}{\lambda_k}\mathbb{E}[\|\Delta g_k\|^2]+\frac{24L_F^2L_H^2\xi_k^2}{\lambda_k}\mathbb{E}[\|\Delta h_k\|^2]+\frac{48L_F^2\alpha_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^\pi]+\frac{216L_F^4L_V^2\xi_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^\mu]\\
&\quad+\frac{96L_F^4L_G^2\beta_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^V]+\frac{48L_F^2L_G^2\beta_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^J]+(28L+2|\mathcal{A}|)L_FL_{TV}B_F^3B_GB_H^2\tau_k^2\lambda_k\lambda_{k-\tau_k}.
\end{aligned}$$

■

D.3 Proof of Proposition 3

We first introduce the following lemma, which will be used in the proof of Proposition 3. The proof of Lemma 8 is presented in Sec.E.8.

Lemma 8 *Under Assumption 2, we have for any policy parameter θ and mean field μ*

$$\langle\mu-\mu^\star(\pi_\theta),\bar{H}(\theta,\mu)-\bar{H}(\theta,\mu^\star(\pi_\theta))\rangle\leq-(1-\delta)\|\mu-\mu^\star(\pi_\theta)\|^2.$$

By the definition of ε_k^μ,

$$\varepsilon_{k+1}^\mu=\|\hat{\mu}_{k+1}-\mu^\star(\pi_{\theta_{k+1}})\|^2$$

$$
\begin{aligned}
&= \|\Pi_{\Delta_{\mathcal{S}}}(\hat{\mu}_k + \xi_k h_k) - \mu^\star(\pi_{\theta_{k+1}})\|^2 \\
&\leq \|\hat{\mu}_k + \xi_k h_k - \mu^\star(\pi_{\theta_{k+1}})\|^2 \\
&= \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k \Delta h_k + \xi_k \bar{H}(\theta_k, \hat{\mu}_k) - \left(\mu^\star(\pi_{\theta_{k+1}}) - \mu^\star(\pi_{\theta_k})\right)\|^2 \\
&= \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k \bar{H}(\theta_k, \hat{\mu}_k)\|^2 + \xi_k^2 \|\Delta h_k\|^2 + \|\mu^\star(\pi_{\theta_{k+1}}) - \mu^\star(\pi_{\theta_k})\|^2 \\
&\quad + 2\xi_k \langle \hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k \bar{H}(\theta_k, \hat{\mu}_k), \Delta h_k \rangle + 2\langle \hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k \bar{H}(\theta_k, \hat{\mu}_k), \mu^\star(\pi_{\theta_{k+1}}) - \mu^\star(\pi_{\theta_k}) \rangle \\
&\quad + 2\xi_k \langle \Delta h_k, \mu^\star(\pi_{\theta_{k+1}}) - \mu^\star(\pi_{\theta_k}) \rangle, \qquad (42)
\end{aligned}
$$

where the first inequality is due to the fact that projection to a convex set is a non-expansive operator.

To bound the first term of (42),

$$
\begin{aligned}
&\|\hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k \bar{H}(\theta_k, \hat{\mu}_k)\|^2 \\
&= \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k \left(\bar{H}(\theta_k, \hat{\mu}_k) - \bar{H}(\theta_k, \mu^\star(\pi_{\theta_k}))\right)\|^2 \\
&= \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\|^2 + \xi_k^2 \|\bar{H}(\theta_k, \hat{\mu}_k) - \bar{H}(\theta_k, \mu^\star(\pi_{\theta_k}))\|^2 \\
&\quad + 2\xi_k \langle \hat{\mu}_k - \mu^\star(\pi_{\theta_k}), \bar{H}(\theta_k, \hat{\mu}_k) - \bar{H}(\theta_k, \mu^\star(\pi_{\theta_k})) \rangle \\
&\leq \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\|^2 + L_H^2 \xi_k^2 \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\|^2 - (1-\delta)\xi_k \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\|^2 \\
&\leq (1 - \frac{(1-\delta)\xi_k}{2})\varepsilon_k^\mu, \qquad (43)
\end{aligned}
$$

where the first equation uses $\bar{H}(\theta, \mu^\star(\pi_\theta)) = 0$ for any θ, the first inequality is a result of Lemma 8 and the Lipschitz continuity of $\bar{H}$, and the second inequality follows from the step size condition $\xi_k \leq \beta_k \leq \frac{1-\delta}{2L_H^2}$.

We next treat the second and third term of (42) using the fact that $\|h_k\| \leq B_H$, $\|\bar{H}(\theta_k, \hat{\mu}_k)\| \leq B_H$, $\|f_k\| \leq B_F$ and that the operator $\mu^\star$ is Lipschitz

$$
\begin{aligned}
\xi_k^2 \|\Delta h_k\|^2 + \|\mu^\star(\pi_{\theta_{k+1}}) - \mu^\star(\pi_{\theta_k})\|^2 &\leq 2\xi_k^2 \|h_k\|^2 + 2\xi_k^2 \|\bar{H}(\theta_k, \hat{\mu}_k)\|^2 + L\|\pi_{\theta_{k+1}} - \pi_{\theta_k}\|^2 \\
&\leq 4B_H^2 \xi_k^2 + L^2 \|f_k\|^2 \\
&\leq 4B_H^2 \xi_k^2 + L^2 B_F^2 \alpha_k^2. \qquad (44)
\end{aligned}
$$

The fourth term of (42) can be bounded leveraging the result in (43) as follows

$$
\begin{aligned}
&2\xi_k \langle \hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k \bar{H}(\theta_k, \hat{\mu}_k), \Delta h_k \rangle \\
&\leq \frac{(1-\delta)\xi_k}{8} \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k \bar{H}(\theta_k, \hat{\mu}_k)\|^2 + \frac{8\xi_k}{1-\delta} \|\Delta h_k\|^2 \\
&\leq \frac{(1-\delta)\xi_k}{8} \cdot (1 - \frac{(1-\delta)\xi_k}{2})\varepsilon_k^\mu + \frac{8\xi_k}{1-\delta} \|\Delta h_k\|^2 \\
&\leq \frac{(1-\delta)\xi_k}{8} \varepsilon_k^\mu + \frac{8\xi_k}{1-\delta} \|\Delta h_k\|^2. \qquad (45)
\end{aligned}
$$

Similarly, for the fifth term of (42), we have

$$
\begin{aligned}
&2\langle \hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k \bar{H}(\theta_k, \hat{\mu}_k), \mu^\star(\pi_{\theta_{k+1}}) - \mu^\star(\pi_{\theta_k}) \rangle \\
&\leq \frac{(1-\delta)\xi_k}{8} \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k \bar{H}(\theta_k, \hat{\mu}_k)\|^2 + \frac{8}{(1-\delta)\xi_k} \|\mu^\star(\pi_{\theta_{k+1}}) - \mu^\star(\pi_{\theta_k})\|^2 \\
&\leq \frac{(1-\delta)\xi_k}{8} \varepsilon_k^\mu + \frac{8L^2}{(1-\delta)\xi_k} \|\pi_{\theta_{k+1}} - \pi_{\theta_k}\|^2 \\
&\leq \frac{(1-\delta)\xi_k}{8} \varepsilon_k^\mu + \frac{8L^2 \alpha_k^2}{(1-\delta)\xi_k} \|f_k\|^2 \\
&\leq \frac{(1-\delta)\xi_k}{8} \varepsilon_k^\mu + \frac{8L^2 \alpha_k^2}{(1-\delta)\xi_k} \left(\|\Delta f_k\| + L_F \sqrt{\varepsilon_k^V} + L_F(L_V + 1)\sqrt{\varepsilon_k^\mu} + \sqrt{\varepsilon_k^\pi}\right)^2 \\
&\leq \frac{(1-\delta)\xi_k}{8} \varepsilon_k^\mu + \frac{32L^2 \alpha_k^2}{(1-\delta)\xi_k} \left(\|\Delta f_k\|^2 + L_F^2 \varepsilon_k^V + 4L_F^2 L_V^2 \varepsilon_k^\mu + \varepsilon_k^\pi\right), \qquad (46)
\end{aligned}
$$

where the fourth inequality follows from Lemma 4.

The final term of (42) can be bounded simply with the Cauchy-Schwarz inequality

$$\begin{aligned}2\xi_k\langle\Delta h_k,\mu^\star(\pi_{\theta_{k+1}})-\mu^\star(\pi_{\theta_k})\rangle &\leq 2\xi_k\|\Delta h_k\|\|\mu^\star(\pi_{\theta_{k+1}})-\mu^\star(\pi_{\theta_k})\|\\ &\leq 4B_H\xi_k\cdot L\|\pi_{\theta_{k+1}}-\pi_{\theta_k}\|\\ &\leq 4LB_FB_H\alpha_k\xi_k.\end{aligned}\tag{47}$$

Plugging (43)-(47) into (42), we get

$$\begin{aligned}\varepsilon_{k+1}^\mu &\leq (1-\frac{(1-\delta)\xi_k}{2})\varepsilon_k^\mu+4B_H^2\xi_k^2+L^2B_F^2\alpha_k^2+\frac{(1-\delta)\xi_k}{8}\varepsilon_k^\mu+\frac{8\xi_k}{1-\delta}\|\Delta h_k\|^2\\ &\quad+\frac{(1-\delta)\xi_k}{8}\varepsilon_k^\mu+\frac{32L^2\alpha_k^2}{(1-\delta)\xi_k}\left(\|\Delta f_k\|^2+L_F^2\varepsilon_k^V+4L_F^2L_V^2\varepsilon_k^\mu+\varepsilon_k^\pi\right)+4LB_FB_H\alpha_k\xi_k\\ &\leq (1-\frac{(1-\delta)\xi_k}{8})\varepsilon_k^\mu+\frac{8\xi_k}{1-\delta}\|\Delta h_k\|^2+4B_H^2\xi_k^2+L^2B_F^2\alpha_k^2+\frac{32L^2\alpha_k^2}{(1-\delta)\xi_k}\left(\|\Delta f_k\|^2+L_F^2\varepsilon_k^V+\varepsilon_k^\pi\right)\\ &\quad+4LB_FB_H\alpha_k\xi_k+(-\frac{(1-\delta)\xi_k}{8}+\frac{128L^2L_F^2L_V^2\alpha_k^2}{(1-\delta)\xi_k})\varepsilon_k^\mu\\ &\leq (1-\frac{(1-\delta)\xi_k}{8})\varepsilon_k^\mu+\frac{8\xi_k}{1-\delta}\|\Delta h_k\|^2+\frac{32L^2\alpha_k^2}{(1-\delta)\xi_k}\left(\|\Delta f_k\|^2+L_F^2\varepsilon_k^V+\varepsilon_k^\pi\right)+9L^2B_F^2B_H^2\xi_k^2,\end{aligned}$$

where the last inequality is a result of the step size condition $\alpha_k\leq\xi_k$ and $\alpha_k\leq\frac{1-\delta}{16LL_FL_V}\xi_k$.

■

D.4 Proof of Proposition 4

The proof of Proposition 4 uses an intermediate result established in the lemma below. We defer the proof of the lemma to Sec.E.9.

Lemma 9 *We have for all* $k\geq\tau_k$

$$\mathbb{E}[\langle\Delta h_k,e_{s_k}-\mathbb{E}_{s\sim\nu^{\pi_{\theta_k},\hat{\mu}_k}}[e_s]\rangle]\leq 16LB_FB_H^2\tau_k^2\lambda_{k-\tau_k}.$$

By the update rule of h_k,

$$\begin{aligned}\Delta h_{k+1}&=h_{k+1}-\bar{H}(\theta_{k+1},\hat{\mu}_{k+1})\\ &=(1-\lambda_k)h_k+\lambda_k(e_{s_k}-\hat{\mu}_k)-\bar{H}(\theta_{k+1},\hat{\mu}_{k+1})\\ &=(1-\lambda_k)h_k+\lambda_k\bar{H}(\theta_k,\hat{\mu}_k)-\bar{H}(\theta_{k+1},\hat{\mu}_{k+1})+\lambda_k\Big((e_{s_k}-\hat{\mu}_k)-\bar{H}(\theta_k,\hat{\mu}_k)\Big)\\ &=(1-\lambda_k)\Delta h_k+\Big(\bar{H}(\theta_k,\hat{\mu}_k)-\bar{H}(\theta_{k+1},\hat{\mu}_{k+1})\Big)+\lambda_k\Big((e_{s_k}-\hat{\mu}_k)-\bar{H}(\theta_k,\hat{\mu}_k)\Big).\end{aligned}$$

This implies

$$\begin{aligned}&\|\Delta h_{k+1}\|^2\\ &=(1-\lambda_k)^2\|\Delta h_k\|^2+\|\bar{H}(\theta_k,\hat{\mu}_k)-\bar{H}(\theta_{k+1},\hat{\mu}_{k+1})\|^2+\lambda_k^2\|(e_{s_k}-\hat{\mu}_k)-\bar{H}(\theta_k,\hat{\mu}_k)\|^2\\ &\quad+(1-\lambda_k)\langle\Delta h_k,\bar{H}(\theta_k,\hat{\mu}_k)-\bar{H}(\theta_{k+1},\hat{\mu}_{k+1})\rangle\\ &\quad+(1-\lambda_k)\lambda_k\langle\Delta h_k,(e_{s_k}-\hat{\mu}_k)-\bar{H}(\theta_k,\hat{\mu}_k)\rangle\\ &\quad+\lambda_k\langle\bar{H}(\theta_k,\hat{\mu}_k)-\bar{H}(\theta_{k+1},\hat{\mu}_{k+1}),(e_{s_k}-\hat{\mu}_k)-\bar{H}(\theta_k,\hat{\mu}_k)\rangle\\ &\leq(1-\lambda_k)^2\|\Delta h_k\|^2+2\|\bar{H}(\theta_k,\hat{\mu}_k)-\bar{H}(\theta_{k+1},\hat{\mu}_{k+1})\|^2+2\lambda_k^2\|(e_{s_k}-\hat{\mu}_k)-\bar{H}(\theta_k,\hat{\mu}_k)\|^2\\ &\quad+\frac{\lambda_k}{2}\|\Delta h_k\|^2+\frac{2}{\lambda_k}\|\bar{H}(\theta_k,\hat{\mu}_k)-\bar{H}(\theta_{k+1},\hat{\mu}_{k+1})\|^2\\ &\quad+(1-\lambda_k)\lambda_k\langle\Delta h_k,e_{s_k}-\mathbb{E}_{s\sim\nu^{\pi_{\theta_k},\hat{\mu}_k}}[e_s]\rangle\end{aligned}$$

$$\begin{aligned}
&\leq (1-\lambda_k)\|\Delta h_k\|^2 + (-\frac{\lambda_k}{2}+\lambda_k^2)\|\Delta h_k\|^2 + \frac{4}{\lambda_k}\|\bar{H}(\theta_k,\hat{\mu}_k)-\bar{H}(\theta_{k+1},\hat{\mu}_{k+1})\|^2\\
&\qquad + (1-\lambda_k)\lambda_k\langle \Delta h_k, e_{s_k} - \mathbb{E}_{s\sim\nu^{\pi_{\theta_k},\hat{\mu}_k}}[e_s]\rangle + 8B_H\lambda_k^2,
\end{aligned}$$

where the final inequality follows from the step size choice $\lambda_k \leq 1$. Taking the expectation and applying Lemma 9 and the Lipschitz continuity of operator $\bar{H}$, we further have

$$\begin{aligned}
&\mathbb{E}[\|\Delta h_{k+1}\|^2]\\
&\leq (1-\lambda_k)\mathbb{E}[\|\Delta h_k\|^2] + (-\frac{\lambda_k}{2}+\lambda_k^2)\mathbb{E}[\|\Delta h_k\|^2] + \frac{4}{\lambda_k}\mathbb{E}[(L_H\|\theta_k-\theta_{k+1}\| + L_H\|\hat{\mu}_k-\hat{\mu}_{k+1}\|)^2]\\
&\qquad + (1-\lambda_k)\lambda_k\cdot 16LB_F B_H^2\tau_k^2\lambda_{k-\tau_k} + 8B_H\lambda_k^2\\
&\leq (1-\lambda_k)\mathbb{E}[\|\Delta h_k\|^2] + (-\frac{\lambda_k}{2}+\lambda_k^2)\mathbb{E}[\|\Delta h_k\|^2] + \frac{8L_H^2}{\lambda_k}\mathbb{E}[\alpha_k^2\|f_k\|^2 + \xi_k^2\|h_k\|^2]\\
&\qquad + 16LB_F B_H^2\tau_k^2\lambda_k\lambda_{k-\tau_k} + 8B_H\lambda_k^2\\
&\leq (1-\lambda_k)\mathbb{E}[\|\Delta h_k\|^2] + (-\frac{\lambda_k}{2}+\lambda_k^2)\mathbb{E}[\|\Delta h_k\|^2]\\
&\qquad + \frac{8L_H^2\alpha_k^2}{\lambda_k}\mathbb{E}[\big(\|\Delta f_k\| + L_F\sqrt{\varepsilon_k^V} + L_F(L_V+1)\sqrt{\varepsilon_k^\mu} + \sqrt{\varepsilon_k^\pi}\big)^2]\\
&\qquad + \frac{8L_H^2\xi_k^2}{\lambda_k}\mathbb{E}[\big(\|\Delta h_k\| + L_H\sqrt{\epsilon_k^\mu}\big)^2] + 16LB_F B_H^2\tau_k^2\lambda_k\lambda_{k-\tau_k} + 8B_H\lambda_k^2\\
&\leq (1-\lambda_k)\mathbb{E}[\|\Delta h_k\|^2] + (-\frac{\lambda_k}{2}+\lambda_k^2)\mathbb{E}[\|\Delta h_k\|^2]\\
&\qquad + \frac{32L_H^2\alpha_k^2}{\lambda_k}\mathbb{E}[\|\Delta f_k\|^2 + L_F^2\varepsilon_k^V + 4L_F^2L_V^2\varepsilon_k^\mu + \varepsilon_k^\pi]\\
&\qquad + \frac{16L_H^2\xi_k^2}{\lambda_k}\mathbb{E}[\|\Delta h_k\|^2 + L_H^2\epsilon_k^\mu] + 24LB_F B_H^2\tau_k^2\lambda_k\lambda_{k-\tau_k}\\
&\leq (1-\lambda_k)\mathbb{E}[\|\Delta h_k\|^2] + (-\frac{\lambda_k}{2}+\lambda_k^2+\frac{16L_H^2\xi_k^2}{\lambda_k})\mathbb{E}[\|\Delta h_k\|^2] + \frac{32L_H^2\alpha_k^2}{\lambda_k}\mathbb{E}[\|\Delta f_k\|^2]\\
&\qquad + \frac{32L_H^2L_F^2\alpha_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^V] + \frac{144L_F^2L_V^2L_H^4\xi_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^\mu] + \frac{32L_H^2\alpha_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^\pi] + 24LB_F B_H^2\tau_k^2\lambda_k\lambda_{k-\tau_k},
\end{aligned}$$

where the third inequality bounds $\|f_k\|$ and $\|h_k\|$ with Lemma 4. The step size condition $\alpha_k \leq \xi_k$ is used a few times to simplify and combine terms.

■

D.5 Proof of Proposition 5

We use the following lemma in our analysis. The proof of the lemma is deferred to Sec.E.10.

Lemma 10 *Under Assumption 1, it holds for any θ, μ, and V that*

$$\left\langle \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(V-V^{\pi_\theta,\mu}) \\ J-J(\pi_\theta,\mu)\end{bmatrix}, \begin{bmatrix} \Pi_{\mathcal{E}_\perp}\bar{G}^V(\theta,V,J,\mu) \\ \bar{G}^J(\theta,J,\mu)\end{bmatrix}\right\rangle \leq -\frac{\gamma}{2}(\|\Pi_{\mathcal{E}_\perp}(V-V^{\pi_\theta,\mu})\|^2 + (J-J(\pi_\theta,\mu))^2),$$

where $\gamma\in(0,1)$ is the discount factor in Lemma 5.

By the definition of ε_k^V,

$$\begin{aligned}
&\varepsilon_{k+1}^V + \varepsilon_{k+1}^J\\
&= \left\|\begin{bmatrix} \Pi_{\mathcal{E}_\perp}(\hat{V}_{k+1} - V^{\pi_{\theta_{k+1}},\hat{\mu}_{k+1}}) \\ \hat{J}_{k+1} - J(\pi_{\theta_{k+1}},\hat{\mu}_{k+1})\end{bmatrix}\right\|^2\\
&= \left\|\begin{bmatrix} \Pi_{\mathcal{E}_\perp}(\Pi_{B_V}(\hat{V}_k + \beta_k g_k^V) - V^{\pi_{\theta_{k+1}},\hat{\mu}_{k+1}}) \\ \Pi_{[0,1]}(\hat{J}_k + \beta_k g_k^J) - J(\pi_{\theta_{k+1}},\hat{\mu}_{k+1})\end{bmatrix}\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&\leq \left\| \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(\hat{V}_k + \beta_k g_k^V - V^{\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}}) \\ \hat{J}_k + \beta_k g_k^J - J(\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}) \end{bmatrix} \right\|^2 \\
&= \left\| \begin{bmatrix} \Pi_{\mathcal{E}_\perp}\Big(\hat{V}_k - V^{\pi_{\theta_k}, \hat{\mu}_k} + \beta_k \bar{G}^V(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) + \beta_k \Delta g_k^V - \Big(V^{\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}} - V^{\pi_{\theta_k}, \hat{\mu}_k}\Big)\Big) \\ \hat{J}_k - J(\pi_{\theta_k}, \hat{\mu}_k) + \beta_k \bar{G}^J(\theta_k, \hat{J}_k, \hat{\mu}_k) + \beta_k \Delta g_k^J - (J(\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}) - J(\pi_{\theta_k}, \hat{\mu}_k)) \end{bmatrix} \right\|^2 \\
&\leq \left\| \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k}, \hat{\mu}_k}) \\ \hat{J}_k - J(\pi_{\theta_k}, \hat{\mu}_k) \end{bmatrix} + \beta_k \begin{bmatrix} \Pi_{\mathcal{E}_\perp} \bar{G}^V(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) \\ \bar{G}^J(\theta_k, \hat{J}_k, \hat{\mu}_k) \end{bmatrix} \right\|^2 + \beta_k^2 \|\Delta g_k\|^2 \\
&\quad + \left\| \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}} - V^{\pi_{\theta_k}, \hat{\mu}_k}) \\ J(\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}) - J(\pi_{\theta_k}, \hat{\mu}_k) \end{bmatrix} \right\|^2 \\
&\quad + 2\beta_k \left\langle \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k}, \hat{\mu}_k}) \\ \hat{J}_k - J(\pi_{\theta_k}, \hat{\mu}_k) \end{bmatrix} + \beta_k \begin{bmatrix} \Pi_{\mathcal{E}_\perp} \bar{G}^V(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) \\ \bar{G}^J(\theta_k, \hat{J}_k, \hat{\mu}_k) \end{bmatrix}, \begin{bmatrix} \Pi_{\mathcal{E}_\perp} \Delta g_k^V \\ \Delta g_k^J \end{bmatrix} \right\rangle \\
&\quad + 2 \left\langle \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k}, \hat{\mu}_k}) \\ \hat{J}_k - J(\pi_{\theta_k}, \hat{\mu}_k) \end{bmatrix} + \beta_k \begin{bmatrix} \Pi_{\mathcal{E}_\perp} \bar{G}^V(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) \\ \bar{G}^J(\theta_k, \hat{J}_k, \hat{\mu}_k) \end{bmatrix}, \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}} - V^{\pi_{\theta_k}, \hat{\mu}_k}) \\ J(\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}) - J(\pi_{\theta_k}, \hat{\mu}_k) \end{bmatrix} \right\rangle \\
&\quad + 2\beta_k \left\langle \Delta g_k, \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}} - V^{\pi_{\theta_k}, \hat{\mu}_k}) \\ J(\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}) - J(\pi_{\theta_k}, \hat{\mu}_k) \end{bmatrix} \right\rangle,
\end{aligned}
\tag{48}
$$

where the last inequality follows from the fact that $\Pi_{\mathcal{E}_\perp}$ has all singular values smaller than or equal to 1.

To bound the first term of (48),

$$
\begin{aligned}
&\left\| \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k}, \hat{\mu}_k}) \\ \hat{J}_k - J(\pi_{\theta_k}, \hat{\mu}_k) \end{bmatrix} + \beta_k \begin{bmatrix} \Pi_{\mathcal{E}_\perp} \bar{G}^V(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) \\ \bar{G}^J(\theta_k, \hat{J}_k, \hat{\mu}_k) \end{bmatrix} \right\|^2 \\
&\leq \|\Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k}, \hat{\mu}_k})\|^2 + (\hat{J}_k - J(\pi_{\theta_k}, \hat{\mu}_k))^2 + \beta_k^2 \|\Pi_{\mathcal{E}_\perp} \bar{G}^V(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\|^2 \\
&\quad + \beta_k \Big(\bar{G}^J(\theta_k, \hat{J}_k, \hat{\mu}_k)\Big)^2 + 2\beta_k \left\langle \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k}, \hat{\mu}_k}) \\ \hat{J}_k - J(\pi_{\theta_k}, \hat{\mu}_k) \end{bmatrix}, \begin{bmatrix} \Pi_{\mathcal{E}_\perp} \bar{G}^V(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) \\ \bar{G}^J(\theta_k, \hat{J}_k, \hat{\mu}_k) \end{bmatrix} \right\rangle \\
&\leq \|\Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k}, \hat{\mu}_k})\|^2 + (\hat{J}_k - J(\pi_{\theta_k}, \hat{\mu}_k))^2 + \beta_k^2 \|\bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\|^2 \\
&\quad - \gamma\beta_k \|\Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k}, \hat{\mu}_k})\|^2 - \gamma\beta_k (\hat{J}_k - J(\pi_{\theta_k}, \hat{\mu}_k))^2 \\
&= (1 - \gamma\beta_k)\|\Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k}, \hat{\mu}_k})\|^2 + (1 - \gamma\beta_k)(\hat{J}_k - J(\pi_{\theta_k}, \hat{\mu}_k))^2 \\
&\quad + \beta_k^2 \|\bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) - \bar{G}(\theta_k, V^{\pi_{\theta_k}, \hat{\mu}_k}, J(\pi_{\theta_k}, \hat{\mu}_k), \hat{\mu}_k)\|^2 \\
&\leq (1 - \gamma\beta_k)\|\Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k}, \hat{\mu}_k})\|^2 + (1 - \gamma\beta_k)(\hat{J}_k - J(\pi_{\theta_k}, \hat{\mu}_k))^2 \\
&\quad + L_G^2 \beta_k^2 \Big(\|\Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k}, \hat{\mu}_k})\| + |\hat{J}_k - J(\pi_{\theta_k}, \hat{\mu}_k)|\Big)^2 \\
&\leq (1 - \gamma\beta_k + 2L_G^2\beta_k^2)\|\Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k}, \hat{\mu}_k})\|^2 + (1 - \gamma\beta_k + 2L_G^2\beta_k^2)(\hat{J}_k - J(\pi_{\theta_k}, \hat{\mu}_k))^2 \\
&\leq (1 - \frac{\gamma\beta_k}{2})(\varepsilon_k^V + \varepsilon_k^J),
\end{aligned}
\tag{49}
$$

where the second inequality applies Lemma 10, the first equation uses the $\bar{G}(\theta, V^{\pi_\theta, \mu}, J(\pi_\theta, \mu), \mu) = 0$ for any θ, μ, third inequality follows from the Lipschitz continuity of operator $\bar{G}$ established in Lemma 3, and the final inequality follows from the step size condition $\beta_k \leq \frac{\gamma}{4L_G^2}$.

To treat the second and third term of (48), we use the boundedness of Δg_k and the Lipschitz continuity conditions from Lemma 1

$$
\begin{aligned}
&\beta_k^2 \|\Delta g_k\|^2 + \left\| \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}} - V^{\pi_{\theta_k}, \hat{\mu}_k}) \\ J(\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}) - J(\pi_{\theta_k}, \hat{\mu}_k) \end{bmatrix} \right\|^2 \\
&\leq \beta_k^2 \|\Delta g_k\|^2 + \|\Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}} - V^{\pi_{\theta_k}, \hat{\mu}_k})\|^2 + (J(\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}) - J(\pi_{\theta_k}, \hat{\mu}_k))^2 \\
&\leq 4B_G^2 \beta_k^2 + (L_V \|\theta_{k+1} - \theta_k\| + L_V \|\hat{\mu}_{k+1} - \hat{\mu}_k\|)^2 + (L_V \|\theta_{k+1} - \theta_k\| + L_V \|\hat{\mu}_{k+1} - \hat{\mu}_k\|)^2
\end{aligned}
$$

$$
\begin{aligned}
&= 4B_G^2\beta_k^2 + 2L_V^2\left(\alpha_k\|f_k\| + \xi_k\|h_k\|\right)^2 \\
&= 4B_G^2\beta_k^2 + 2L_V^2\xi_k^2\left(B_F + B_H\right)^2 \\
&\leq 4L_V^2(B_F^2 + B_G^2 + B_H^2)\beta_k^2,
\end{aligned}
\tag{50}
$$

where we combine terms using the step size condition $\alpha_k \leq \xi_k \leq \beta_k$.

The fourth term of (48) can be bounded leveraging the result in (49) as follows

$$
\begin{aligned}
&2\beta_k\left\langle\begin{bmatrix}\Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k},\hat{\mu}_k}) \\ \hat{J}_k - J(\pi_{\theta_k},\hat{\mu}_k)\end{bmatrix} + \beta_k\begin{bmatrix}\Pi_{\mathcal{E}_\perp}\bar{G}^V(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k) \\ \bar{G}^J(\theta_k,\hat{J}_k,\hat{\mu}_k)\end{bmatrix}, \begin{bmatrix}\Pi_{\mathcal{E}_\perp}\Delta g_k^V \\ \Delta g_k^J\end{bmatrix}\right\rangle \\
&\leq \frac{\gamma\beta_k}{8}\left\|\begin{bmatrix}\Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k},\hat{\mu}_k}) \\ \hat{J}_k - J(\pi_{\theta_k},\hat{\mu}_k)\end{bmatrix} + \beta_k\begin{bmatrix}\Pi_{\mathcal{E}_\perp}\bar{G}^V(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k) \\ \bar{G}^J(\theta_k,\hat{J}_k,\hat{\mu}_k)\end{bmatrix}\right\|^2 + \frac{8\beta_k}{\gamma}\left\|\begin{bmatrix}\Pi_{\mathcal{E}_\perp}\Delta g_k^V \\ g_k^J\end{bmatrix}\right\|^2 \\
&\leq \frac{\gamma\beta_k}{8}(1 - \frac{\gamma\beta_k}{2})(\varepsilon_k^V + \varepsilon_k^J) + \frac{8\beta_k}{\gamma}\|\Delta g_k\|^2 \\
&\leq \frac{\gamma\beta_k}{8}(\varepsilon_k^V + \varepsilon_k^J) + \frac{8\beta_k}{\gamma}\|\Delta g_k\|^2.
\end{aligned}
\tag{51}
$$

Similarly, for the fifth term of (48), we have

$$
\begin{aligned}
&2\left\langle\begin{bmatrix}\Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k},\hat{\mu}_k}) \\ \hat{J}_k - J(\pi_{\theta_k},\hat{\mu}_k)\end{bmatrix} + \beta_k\begin{bmatrix}\Pi_{\mathcal{E}_\perp}\bar{G}^V(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k) \\ \bar{G}^J(\theta_k,\hat{J}_k,\hat{\mu}_k)\end{bmatrix}, \begin{bmatrix}\Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_{k+1}},\hat{\mu}_{k+1}} - V^{\pi_{\theta_k},\hat{\mu}_k}) \\ J(\pi_{\theta_{k+1}},\hat{\mu}_{k+1}) - J(\pi_{\theta_k},\hat{\mu}_k)\end{bmatrix}\right\rangle \\
&\leq \frac{\gamma\beta_k}{8}\left\|\begin{bmatrix}\Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k},\hat{\mu}_k}) \\ \hat{J}_k - J(\pi_{\theta_k},\hat{\mu}_k)\end{bmatrix} + \beta_k\begin{bmatrix}\Pi_{\mathcal{E}_\perp}\bar{G}^V(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k) \\ \bar{G}^J(\theta_k,\hat{J}_k,\hat{\mu}_k)\end{bmatrix}\right\|^2 \\
&\quad + \frac{8}{\gamma\beta_k}\left\|\begin{bmatrix}\Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_{k+1}},\hat{\mu}_{k+1}} - V^{\pi_{\theta_k},\hat{\mu}_k}) \\ J(\pi_{\theta_{k+1}},\hat{\mu}_{k+1}) - J(\pi_{\theta_k},\hat{\mu}_k)\end{bmatrix}\right\|^2 \\
&\leq \frac{\gamma\beta_k}{8}(\varepsilon_k^V + \varepsilon_k^J) + \frac{8}{\gamma\beta_k}\|V^{\pi_{\theta_{k+1}},\hat{\mu}_{k+1}} - V^{\pi_{\theta_k},\hat{\mu}_k}\|^2 \\
&\quad + \frac{8}{\gamma\beta_k}(J(\pi_{\theta_{k+1}},\hat{\mu}_{k+1}) - J(\pi_{\theta_k},\,\hat{\mu}_k))^2 \\
&\leq \frac{\gamma\beta_k}{8}(\varepsilon_k^V + \varepsilon_k^J) + \frac{16L_V^2}{\gamma\beta_k}\left(\|\pi_{\theta_{k+1}} - \pi_{\theta_k}\|^2 + \|\hat{\mu}_{k+1} - \hat{\mu}_k\|^2\right) \\
&\quad + \frac{16L_V^2}{\gamma\beta_k}\left(\|\pi_{\theta_{k+1}} - \pi_{\theta_k}\|^2 + \|\hat{\mu}_{k+1} - \hat{\mu}_k\|^2\right) \\
&\leq \frac{\gamma\beta_k}{8}(\varepsilon_k^V + \varepsilon_k^J) + \frac{32L_V^2}{\gamma\beta_k}(\alpha_k^2\|f_k\|^2 + \xi_k^2\|h_k\|^2) \\
&\leq \frac{\gamma\beta_k}{8}(\varepsilon_k^V + \varepsilon_k^J) \\
&\quad + \frac{32L_V^2}{\gamma\beta_k}\left(4\alpha_k^2(\|\Delta f_k\|^2 + L_F^2\varepsilon_k^V + L_F^2(L_V + 1)^2\varepsilon_k^\mu + \varepsilon_k^\pi) + 2\xi_k^2(\|\Delta h_k\|^2 + L_H^2\varepsilon_k^\mu)\right) \\
&\leq \frac{\gamma\beta_k}{8}(\varepsilon_k^V + \varepsilon_k^J) + \frac{128L_V^2\alpha_k^2}{\gamma\beta_k}\left(\|\Delta f_k\|^2 + L_F^2\varepsilon_k^V + 4L_F^2L_V^2\varepsilon_k^\mu + \varepsilon_k^\pi\right) \\
&\quad + \frac{64L_V^2\xi_k^2}{\gamma\beta_k}\left(\|\Delta h_k\|^2 + L_H^2\varepsilon_k^\mu\right),
\end{aligned}
\tag{52}
$$

where the third inequality applies Lemma 1 and the fifth inequality applies Lemma 4.

The final term of (48) can be bounded simply with the Cauchy-Schwarz inequality

$$
2\beta_k\left\langle\Delta g_k, \begin{bmatrix}\Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_{k+1}},\hat{\mu}_{k+1}} - V^{\pi_{\theta_k},\hat{\mu}_k}) \\ J(\pi_{\theta_{k+1}},\hat{\mu}_{k+1}) - J(\pi_{\theta_k},\hat{\mu}_k)\end{bmatrix}\right\rangle
$$

$$
\begin{aligned}
&\leq 2\beta_k \left\|\Delta g_k\right\| \left\| \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}} - V^{\pi_{\theta_k}, \hat{\mu}_k}) \\ J(\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}) - J(\pi_{\theta_k}, \hat{\mu}_k) \end{bmatrix} \right\| \\
&\leq 2\beta_k \left\|\Delta g_k\right\| \Big(\|\Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}} - V^{\pi_{\theta_k}, \hat{\mu}_k})\| + |J(\pi_{\theta_{k+1}}, \hat{\mu}_{k+1}) - J(\pi_{\theta_k}, \hat{\mu}_k)| \Big) \\
&\leq 4B_G\beta_k \cdot \big(L_V(\|\pi_{\theta_{k+1}} - \pi_{\theta_k}\| + \|\hat{\mu}_{k+1} - \hat{\mu}_k\|) + L_V(\|\pi_{\theta_{k+1}} - \pi_{\theta_k}\| + \|\hat{\mu}_{k+1} - \hat{\mu}_k\|)\big) \\
&\leq 8L_V B_G \beta_k (B_F \alpha_k + B_H \xi_k) \\
&\leq 16 L_V B_F B_G B_H \beta_k \xi_k. \qquad (53)
\end{aligned}
$$

Plugging (49)-(53) into (48), we get

$$
\begin{aligned}
&\varepsilon_{k+1}^V + \varepsilon_{k+1}^J \\
&\leq (1 - \frac{\gamma\beta_k}{2})(\varepsilon_k^V + \varepsilon_k^J) + 4L_V^2(B_F^2 + B_G^2 + B_H^2)\beta_k^2 + \frac{\gamma\beta_k}{8}(\varepsilon_k^V + \varepsilon_k^J) + \frac{8\beta_k}{\gamma}\|\Delta g_k\|^2 \\
&\quad + \frac{\gamma\beta_k}{8}(\varepsilon_k^V + \varepsilon_k^J) + \frac{128L_V^2\alpha_k^2}{\gamma\beta_k}\left(\|\Delta f_k\|^2 + L_F^2\varepsilon_k^V + 4L_F^2L_V^2\varepsilon_k^\mu + \varepsilon_k^\pi\right) \\
&\quad + \frac{64L_V^2\xi_k^2}{\gamma\beta_k}\left(\|\Delta h_k\|^2 + L_H^2\varepsilon_k^\mu\right) + 16L_V B_F B_G B_H \beta_k \xi_k \\
&\leq (1 - \frac{\gamma\beta_k}{4})(\varepsilon_k^V + \varepsilon_k^J) + \frac{128L_V^2\alpha_k^2}{\gamma\beta_k}\|\Delta f_k\|^2 + \frac{8\beta_k}{\gamma}\|\Delta g_k\|^2 + \frac{64L_V^2\xi_k^2}{\gamma\beta_k}\|\Delta h_k\|^2 \\
&\quad + \frac{128L_V^2\alpha_k^2}{\gamma\beta_k}(L_F^2\varepsilon_k^V + \varepsilon_k^\pi) + \frac{192L_V^2\xi_k^2}{\gamma\beta_k}\varepsilon_k^\mu + 28L_V^2B_F^2B_G^2B_H^2\beta_k^2,
\end{aligned}
$$

where we use the conditions $\xi_k \leq \beta_k$ and $\alpha_k \leq \frac{L_H}{2L_F L_V}\xi_k$ in the last inequality to simplify and combine terms.

■

D.6 Proof of Proposition 6

The proof of Proposition 6 relies on the following lemma, the proof of which is presented in Sec.E.11.

Lemma 11 *We have for all* $k \geq \tau_k$

$$
\begin{aligned}
&\mathbb{E}[\langle \Delta g_k, G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\rangle] \\
&\qquad\qquad \leq (22L + 2|\mathcal{A}|)L_F L_{TV} B_F B_G^2 B_H \tau_k^2 \lambda_{k-\tau_k}.
\end{aligned}
$$

By the update rule of f_k,

$$
\begin{aligned}
\Delta g_{k+1} &= g_{k+1} - \bar{G}(\theta_{k+1}, \hat{V}_{k+1}, \hat{J}_{k+1}, \hat{\mu}_{k+1}) \\
&= (1 - \lambda_k)g_k + \lambda_k\bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_{k+1}, \hat{V}_{k+1}, \hat{J}_{k+1}, \hat{\mu}_{k+1}) \\
&= (1 - \lambda_k)g_k + \lambda_k\bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) - \bar{G}(\theta_{k+1}, \hat{V}_{k+1}, \hat{J}_{k+1}, \hat{\mu}_{k+1}) \\
&\qquad + \lambda_k\Big(G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\Big) \\
&= (1 - \lambda_k)\Delta g_k + \Big(\bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) - \bar{G}(\theta_{k+1}, \hat{V}_{k+1}, \hat{J}_{k+1}, \hat{\mu}_{k+1})\Big) \\
&\qquad + \lambda_k\Big(G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\Big).
\end{aligned}
$$

Taking the norm, we have

$$
\begin{aligned}
&\|\Delta g_{k+1}\|^2 \\
&= (1 - \lambda_k)^2\|\Delta G_k\|^2 + \|\bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) - \bar{G}(\theta_{k+1}, \hat{V}_{k+1}, \hat{J}_{k+1}, \hat{\mu}_{k+1})\|^2 \\
&\qquad + \lambda_k^2\|G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\|^2
\end{aligned}
$$

$$\begin{aligned}
&\quad + (1-\lambda_k)\langle \Delta g_k, \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) - \bar{G}(\theta_{k+1}, \hat{V}_{k+1}, \hat{J}_{k+1}, \hat{\mu}_{k+1})\rangle \\
&\quad + (1-\lambda_k)\lambda_k\langle \Delta g_k, G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\rangle \\
&\quad + \lambda_k\langle \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) - \bar{G}(\theta_{k+1}, \hat{V}_{k+1}, \hat{J}_{k+1}, \hat{\mu}_{k+1}), G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\rangle \\
&\leq (1-\lambda_k)^2\|\Delta g_k\|^2 + 2\|\bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) - \bar{G}(\theta_{k+1}, \hat{V}_{k+1}, \hat{J}_{k+1}, \hat{\mu}_{k+1})\|^2 \\
&\quad + 2\lambda_k^2\|G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\|^2 \\
&\quad + \frac{\lambda_k}{2}\|\Delta g_k\|^2 + \frac{2}{\lambda_k}\|\bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) - \bar{G}(\theta_{k+1}, \hat{V}_{k+1}, \hat{J}_{k+1}, \hat{\mu}_{k+1})\|^2 \\
&\quad + (1-\lambda_k)\lambda_k\langle \Delta g_k, G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\rangle \\
&\leq (1-\lambda_k)\|\Delta g_k\|^2 + (-\frac{\lambda_k}{2} + \lambda_k^2)\|\Delta g_k\|^2 \\
&\quad + \frac{4}{\lambda_k}\|\bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) - \bar{G}(\theta_{k+1}, \hat{V}_{k+1}, \hat{J}_{k+1}, \hat{\mu}_{k+1})\|^2 \\
&\quad + 8B_G^2\lambda_k^2 + (1-\lambda_k)\lambda_k\langle \Delta g_k, G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\rangle, \qquad (54)
\end{aligned}$$

where the final inequality follows from the step size condition $\lambda_k \leq 1$ and the boundedness of operator F.

Taking expectation and plugging in the result of Lemma 7, we can simplify (54) as

$$\begin{aligned}
&\mathbb{E}[\|\Delta g_{k+1}\|^2] \\
&\leq \mathbb{E}\Big[(1-\lambda_k)\|\Delta g_k\|^2 + (-\frac{\lambda_k}{2} + \lambda_k^2)\|\Delta g_k\|^2 \\
&\qquad + \frac{4}{\lambda_k}\|\bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) - \bar{G}(\theta_{k+1}, \hat{V}_{k+1}, \hat{J}_{k+1}, \hat{\mu}_{k+1})\|^2 \\
&\qquad + 8B_G^2\lambda_k^2 + (1-\lambda_k)\lambda_k\langle \Delta g_k, G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\rangle\Big] \\
&\leq (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] + (-\frac{\lambda_k}{2} + \lambda_k^2)\mathbb{E}[\|\Delta g_k\|^2] + 8B_G^2\lambda_k^2 \\
&\qquad + \frac{4L_G^2}{\lambda_k}\mathbb{E}\left[\Big(\|\theta_k - \theta_{k+1}\| + \|\hat{V}_k - \hat{V}_{k+1}\| + |\hat{J}_k - \hat{J}_{k+1}| + \|\hat{\mu}_k - \hat{\mu}_{k+1}\|\Big)^2\right] \\
&\qquad + (1-\lambda_k)\lambda_k \cdot (22L + 2|\mathcal{A}|)L_F L_{TV} B_F B_G^2 B_H \tau_k^2 \lambda_{k-\tau_k} \\
&\leq (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] + (-\frac{\lambda_k}{2} + \lambda_k^2)\mathbb{E}[\|\Delta g_k\|^2] + (30L + 2|\mathcal{A}|)L_F L_{TV} B_F B_G^2 B_H \tau_k^2 \lambda_k \lambda_{k-\tau_k} \\
&\qquad + \frac{4L_G^2}{\lambda_k}\mathbb{E}[\big(\alpha_k\|f_k\| + \beta_k\|g_k^V\| + \beta_k|g_k^J| + \xi_k\|h_k\|\big)^2] \\
&\leq (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] + (-\frac{\lambda_k}{2} + \lambda_k^2)\mathbb{E}[\|\Delta g_k\|^2] + (30L + 2|\mathcal{A}|)L_F L_{TV} B_F B_G^2 B_H \tau_k^2 \lambda_k \lambda_{k-\tau_k} \\
&\qquad + \frac{4L_G^2}{\lambda_k}\mathbb{E}[\Big(\alpha_k\|f_k\| + \sqrt{|\mathcal{S}|+1}\beta_k\|g_k\| + \xi_k\|h_k\|\Big)^2] \\
&\leq (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] + (-\frac{\lambda_k}{2} + \lambda_k^2)\mathbb{E}[\|\Delta g_k\|^2] + (30L + 2|\mathcal{A}|)L_F L_{TV} B_F B_G^2 B_H \tau_k^2 \lambda_k \lambda_{k-\tau_k} \\
&\qquad + \frac{12L_G^2\alpha_k^2}{\lambda_k}\mathbb{E}[\bigg(\|\Delta f_k\| + L_F\sqrt{\varepsilon_k^V} + L_F(L_V+1)\sqrt{\varepsilon_k^\mu} + \sqrt{\varepsilon_k^\pi}\bigg)^2] \\
&\qquad + \frac{24|\mathcal{S}|L_G^2\beta_k^2}{\lambda_k}\mathbb{E}[\bigg(\|\Delta g_k\| + L_G\sqrt{\varepsilon_k^V} + L_G\sqrt{\varepsilon_k^J}\bigg)^2] + \frac{12L_G^2\xi_k^2}{\lambda_k}\mathbb{E}[\bigg(\|\Delta h_k\| + L_H\sqrt{\varepsilon_k^\mu}\bigg)^2], \qquad (55)
\end{aligned}$$

where the fourth inequality follows from $\|g_k^V\| + |g_k^J| \leq \|g_k^V\|_1 + |g_k^J| = \|g_k\|_1 \leq \sqrt{|\mathcal{S}|+1}\|g_k\|$.

We can simplify the sum of the last three terms as follows

$$\frac{12L_G^2\alpha_k^2}{\lambda_k}\mathbb{E}\left[\bigg(\|\Delta f_k\| + L_F\sqrt{\varepsilon_k^V} + L_F(L_V+1)\sqrt{\varepsilon_k^\mu} + \sqrt{\varepsilon_k^\pi}\bigg)^2\right]$$

$$
\begin{aligned}
&+ \frac{24|\mathcal{S}|L_G^2\beta_k^2}{\lambda_k}\mathbb{E}[\Big(\|\Delta g_k\| + L_G\sqrt{\varepsilon_k^V} + L_G\sqrt{\varepsilon_k^J}\Big)^2] + \frac{12L_G^2\xi_k^2}{\lambda_k}\mathbb{E}[\Big(\|\Delta h_k\| + L_H\sqrt{\varepsilon_k^\mu}\Big)^2] \\
&\leq \frac{48L_G^2\alpha_k^2}{\lambda_k}\mathbb{E}[\|\Delta f_k\|^2 + L_F^2\varepsilon_k^V + 4L_F^2L_V^2\varepsilon_k^\mu + \varepsilon_k^\pi] + \frac{72|\mathcal{S}|L_G^2\beta_k^2}{\lambda_k}\mathbb{E}[\|\Delta g_k\|^2 + L_G^2\varepsilon_k^V + L_G^2\varepsilon_k^J] \\
&\quad + \frac{24L_G^2\xi_k^2}{\lambda_k}\mathbb{E}[\|\Delta h_k\|^2 + L_H^2\varepsilon_k^\mu] \\
&\leq \mathbb{E}\Big[\frac{48L_G^2\alpha_k^2}{\lambda_k}\|\Delta f_k\|^2 + \frac{72|\mathcal{S}|L_G^2\beta_k^2}{\lambda_k}\|\Delta g_k\|^2 + \frac{24L_G^2\xi_k^2}{\lambda_k}\|\Delta h_k\|^2 + \frac{48L_G^2\alpha_k^2}{\lambda_k}\varepsilon_k^\pi \\
&\quad + \frac{216L_F^2L_G^2L_H^2L_V^2\xi_k^2}{\lambda_k}\varepsilon_k^\mu + \frac{120|\mathcal{S}|L_F^2L_G^4\beta_k^2}{\lambda_k}(\varepsilon_k^V + \varepsilon_k^J)\Big]. \qquad (56)
\end{aligned}
$$

Combining (55) and (56), we have

$$
\begin{aligned}
&\mathbb{E}[\|\Delta g_{k+1}\|^2] \\
&\leq (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] + (-\frac{\lambda_k}{2} + \lambda_k^2)\mathbb{E}[\|\Delta g_k\|^2] + (30L + 2|\mathcal{A}|)L_F L_{TV} B_F B_G^2 B_H \tau_k^2 \lambda_k \lambda_{k-\tau_k} \\
&\quad + \frac{12L_G^2\alpha_k^2}{\lambda_k}\left(\|\Delta f_k\| + L_F\sqrt{\varepsilon_k^V} + L_F(L_V+1)\sqrt{\varepsilon_k^\mu} + \sqrt{\varepsilon_k^\pi}\right)^2 \\
&\quad + \frac{24|\mathcal{S}|L_G^2\beta_k^2}{\lambda_k}\left(\|\Delta g_k\| + L_G\sqrt{\varepsilon_k^V} + L_G\sqrt{\varepsilon_k^J}\right)^2 + \frac{12L_G^2\xi_k^2}{\lambda_k}\left(\|\Delta h_k\| + L_H\sqrt{\varepsilon_k^\mu}\right)^2 \\
&\leq (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] + (-\frac{\lambda_k}{2} + \lambda_k^2)\mathbb{E}[\|\Delta g_k\|^2] + (30L + 2|\mathcal{A}|)L_F L_{TV} B_F B_G^2 B_H \tau_k^2 \lambda_k \lambda_{k-\tau_k} \\
&\quad + \frac{48L_G^2\alpha_k^2}{\lambda_k}\mathbb{E}[\|\Delta f_k\|^2] + \frac{72|\mathcal{S}|L_G^2\beta_k^2}{\lambda_k}\mathbb{E}[\|\Delta g_k\|^2] + \frac{24L_G^2\xi_k^2}{\lambda_k}\mathbb{E}[\|\Delta h_k\|^2] + \frac{48L_G^2\alpha_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^\pi] \\
&\quad + \frac{216L_F^2L_G^2L_H^2L_V^2\xi_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^\mu] + \frac{120|\mathcal{S}|L_F^2L_G^4\beta_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^V + \varepsilon_k^J] \\
&\leq (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] + (-\frac{\lambda_k}{2} + \lambda_k^2 + \frac{72|\mathcal{S}|L_G^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta g_k\|^2] + \frac{48L_G^2\alpha_k^2}{\lambda_k}\mathbb{E}[\|\Delta f_k\|^2] \\
&\quad + \frac{24L_G^2\xi_k^2}{\lambda_k}\mathbb{E}[\|\Delta h_k\|^2] + \frac{48L_G^2\alpha_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^\pi] + \frac{216L_F^2L_G^2L_H^2L_V^2\xi_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^\mu] \\
&\quad + \frac{120|\mathcal{S}|L_F^2L_G^4\beta_k^2}{\lambda_k}\mathbb{E}[\varepsilon_k^V + \varepsilon_k^J] + (30L + 2|\mathcal{A}|)L_F L_{TV} B_F B_G^2 B_H \tau_k^2 \lambda_k \lambda_{k-\tau_k}.
\end{aligned}
$$

■

E Proof of Lemmas

E.1 Proof of Lemma 1

The Lipschitz continuity conditions of the value function and J function in the policy are proved in Lemma 3 and Lemma 2 of Kumar et al. [2024], respectively. The Lipschitz continuity in the mean field can be proved using the same line of argument under Assumption 3.

The Lipschitz gradient condition of J in θ is proved in Lemma 4 of Kumar et al. [2024] and can be extended to the gradient of J in μ by a similar argument.

■

E.2 Proof of Lemma 2

First, by definition in (22),

$$
\|F(\theta, V, \mu, s, a, s')\| = \|(r(s,a,\mu) + V(s'))\nabla_\theta \log \pi_\theta(a \mid s)\|
$$

$$\begin{aligned}&\leq (|r(s,a,\mu)| + |V(s')|)\|\nabla_\theta \log \pi_\theta(a \mid s)\|\\&\leq (1+B_V)\cdot 1\\&\leq B_V + 1,\end{aligned}$$

where the second inequality is due to the softmax function being Lipschitz with constant 1.

Similarly, we have

$$\begin{aligned}\|G^V(V,J,\mu,s,a,s')\| &= \|(r(s,a,\mu) - J + V(s') - V(s))e_s\|\\&\leq (|r(s,a,\mu)| + |J| + |V(s')| - |V(s)|)\|e_s\|\\&\leq (1+1+B_V+B_V)\cdot 1]\\&\leq 2B_V + 2,\end{aligned}$$

and

$$|G^J(J,\mu,s,a)| = |c_J(r(s,a,\mu) - J)| \leq 2c_J,$$

which implies

$$\|G(V,J,\mu,s,a,s')\| \leq \|G^V(V,J,\mu,s,a,s')\| + |G^J(J,\mu,s,a)| \leq 2(B_V + c_J + 2).$$

Finally, we have

$$\|H(\mu,s)\| = \|e_s - \mu\| \leq \|e_s\| + \|\mu\| \leq 2.$$

■

E.3 Proof of Lemma 3

By the definition of $\bar{F}(\theta,V,\mu)$ in (23),

$$\begin{aligned}&\|\bar{F}(\theta_1,V_1,\mu_1) - \bar{F}(\theta_2,V_2,\mu_2)\|\\&= \|\mathbb{E}_{s\sim\nu^{\pi_{\theta_1},\mu_1},a\sim\pi_{\theta_1}(\cdot|s),s'\sim\mathcal{P}^{\mu_1}(\cdot|s,a)}[F(\theta_1,V_1,\mu_1,s,a,s')]\\&\qquad - \mathbb{E}_{s\sim\nu^{\pi_{\theta_2},\mu_2},a\sim\pi_{\theta_2}(\cdot|s),s'\sim\mathcal{P}^{\mu_2}(\cdot|s,a)}[F(\theta_2,V_2,\mu_2,s,a,s')]\|\\&= \|\mathbb{E}_{s\sim\nu^{\pi_{\theta_1},\mu_1},a\sim\pi_{\theta_1}(\cdot|s),s'\sim\mathcal{P}^{\mu_1}(\cdot|s,a)}[F(\theta_1,\Pi_{\mathcal{E}_\perp}V_1,\mu_1,s,a,s')]\\&\qquad - \mathbb{E}_{s\sim\nu^{\pi_{\theta_2},\mu_2},a\sim\pi_{\theta_2}(\cdot|s),s'\sim\mathcal{P}^{\mu_2}(\cdot|s,a)}[F(\theta_2,\Pi_{\mathcal{E}_\perp}V_2,\mu_2,s,a,s')]\|\\&= \|\sum_{s,a,s'}(\nu^{\pi_{\theta_1},\mu_1}(s)\pi_{\theta_1}(a\mid s)\mathcal{P}^{\mu_1}(\cdot\mid s,a) - \nu^{\pi_{\theta_2},\mu_2}(s)\pi_{\theta_2}(a\mid s)\mathcal{P}^{\mu_2}(\cdot\mid s,a))F(\theta_2,\Pi_{\mathcal{E}_\perp}V_2,\mu_2,s,a,s')\\&\qquad + \mathbb{E}_{s\sim\nu^{\pi_{\theta_1},\mu_1},a\sim\pi_{\theta_1}(\cdot|s),s'\sim\mathcal{P}^{\mu_1}(\cdot|s,a)}[F(\theta_1,\Pi_{\mathcal{E}_\perp}V_1,\mu_1,s,a,s') - F(\theta_2,\Pi_{\mathcal{E}_\perp}V_2,\mu_2,s,a,s')]\|\\&\leq \|\mathbb{E}_{s\sim\nu^{\pi_{\theta_1},\mu_1},a\sim\pi_{\theta_1}(\cdot|s),s'\sim\mathcal{P}^{\mu_1}(\cdot|s,a,\mu_1)}[F(\theta_1,\Pi_{\mathcal{E}_\perp}V_1,\mu_1,s,a,s') - F(\theta_2,\Pi_{\mathcal{E}_\perp}V_2,\mu_2,s,a,s')]\|\\&\qquad + 2B_F d_{TV}(\nu^{\pi_{\theta_1},\mu_1}\otimes\pi_{\theta_1}\otimes\mathcal{P}^{\mu_1},\nu^{\pi_{\theta_2},\mu_2}\otimes\pi_{\theta_2}\otimes\mathcal{P}^{\mu_2}),\end{aligned}\tag{57}$$

where the inequality comes from the definition of TV distance in (24) and the second equation is a result of the fact that for any constant c

$$\begin{aligned}&\mathbb{E}_{s\sim\nu^{\pi_\theta,\mu},a\sim\pi_\theta(\cdot|s),s'\sim\mathcal{P}^\mu(\cdot|s,a)}[F(\theta,V+c\mathbf{1}_{|\mathcal{S}|},\mu,s,a,s')]\\&= \mathbb{E}_{s\sim\nu^{\pi_\theta,\mu},a\sim\pi_\theta(\cdot|s),s'\sim\mathcal{P}^\mu(\cdot|s,a)}[(r(s,a,\mu) + (V(s')+c) - (V(s)+c))\nabla_\theta\log\pi_\theta(a\mid s)]\\&= \mathbb{E}_{s\sim\nu^{\pi_\theta,\mu},a\sim\pi_\theta(\cdot|s),s'\sim\mathcal{P}^\mu(\cdot|s,a)}[(r(s,a,\mu) + V(s') - V(s))\nabla_\theta\log\pi_\theta(a\mid s)]\\&= \mathbb{E}_{s\sim\nu^{\pi_\theta,\mu},a\sim\pi_\theta(\cdot|s),s'\sim\mathcal{P}^\mu(\cdot|s,a)}[F(\theta,V,\mu,s,a,s')].\end{aligned}$$

For any s,a,s' we have from (22)

$$\|F(\theta_1,\Pi_{\mathcal{E}_\perp}V_1,\mu_1,s,a,s') - F(\theta_2,\Pi_{\mathcal{E}_\perp}V_2,\mu_2,s,a,s')\|$$

$$\begin{aligned}
&= \|(r(s,a,\mu_1) + \Pi_{\mathcal{E}_\perp} V_1(s') - \Pi_{\mathcal{E}_\perp} V_1(s))\nabla_\theta \log \pi_{\theta_1}(a \mid s) \\
&\qquad - (r(s,a,\mu_2) + \Pi_{\mathcal{E}_\perp} V_2(s') - \Pi_{\mathcal{E}_\perp} V_2(s))\nabla_\theta \log \pi_{\theta_2}(a \mid s)\| \\
&\leq |r(s,a,\mu_1) - r(s,a,\mu_2)| \|\nabla_\theta \log \pi_{\theta_1}(a \mid s)\| \\
&\qquad + |r(s,a,\mu_2)| \|\nabla_\theta \log \pi_{\theta_1}(a \mid s) - \nabla_\theta \log \pi_{\theta_2}(a \mid s)\| \\
&\qquad + |\Pi_{\mathcal{E}_\perp} V_1(s') - \Pi_{\mathcal{E}_\perp} V_1(s) - \Pi_{\mathcal{E}_\perp} V_2(s') + \Pi_{\mathcal{E}_\perp} V_2(s)| \|\nabla_\theta \log \pi_{\theta_1}(a \mid s)\| \\
&\qquad + |\Pi_{\mathcal{E}_\perp} V_2(s') - \Pi_{\mathcal{E}_\perp} V_2(s)| \|\nabla_\theta \log \pi_{\theta_1}(a \mid s) - \nabla_\theta \log \pi_{\theta_2}(a \mid s)\| \\
&\leq |r(s,a,\mu_1) - r(s,a,\mu_2)| + (1 + 2\|V\|)\|\nabla_\theta \log \pi_{\theta_1}(a \mid s) - \nabla_\theta \log \pi_{\theta_2}(a \mid s)\| \\
&\leq L\|\mu_1 - \mu_2\| + \|\nabla_\theta \log \pi_{\theta_1}(a \mid s) - \nabla_\theta \log \pi_{\theta_2}(a \mid s)\| \\
&\qquad + 2\|\Pi_{\mathcal{E}_\perp} V_1 - \Pi_{\mathcal{E}_\perp} V_2\| + 2\|\Pi_{\mathcal{E}_\perp} V_2\| \|\nabla_\theta \log \pi_{\theta_1}(a \mid s) - \nabla_\theta \log \pi_{\theta_2}(a \mid s)\| \\
&\leq 5(2B_V + 1)\|\theta_1 - \theta_2\| + L\|\mu_1 - \mu_2\| + 2\|\Pi_{\mathcal{E}_\perp} V_1 - \Pi_{\mathcal{E}_\perp} V_2\|, \qquad (58)
\end{aligned}$$

where the second inequality bounds $\|\log \pi_{\theta_1}(a \mid s)\|$ by 1 due to the softmax function being Lipschitz with constant 1, the third inequality follows from Assumption 3, and the final inequality is a result of the fact that the softmax function is smooth with constant 5 (see Agarwal et al. [2021][Lemma 52]).

Applying (58) and the relationship in (25) to (57), we have

$$\begin{aligned}
&\|\bar{F}(\theta_1, V_1, \mu_1) - \bar{F}(\theta_2, V_2, \mu_2)\| \\
&\leq \|\mathbb{E}_{s \sim \nu^{\pi_{\theta_1}, \mu_1}, a \sim \pi_{\theta_1}(\cdot|s), s' \sim \mathcal{P}^{\mu_1}(\cdot|s,a,\mu_1)}[F(\theta_1, \Pi_{\mathcal{E}_\perp} V_1, \mu_1, s, a, s') - F(\theta_2, \Pi_{\mathcal{E}_\perp} V_2, \mu_2, s, a, s')]\| \\
&\qquad + 2B_F d_{TV}(\nu^{\pi_{\theta_1}, \mu_1} \otimes \pi_{\theta_1} \otimes \mathcal{P}^{\mu_1}, \nu^{\pi_{\theta_2}, \mu_2} \otimes \pi_{\theta_2} \otimes \mathcal{P}^{\mu_2}) \\
&\leq 5(2B_V + 1)\|\theta_1 - \theta_2\| + L\|\mu_1 - \mu_2\| + 2\|\Pi_{\mathcal{E}_\perp} V_1 - \Pi_{\mathcal{E}_\perp} V_2\| \\
&\qquad + 2B_F L_{TV}(\|\theta_1 - \theta_2\| + \|\mu_1 - \mu_2\|) \\
&\leq (10B_V + L + 2B_F L_{TV} + 5)\left(\|\theta_1 - \theta_2\| + \|\mu_1 - \mu_2\| + \|\Pi_{\mathcal{E}_\perp} V_1 - \Pi_{\mathcal{E}_\perp} V_2\|\right).
\end{aligned}$$

Following a line of argument similar to (57),

$$\begin{aligned}
&\|\bar{G}(\theta_1, V_1, J_1, \mu_1) - \bar{G}(\theta_2, V_2, J_2, \mu_2)\| \\
&\leq \|\mathbb{E}_{s \sim \nu^{\pi_{\theta_1}, \mu_1}, a \sim \pi_{\theta_1}(\cdot|s), s' \sim \mathcal{P}^{\mu_1}(\cdot|s,a,\mu_1)}[G(\Pi_{\mathcal{E}_\perp} V_1, J_1, \mu_1, s, a, s') - G(\Pi_{\mathcal{E}_\perp} V_2, J_2, \mu_2, s, a, s')]\| \\
&\qquad + 2B_G d_{TV}(\nu^{\pi_{\theta_1}, \mu_1} \otimes \pi_{\theta_1} \otimes \mathcal{P}^{\mu_1}, \nu^{\pi_{\theta_2}, \mu_2} \otimes \pi_{\theta_2} \otimes \mathcal{P}^{\mu_2}). \qquad (59)
\end{aligned}$$

The first term of (59) can be bounded in a manner similar to (58). For any s, a, s', we have

$$\begin{aligned}
&\|G(\Pi_{\mathcal{E}_\perp} V_1, J_1, \mu_1, s, a, s') - G(\Pi_{\mathcal{E}_\perp} V_2, J_2, \mu_2, s, a, s')\| \\
&\leq \|(r(s,a,\mu_1) - J_1 + \Pi_{\mathcal{E}_\perp} V_1(s') - \Pi_{\mathcal{E}_\perp} V_1(s))e_s \\
&\qquad\qquad - (r(s,a,\mu_2) - J_2 + \Pi_{\mathcal{E}_\perp} V_2(s') - \Pi_{\mathcal{E}_\perp} V_2(s))e_s\| \\
&\qquad + c_J|r(s,a,\mu_1) - J_1 - r(s,a,\mu_2) + J_2| \\
&\leq |r(s,a,\mu_1) - r(s,a,\mu_2)| \|e_s\| + |J_1 - J_2| \|e_s\| + 2\|\Pi_{\mathcal{E}_\perp} V_1 - \Pi_{\mathcal{E}_\perp} V_2\| \|e_s\| \\
&\qquad + c_J|r(s,a,\mu_1) - r(s,a,\mu_2)| + c_J|J_1 - J_2| \\
&\leq (c_J + 1)|r(s,a,\mu_1) - r(s,a,\mu_2)| + (c_J + 1)|J_1 - J_2| + 2\|\Pi_{\mathcal{E}_\perp} V_1 - \Pi_{\mathcal{E}_\perp} V_2\| \\
&\leq (c_J + 1)L|\mu_1 - \mu_2| + (c_J + 1)|J_1 - J_2| + 2\|\Pi_{\mathcal{E}_\perp} V_1 - \Pi_{\mathcal{E}_\perp} V_2\|. \qquad (60)
\end{aligned}$$

Plugging (60) into (59), we get

$$\begin{aligned}
&\|\bar{G}(\theta_1, V_1, J_1, \mu_1) - \bar{G}(\theta_2, V_2, J_2, \mu_2)\| \\
&\leq \|\mathbb{E}_{s \sim \nu^{\pi_{\theta_1}, \mu_1}, a \sim \pi_{\theta_1}(\cdot|s), s' \sim \mathcal{P}^{\mu_1}(\cdot|s,a,\mu_1)}[G(\Pi_{\mathcal{E}_\perp} V_1, J_1, \mu_1, s, a, s') - G(\Pi_{\mathcal{E}_\perp} V_2, J_2, \mu_2, s, a, s')]\| \\
&\qquad + 2B_G d_{TV}(\nu^{\pi_{\theta_1}, \mu_1} \otimes \pi_{\theta_1} \otimes \mathcal{P}^{\mu_1}, \nu^{\pi_{\theta_2}, \mu_2} \otimes \pi_{\theta_2} \otimes \mathcal{P}^{\mu_2}) \\
&\leq (c_J + 1)L|\mu_1 - \mu_2| + (c_J + 1)|J_1 - J_2| + 2\|\Pi_{\mathcal{E}_\perp} V_1 - \Pi_{\mathcal{E}_\perp} V_2\| \\
&\qquad + 2B_G L_{TV}(\|\theta_1 - \theta_2\| + \|\mu_1 - \mu_2\|) \\
&\leq L_G\left(\|\theta_1 - \theta_2\| + \|\mu_1 - \mu_2\| + 2\|\Pi_{\mathcal{E}_\perp} V_1 - \Pi_{\mathcal{E}_\perp} V_2\| + |J_1 - J_2|\right),
\end{aligned}$$

with $L_G = 2B_G L_{TV} + (L+1)(c_J+1) + 2$.

Finally, again following steps similar to (57) we can show

$$\begin{aligned} &\|\bar{H}(\theta_1,\mu_1) - \bar{H}(\theta_2,\mu_2)\| \\ &\leq \|\mathbb{E}_{s\sim\nu^{\pi_{\theta_1},\mu_1}}[H(\mu_1,s) - H(\mu_2,s)]\| + 2B_H d_{TV}(\nu^{\pi_{\theta_1},\mu_1},\nu^{\pi_{\theta_2},\mu_2}). \end{aligned} \tag{61}$$

From the definition of $H(\mu,s)$ in (22), we have for any s

$$\|H(\mu_1,s) - H(\mu_2,s)\| = \|(e_s - \mu_1) - (e_s - \mu_2)\| = \|\mu_1 - \mu_2\|. \tag{62}$$

By Assumption 3,

$$\begin{aligned} d_{TV}(\nu^{\pi_{\theta_1},\mu_1},\nu^{\pi_{\theta_2},\mu_2}) &= \frac{1}{2}\|\nu^{\pi_{\theta_1},\mu_1} - \nu^{\pi_{\theta_2},\mu_2}\|_1 \\ &\leq L(\|\pi_{\theta_1} - \pi_{\theta_2}\| + \|\mu_1 - \mu_2\|) \\ &\leq L(\|\theta_1 - \theta_2\| + \|\mu_1 - \mu_2\|), \end{aligned} \tag{63}$$

where the final inequality is a result of the 1-Lipschitz continuity of the softmax function.

Plugging (62) and (63) into (61), we have

$$\begin{aligned} \|\bar{H}(\theta_1,\mu_1) - \bar{H}(\theta_2,\mu_2)\| &\leq \|\mathbb{E}_{s\sim\nu^{\pi_{\theta_1},\mu_1}}[H(\mu_1,s) - H(\mu_2,s)]\| + 2B_H d_{TV}(\nu^{\pi_{\theta_1},\mu_1},\nu^{\pi_{\theta_2},\mu_2}) \\ &\leq \|\mu_1 - \mu_2\| + L(\|\theta_1 - \theta_2\| + \|\mu_1 - \mu_2\|) \\ &\leq (L+1)(\|\theta_1 - \theta_2\| + \|\mu_1 - \mu_2\|). \end{aligned} \tag{64}$$

■

E.4 Proof of Lemma 4

By the definition Δf_k,

$$\begin{aligned} &\|f_k\| \\ &= \|\Delta f_k + \bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k) - \bar{F}(\theta_k,V^{\pi_{\theta_k},\mu^\star(\pi_{\theta_k})},\mu^\star(\pi_{\theta_k})) + \bar{F}(\theta_k,V^{\pi_{\theta_k},\mu^\star(\pi_{\theta_k})},\mu^\star(\pi_{\theta_k}))\| \\ &\leq \|\Delta f_k\| + \|\bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k) - \bar{F}(\theta_k,V^{\pi_{\theta_k},\mu^\star(\pi_{\theta_k})},\mu^\star(\pi_{\theta_k}))\| + \|\bar{F}(\theta_k,V^{\pi_{\theta_k},\mu^\star(\pi_{\theta_k})},\mu^\star(\pi_{\theta_k}))\| \\ &\leq \|\Delta f_k\| + L_F\|\Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_k},\mu^\star(\pi_{\theta_k})} - \hat{V}_k)\| + L_F\|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\| + \|\nabla_\theta J(\pi_{\theta_k},\mu)\mid_{\mu=\mu^\star(\pi_{\theta_k})}\| \\ &\leq \|\Delta f_k\| + L_F\|\Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_k},\mu^\star(\pi_{\theta_k})} - V^{\pi_{\theta_k},\hat{\mu}_k})\| + L_F\|\Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_k},\hat{\mu}_k} - \hat{V}_k)\| \\ &\qquad + L_F\|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\| + \sqrt{\varepsilon_k^\pi} \\ &\leq \|\Delta f_k\| + L_F\sqrt{\varepsilon_k^V} + L_F(L_V+1)\sqrt{\varepsilon_k^\mu} + \sqrt{\varepsilon_k^\pi}, \end{aligned}$$

where the last inequality follows from the Lipschitz continuity of the value function in the mean field and the fact that linear projection is non-expansive, and the second inequality follows from the Lipschitz continuity of operator F and the relationship

$$\nabla_\theta J(\pi_{\theta_k},\mu)\mid_{\mu=\mu^\star(\pi_{\theta_k})} = \bar{F}(\theta_k,V^{\pi_{\theta_k},\mu^\star(\pi_{\theta_k})},\mu^\star(\pi_{\theta_k})).$$

Similarly, by the definition of Δg_k, we have

$$\begin{aligned} \|g_k\| &= \|\Delta g_k + \bar{G}(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k) - \bar{G}(\theta_k,V^{\pi_{\theta_k},\hat{\mu}_k},J(\pi_{\theta_k},\hat{\mu}_k),\hat{\mu}_k)\| \\ &\leq \|\Delta g_k\| + \|\bar{G}(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k) - \bar{G}(\theta_k,V^{\pi_{\theta_k},\hat{\mu}_k},J(\pi_{\theta_k},\hat{\mu}_k),\hat{\mu}_k)\| \\ &\leq \|\Delta g_k\| + L_G\|\Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_k},\hat{\mu}_k} - \hat{V}_k)\| + L_G|J(\pi_{\theta_k},\hat{\mu}_k) - \hat{J}_k| \\ &= \|\Delta g_k\| + L_G\sqrt{\varepsilon_k^V} + L_G\sqrt{\varepsilon_k^J}, \end{aligned}$$

where the first equation follows from the fact that $G(\theta_k, V^{\pi_{\theta_k}, \hat{\mu}_k}, J(\pi_{\theta_k}, \hat{\mu}_k), \hat{\mu}_k) = 0$.

Finally, by the definition of Δh_k, we have

$$\begin{aligned}
\|h_k\| &= \|\Delta h_k + \bar{H}(\theta_k, \hat{\mu}_k)\| \\
&= \|\Delta h_k + \bar{H}(\theta_k, \hat{\mu}_k) - \bar{H}(\theta_k, \mu^\star(\pi_{\theta_k}))\| \\
&\leq \|\Delta h_k\| + \|\bar{H}(\theta_k, \hat{\mu}_k) - \bar{H}(\theta_k, \mu^\star(\pi_{\theta_k}))\| \\
&\leq \|\Delta h_k\| + L_H \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\| \\
&= \|\Delta h_k\| + L_H \sqrt{\epsilon_k^\mu},
\end{aligned}$$

where the second equation follows from the fact that $H(\theta_k, \mu^\star(\pi_{\theta_k})) = 0$.

■

E.5 Proof of Lemma 5

See Zhang et al. [2021a][Lemma 2] or Tsitsiklis and Van Roy [1999][Lemma 7].

E.6 Proof of Lemma 6

Adapted from Lemma 19 of Ganesh et al. [2024].

E.7 Proof of Lemma 7

The proof of this lemma proceeds in a manner similar to that of Lemma 9. We note that the samples generated in the algorithm follow the time-varying Markov chain

$$s_{k-\tau_k} \xrightarrow{\theta_{k-\tau_k}} a_{k-\tau_k} \xrightarrow{\hat{\mu}_{k-\tau_k}} s_{k-\tau_k+1} \xrightarrow{\theta_{k-\tau_k+1}} a_{k-\tau_k+1} \xrightarrow{\hat{\mu}_{k-\tau_k+1}} \cdots s_{k-1} \xrightarrow{\theta_{k-1}} a_{k-1} \xrightarrow{\hat{\mu}_{k-1}} s_k. \tag{65}$$

We construct an auxiliary Markov chain generated under a constant control

$$s_{k-\tau_k} \xrightarrow{\theta_{k-\tau_k}} a_{k-\tau_k} \xrightarrow{\hat{\mu}_{k-\tau_k}} \widetilde{s}_{k-\tau_k+1} \xrightarrow{\theta_{k-\tau_k}} \widetilde{a}_{k-\tau_k+1} \xrightarrow{\hat{\mu}_{k-\tau_k}} \cdots \widetilde{s}_{k-1} \xrightarrow{\theta_{k-\tau_k}} \widetilde{a}_{k-1} \xrightarrow{\hat{\mu}_{k-\tau_k}} \widetilde{s}_k \tag{66}$$

Let $\widetilde{\mu}$ denote the stationary distribution of state, action, and next state under (66). We denote $p_k(s,a,s') = \mathbb{P}(s_k = s, a_k = a, s_{k+1} = s')$ and $\widetilde{p}_k(s,a,s') = \mathbb{P}(\widetilde{s}_k = s, \widetilde{a}_k = a, \widetilde{s}_{k+1} = s')$ and define

$$\begin{aligned}
T_1 &\triangleq \mathbb{E}[\langle \Delta f_k - \Delta f_{k-\tau_k}, F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k)\rangle], \\
T_2 &\triangleq \mathbb{E}[\langle \Delta f_{k-\tau_k}, F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - F(\theta_k, \hat{V}_k, \hat{\mu}_k, \widetilde{s}_k, \widetilde{a}_k, \widetilde{s}_{k+1})\rangle], \\
T_3 &\triangleq \mathbb{E}[\langle \Delta f_{k-\tau_k}, F(\theta_k, \hat{V}_k, \hat{\mu}_k, \widetilde{s}_k, \widetilde{a}_k, \widetilde{s}_{k+1}) - \mathbb{E}_{(s,a,s')\sim\widetilde{\mu}}[F(\theta_k, \hat{V}_k, \hat{\mu}_k, s, a, s')]\rangle] \\
T_4 &\triangleq \mathbb{E}[\langle \Delta f_{k-\tau_k}, \mathbb{E}_{(s,a,s')\sim\widetilde{\mu}}[F(\theta_k, \hat{V}_k, \hat{\mu}_k, s, a, s')] - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k)\rangle].
\end{aligned}$$

It is obvious to see

$$\mathbb{E}[\langle \Delta f_k, F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k)\rangle] = T_1 + T_2 + T_3 + T_4. \tag{67}$$

We bound the terms individually. First, we treat T_1

$$\begin{aligned}
T_1 &= \mathbb{E}[\langle \Delta f_k - \Delta f_{k-\tau_k}, F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k)\rangle] \\
&\leq \mathbb{E}[\|f_k - f_{k-\tau_k}\| \|F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k)\|] \\
&\quad + \mathbb{E}\Big[\|\bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) - \bar{F}(\theta_{k-\tau_k}, \hat{V}_{k-\tau_k}, \hat{\mu}_{k-\tau_k})\| \\
&\qquad\qquad \cdot \|F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k)\|\Big]
\end{aligned}$$

$$
\begin{aligned}
&\leq 2B_F\sum_{t=0}^{\tau_k-1}\mathbb{E}[\|f_{k-t}-f_{k-t-1}\|]\\
&\qquad+2L_F B_F\sum_{t=0}^{\tau_k-1}\mathbb{E}[\|\theta_{k-t}-\theta_{k-t-1}\|+\|\hat{V}_{k-t}-\hat{V}_{k-t-1}\|+\|\hat{\mu}_{k-t}-\hat{\mu}_{k-t-1}\|]\\
&\leq 4B_F^2\tau_k\lambda_{k-\tau_k}+2L_F B_F\tau_k(B_F\alpha_{k-\tau_k}+B_G\beta_{k-\tau_k}+B_H\xi_{k-\tau_k})\\
&\leq 10L_F B_F^2 B_G B_H\tau_k\lambda_{k-\tau_k},
\end{aligned}
$$

where the second inequality bounds $\|F(\theta_k,\hat{V}_k,\hat{\mu}_k,s_k,a_k,s_{k+1})-\bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k)\|$ by $2B_F$ and $\|\bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k)-\bar{F}(\theta_{k-\tau_k},\hat{V}_{k-\tau_k},\hat{\mu}_{k-\tau_k})\|$ using the Lipschitz continuity established in Lemma 3. The last inequality follows from the step size condition $\alpha_k\leq\xi_k\leq\beta_k\leq\lambda_k$ for all k. The third inequality follows from the fact that $\|f_{k+1}-f_k\|=\lambda_k\|f_k-F(\theta_k,\hat{V}_k,\hat{\mu}_k,s_k,a_k,s_{k+1})\|\leq 2B_F\lambda_k$ for all k and that the per-iteration drift of θ_k, $\hat{V}_k$, and $\hat{\mu}_k$ can be similarly bounded

$$
\|\theta_{k+1}-\theta_k\|\leq B_F\alpha_k,\quad\|\hat{V}_{k+1}-\hat{V}_k\|\leq B_G\beta_k,\quad\|\hat{\mu}_{k+1}-\hat{\mu}_k\|\leq B_H\xi_k.
$$

We next bound T_2

$$
\begin{aligned}
T_2&=\mathbb{E}[\langle\Delta f_{k-\tau_k},F(\theta_k,\hat{V}_k,\hat{\mu}_k,s_k,a_k,s_{k+1})-F(\theta_k,\hat{V}_k,\hat{\mu}_k,\widetilde{s}_k,\widetilde{a}_k,\widetilde{s}_{k+1})\rangle]\\
&\leq 2B_F\mathbb{E}_{\mathcal{F}_{k-\tau_k}}[\mathbb{E}[\|F(\theta_k,\hat{V}_k,\hat{\mu}_k,s_k,a_k,s_{k+1})-F(\theta_k,\hat{V}_k,\hat{\mu}_k,\widetilde{s}_k,\widetilde{a}_k,\widetilde{s}_{k+1})\|\mid\mathcal{F}_{k-\tau_k}]]\\
&\leq 2B_F\mathbb{E}[\int_{\mathcal{S}}\int_{\mathcal{A}}\int_{\mathcal{S}}F(\theta_k,\hat{V}_k,\hat{\mu}_k,s,a,s')\,(p_k(s,a,s')-\widetilde{p}_k(s,a,s'))\,ds\,da\,ds']\\
&\leq 2B_F^2\mathbb{E}[d_{TV}(p_k,\widetilde{p}_k)].
\end{aligned}
$$

where the last inequality follows from the definition of TV distance in (24).

Applying Lemma B.2 from Wu et al. [2020], we then have

$$
\begin{aligned}
T_2&\leq 2B_F^2\mathbb{E}[d_{TV}(p_k,\widetilde{p}_k)]\\
&\leq 2B_F^2\mathbb{E}[d_{TV}(\mathbb{P}(s_k=\cdot),\mathbb{P}(\widetilde{s}_k=\cdot))+\frac{|\mathcal{A}|}{2}\|\theta_{k-1}-\theta_{k-\tau_k}\|]\\
&\leq 2B_F^2\mathbb{E}\Big[d_{TV}(\mathbb{P}(s_{k-1}=\cdot),\mathbb{P}(\widetilde{s}_{k-1}=\cdot))+L\|\theta_{k-1}-\theta_{k-\tau_k}\|+L\|\hat{\mu}_{k-1}-\hat{\mu}_{k-\tau_k}\|\\
&\qquad+\frac{|\mathcal{A}|}{2}\|\theta_{k-1}-\theta_{k-\tau_k}\|\Big]\\
&\leq|\mathcal{A}|B_F^2\mathbb{E}[\|\theta_{k-1}-\theta_{k-\tau_k}\|]+2LB_F^2\sum_{t=k-\tau_k}^{k-1}\mathbb{E}[\|\theta_t-\theta_{k-\tau_k}\|+\|\hat{\mu}_t-\hat{\mu}_{k-\tau_k}\|]\\
&\leq(2L+|\mathcal{A}|)B_F^2\tau_k^2(B_F\alpha_{k-\tau_k}+B_H\xi_{k-\tau_k})\\
&\leq(4L+2|\mathcal{A}|)B_F^3B_H\tau_k^2\lambda_{k-\tau_k},
\end{aligned}
$$

where the third inequality is a result of (16), and the fourth inequality recursively applies the inequality above it.

The term T_3 is proportional to the distance between the distribution of the auxiliary Markov chain (66) at time k and its stationary distribution. To bound T_3,

$$
\begin{aligned}
T_3&=\mathbb{E}[\langle\Delta f_{k-\tau_k},F(\theta_k,\hat{V}_k,\hat{\mu}_k,\widetilde{s}_k,\widetilde{a}_k,\widetilde{s}_{k+1})-\mathbb{E}_{(s,a,s')\sim\widetilde{\mu}}[F(\theta_k,\hat{V}_k,\hat{\mu}_k,s,a,s')]\rangle]\\
&\leq 2B_F\mathbb{E}_{\mathcal{F}_{k-\tau_k}}[\mathbb{E}[\|F(\theta_k,\hat{V}_k,\hat{\mu}_k,\widetilde{s}_k,\widetilde{a}_k,\widetilde{s}_{k+1})-\mathbb{E}_{(s,a,s')\sim\widetilde{\mu}}[F(\theta_k,\hat{V}_k,\hat{\mu}_k,s,a,s')]\|\mid\mathcal{F}_{k-\tau_k}]]\\
&\leq 2B_F\mathbb{E}[\int_{\mathcal{S}}\int_{\mathcal{A}}\int_{\mathcal{S}}F(\theta_k,\hat{V}_k,\hat{\mu}_k,s,a,s')\,(\widetilde{p}_k(s)-\widetilde{\mu}(s))\,ds\,da\,ds']\\
&\leq 2B_F^2\mathbb{E}[d_{TV}(\widetilde{p}_k,\widetilde{\mu})]\\
&\leq 2B_F^2\alpha_k,
\end{aligned}
$$

where the final inequality follows from the definition of the mixing time τ_k as the number of iterations for the TV distance between $\widetilde{p}_k$ and $\widetilde{\mu}$ to drop below α_k.

Finally, we bound the term T_4

$$
\begin{aligned}
T_4 &= \mathbb{E}[\langle \Delta f_{k-\tau_k}, \mathbb{E}_{(s,a,s')\sim\widetilde{\mu}}[F(\theta_k,\hat{V}_k,\hat{\mu}_k,s,a,s')] - \bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k)\rangle]\\
&\leq 2B_F\mathbb{E}[\|\mathbb{E}_{(s,a,s')\sim\widetilde{\mu}}[F(\theta_k,\hat{V}_k,\hat{\mu}_k,s,a,s')] - \bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k)\|]\\
&\leq 2B_F^2\mathbb{E}[d_{TV}(\widetilde{\mu},\nu^{\pi_{\theta_k},\hat{\mu}_k}\otimes\pi_{\theta_k}\otimes\mathcal{P}^{\hat{\mu}_k})]\\
&\leq 2L_{TV}B_F^2\mathbb{E}[\|\pi_{\theta_k}-\pi_{\theta_{k-\tau_k}}\| + \|\hat{\mu}_k-\hat{\mu}_{k-\tau_k}\|]\\
&\leq 2L_{TV}B_F^2\tau_k(B_F\alpha_{k-\tau_k}+B_H\xi_{k-\tau_k})\\
&\leq 4L_{TV}B_F^3B_H\xi_{k-\tau_k},
\end{aligned}
$$

where the third inequality applies the result in (25).

Collecting the bounds on T_1-T_4 and plugging them into (67), we get

$$
\begin{aligned}
&\mathbb{E}[\langle \Delta f_k, F(\theta_k,\hat{V}_k,\hat{\mu}_k,s_k,a_k,s_{k+1}) - \bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k)\rangle]\\
&= T_1+T_2+T_3+T_4\\
&\leq 10L_FB_F^2B_GB_H\tau_k\lambda_{k-\tau_k} + (4L+2|\mathcal{A}|)B_F^3B_H\tau_k^2\lambda_{k-\tau_k} + 2B_F^2\alpha_k + 4L_{TV}B_F^3B_H\xi_{k-\tau_k}\\
&\leq (20L+2|\mathcal{A}|)L_FL_{TV}B_F^3B_GB_H^2\tau_k^2\lambda_{k-\tau_k}.
\end{aligned}
$$

■

E.8 Proof of Lemma 8

By the definition of $\bar{H}$, we have for any $\mu\in\Delta_{\mathcal{S}}$

$$
\begin{aligned}
&\langle \mu-\mu^\star(\pi_\theta), \bar{H}(\theta,\mu)-\bar{H}(\theta,\mu^\star(\pi_\theta))\rangle\\
&= \langle \mu-\mu^\star(\pi_\theta), \mu^\star(\pi_\theta)-\mu\rangle + \langle \mu-\mu^\star(\pi_\theta), \nu^{\pi_\theta,\mu}-\nu^{\pi_\theta,\mu^\star(\pi_\theta)}\rangle\\
&\leq -\|\mu-\mu^\star(\pi_\theta)\|^2 + \|\mu-\mu^\star(\pi_\theta)\|\|\nu^{\pi_\theta,\mu}-\nu^{\pi_\theta,\mu^\star(\pi_\theta)}\|\\
&\leq -(1-\delta)\|\mu-\mu^\star(\pi_\theta)\|^2,
\end{aligned}
$$

where the second inequality follows from Assumption 2.

■

E.9 Proof of Lemma 9

The cause of the gap between $\mathbb{E}[e_{s_k}]$ and $\mathbb{E}_{s\sim\nu^{\pi_{\theta_k},\hat{\mu}_k}}[e_s]$ is a time-varying Markovian noise. To elaborate, we first show how the sample s_k is generated below

$$
s_{k-\tau_k}\xrightarrow{\theta_{k-\tau_k},\hat{\mu}_{k-\tau_k}} s_{k-\tau_k+1}\xrightarrow{\theta_{k-\tau_k+1},\hat{\mu}_{k-\tau_k+1}}\cdots s_{k-1}\xrightarrow{\theta_{k-1},\hat{\mu}_{k-1}} s_k. \tag{68}
$$

This Markov chain is "time-varying" as its stationary distribution changes over iterations as the control changes. We introduce an auxiliary Markov chain, which is "time-invariant" in the sense that it is generated under a constant control, starting from state $s_{k-\tau_k}$.

$$
s_{k-\tau_k}\xrightarrow{\theta_{k-\tau_k},\hat{\mu}_{k-\tau_k}} \widetilde{s}_{k-\tau_k+1}\xrightarrow{\theta_{k-\tau_k},\hat{\mu}_{k-\tau_k}}\cdots \widetilde{s}_{k-1}\xrightarrow{\theta_{k-\tau_k},\hat{\mu}_{k-\tau_k}} \widetilde{s}_k. \tag{69}
$$

Defining

$$
\begin{aligned}
T_1 &\triangleq \mathbb{E}[\langle \Delta h_k-\Delta h_{k-\tau_k}, e_{s_k}-\mathbb{E}_{s\sim\nu^{\pi_{\theta_k},\hat{\mu}_k}}[e_s]\rangle]\\
T_2 &\triangleq \mathbb{E}[\langle \Delta h_{k-\tau_k}, e_{s_k}-e_{\widetilde{s}_k}\rangle]\\
T_3 &\triangleq \mathbb{E}[\langle \Delta h_{k-\tau_k}, e_{\widetilde{s}_k}-\mathbb{E}_{s\sim\nu^{\pi_{\theta_{k-\tau_k}},\hat{\mu}_{k-\tau_k}}}[e_s]\rangle]\\
T_4 &\triangleq \mathbb{E}[\langle \Delta h_{k-\tau_k}, \mathbb{E}_{s\sim\nu^{\pi_{\theta_{k-\tau_k}},\hat{\mu}_{k-\tau_k}}}[e_s]-\mathbb{E}_{s\sim\nu^{\pi_{\theta_k},\hat{\mu}_k}}[e_s]\rangle],
\end{aligned}
$$

we see that

$$\mathbb{E}[\langle \Delta h_k, e_{s_k} - \mathbb{E}_{s\sim\nu^{\pi_{\theta_k},\hat{\mu}_k}}[e_s]\rangle] = T_1 + T_2 + T_3 + T_4. \tag{70}$$

We bound the terms individually. First, we treat T_1

$$\begin{aligned}
T_1 &= \mathbb{E}[\langle h_k - h_{k-\tau_k} + \bar{H}(\theta_{k-\tau_k}, \hat{\mu}_{k-\tau_k}) - \bar{H}(\theta_k, \hat{\mu}_k), e_{s_k} - \mathbb{E}_{s\sim\nu^{\pi_{\theta_k},\hat{\mu}_k}}[e_s]\rangle] \\
&\leq \mathbb{E}[\|h_k - h_{k-\tau_k}\| \|e_{s_k} - \mathbb{E}_{s\sim\nu^{\pi_{\theta_k},\hat{\mu}_k}}[e_s]\|] \\
&\qquad + \mathbb{E}[\|\bar{H}(\theta_k, \hat{\mu}_k) - \bar{H}(\theta_{k-\tau_k}, \hat{\mu}_{k-\tau_k})\| \|e_{s_k} - \mathbb{E}_{s\sim\nu^{\pi_{\theta_k},\hat{\mu}_k}}[e_s]\|] \\
&\leq 2\sum_{t=0}^{\tau_k-1} \mathbb{E}[\|h_{k-t} - h_{k-t-1}\|] + 2L_H \sum_{t=0}^{\tau_k-1} \mathbb{E}[\|\theta_{k-t} - \theta_{k-t-1}\| + \|\hat{\mu}_{k-t} - \hat{\mu}_{k-t-1}\|] \\
&\leq 4B_H\tau_k\lambda_{k-\tau_k} + 2B_F\tau_k\alpha_{k-\tau_k} + 2B_H\tau_k\xi_{k-\tau_k} \\
&\leq 8B_F B_H \tau_k \lambda_{k-\tau_k},
\end{aligned}$$

where the last inequality follows from the step size condition $\alpha_k \leq \xi_k \leq \lambda_k$ for all k, and the third inequality follows from the fact that $\|h_{k+1} - h_k\| \leq \lambda_k \|h_k + \hat{\mu}_k - e_{s_k}\| \leq 2B_H\lambda_k$ for all k and that the per-iteration drift of θ_k and $\hat{\mu}_k$ can be similarly bounded

$$\|\theta_{k+1} - \theta_k\| \leq B_F\alpha_k, \quad \|\hat{\mu}_{k+1} - \hat{\mu}_k\| \leq B_H\xi_k.$$

We next bound T_2. We denote $p_k(s) = \mathbb{P}(s_k = s)$ and $\widetilde{p}_k(s) = \mathbb{P}(\widetilde{s}_k = s)$.

$$\begin{aligned}
T_2 &= \mathbb{E}_{\mathcal{F}_{k-\tau_k}}[\mathbb{E}[\langle h_{k-\tau_k} - \bar{H}(\theta_{k-\tau_k}, \hat{\mu}_{k-\tau_k}), e_{s_k} - e_{\widetilde{s}_k}\rangle \mid \mathcal{F}_{k-\tau_k}]] \\
&\leq 2B_H \mathbb{E}_{\mathcal{F}_{k-\tau_k}}[\mathbb{E}[\|e_{s_k} - e_{\widetilde{s}_k}\| \mid \mathcal{F}_{k-\tau_k}]] \\
&\leq 2B_H \mathbb{E}[\int_{\mathcal{S}} e_s\,(p_k(s) - \widetilde{p}_k(s))\,ds] \\
&\leq 2B_H \mathbb{E}[d_{TV}(p_k, \widetilde{p}_k)] \\
&\leq 2B_H \mathbb{E}[d_{TV}(p_{k-1}, \widetilde{p}_{k-1}) + L\|\theta_{k-1} - \theta_{k-\tau_k}\| + L\|\hat{\mu}_{k-1} - \hat{\mu}_{k-\tau_k}\|] \\
&\leq 2LB_H \sum_{t=k-\tau_k}^{k-1} \mathbb{E}[\|\theta_t - \theta_{k-\tau_k}\| + \|\hat{\mu}_t - \hat{\mu}_{k-\tau_k}\|] \\
&\leq 2LB_H\tau_k^2(B_F\alpha_{k-\tau_k} + B_H\xi_{k-\tau_k}) \\
&\leq 4LB_F B_H^2 \tau_k^2 \lambda_{k-\tau_k},
\end{aligned}$$

where the third inequality follows from the definition of TV distance in (24), and the fourth and fifth inequalities are a result of (16).

The term T_3 is proportional to the distance between the distribution of the auxiliary Markov chain (69) at time k and its stationary distribution. Let $\widetilde{\mu}$ denote the stationary distribution of (69). We can bound this term as follows under Assumption 1

$$\begin{aligned}
T_3 &= \mathbb{E}[\langle \Delta h_{k-\tau_k}, e_{\widetilde{s}_k} - \mathbb{E}_{s\sim\nu^{\pi_{\theta_{k-\tau_k}},\hat{\mu}_{k-\tau_k}}}[e_s]\rangle] \\
&\leq 2B_H \mathbb{E}_{\mathcal{F}_{k-\tau_k}}[\mathbb{E}[\|e_{\widetilde{s}_k} - \mathbb{E}_{s\sim\nu^{\pi_{\theta_{k-\tau_k}},\hat{\mu}_{k-\tau_k}}}[e_s]\| \mid \mathcal{F}_{k-\tau_k}]] \\
&\leq 2B_H \mathbb{E}[\int_{\mathcal{S}} e_s\,(\widetilde{p}_k(s) - \widetilde{\mu}(s))\,ds] \\
&\leq 2B_H \mathbb{E}[d_{TV}(\widetilde{p}_k, \widetilde{\mu})] \\
&\leq 2B_H\alpha_k,
\end{aligned}$$

where the final inequality follows from the definition of the mixing time τ_k as the number of iterations for the TV distance between $\widetilde{p}_k$ and $\widetilde{\mu}$ to drop below α_k.

The term T_4 can be treated by the Lipschitz continuity of ν

$$T_4 = \mathbb{E}[\langle \Delta h_{k-\tau_k}, \mathbb{E}_{s\sim\nu^{\pi_{\theta_{k-\tau_k}},\hat{\mu}_{k-\tau_k}}}[e_s] - \mathbb{E}_{s\sim\nu^{\pi_{\theta_k},\hat{\mu}_k}}[e_s]\rangle]$$

$$
\begin{aligned}
&\leq 2B_H \mathbb{E}[\|\nu^{\pi_{\theta_{k-\tau_k}}, \hat{\mu}_{k-\tau_k}} - \nu^{\pi_{\theta_k}, \hat{\mu}_k}\|] \\
&\leq 2B_H L \mathbb{E}[\|\pi_{\theta_k} - \pi_{\theta_{k-\tau_k}}\|] + 2B_H \delta \mathbb{E}[\|\hat{\mu}_{k-\tau_k} - \hat{\mu}_k\|] \\
&\leq 2B_H L \sum_{t=k-\tau_k}^{k} \mathbb{E}[\|\alpha_t f_t\|] + 2B_H L \sum_{t=k-\tau_k}^{k} \mathbb{E}[\|\xi_t h_t\|] \\
&\leq 2B_H L \tau_k \left(B_F \alpha_{k-\tau_k} + B_H \xi_{k-\tau_k}\right) \\
&\leq 2L B_F B_H^2 \xi_{k-\tau_k}
\end{aligned}
$$

where the last inequality follows from the step size condition $\alpha_k \leq \xi_k$ for all k.

Collecting the bounds on T_1-T_4 and plugging them into (70), we get

$$
\begin{aligned}
&\mathbb{E}[\langle \Delta h_k, e_{s_k} - \mathbb{E}_{s \sim \nu^{\pi_{\theta_k}, \hat{\mu}_k}}[e_s] \rangle] \\
&= T_1 + T_2 + T_3 + T_4 \\
&\leq 8 B_F B_H \tau_k \lambda_{k-\tau_k} + 4 L B_F B_H^2 \tau_k^2 \lambda_{k-\tau_k} + 2 B_H \alpha_k + 2 L B_F B_H^2 \xi_{k-\tau_k} \\
&\leq 16 L B_F B_H^2 \tau_k^2 \lambda_{k-\tau_k}.
\end{aligned}
$$

■

E.10 Proof of Lemma 10

By the definition of operators G^V and G^J in (22), for any $V \in \mathbb{R}^{|\mathcal{S}|}$ and $J \in \mathbb{R}$

$$
\begin{aligned}
&\left\langle \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(V - V^{\pi_\theta, \mu}) \\ J - J(\pi_\theta, \mu) \end{bmatrix}, \begin{bmatrix} \Pi_{\mathcal{E}_\perp} \bar{G}^V(\theta, V, J, \mu) \\ \bar{G}^J(\theta, J, \mu) \end{bmatrix} \right\rangle \\
&\leq \langle \Pi_{\mathcal{E}_\perp}(V - V^{\pi_\theta, \mu}), \Pi_{\mathcal{E}_\perp} \mathbb{E}_{s \sim \nu^{\pi_\theta, \mu}, a \sim \pi_\theta(\cdot|s), s' \sim \mathcal{P}^\mu(\cdot|s,a)}[r(s,a,\mu) - J + e_s(e_{s'} - e_s)^\top V] \rangle \\
&\quad + c_J \langle J - J(\pi_\theta, \mu), \mathbb{E}_{s \sim \nu^{\pi_\theta, \mu}, a \sim \pi_\theta(\cdot|s)}[r(s,a,\mu) - J] \rangle \\
&= \langle \Pi_{\mathcal{E}_\perp}(V - V^{\pi_\theta, \mu}), \Pi_{\mathcal{E}_\perp} \mathbb{E}_{s \sim \nu^{\pi_\theta, \mu}, a \sim \pi_\theta(\cdot|s), s' \sim \mathcal{P}^\mu(\cdot|s,a)} \left[\left(r(s,a,\mu) - J(\pi_\theta, \mu) + (e_{s'} - e_s)^\top \Pi_{\mathcal{E}_\perp} V \right) e_s \right] \rangle \\
&\quad + \langle \Pi_{\mathcal{E}_\perp}(V - V^{\pi_\theta, \mu}), \Pi_{\mathcal{E}_\perp} \mathbb{E}_{s \sim \nu^{\pi_\theta, \mu}}[(J(\pi_\theta, \mu) - J) e_s] \rangle \\
&\quad + c_J \langle J - J(\pi_\theta, \mu), \mathbb{E}_{s \sim \nu^{\pi_\theta, \mu}, a \sim \pi_\theta(\cdot|s)}[r(s,a,\mu) - J] \rangle \\
&= \langle \Pi_{\mathcal{E}_\perp}(V - V^{\pi_\theta, \mu}), \Pi_{\mathcal{E}_\perp} \mathbb{E}_{s \sim \nu^{\pi_\theta, \mu}, a \sim \pi_\theta(\cdot|s), s' \sim \mathcal{P}^\mu(\cdot|s,a)} \left[e_s (e_{s'} - e_s)^\top \right] \Pi_{\mathcal{E}_\perp} \left(V - V^{\pi_\theta, \mu} \right) \rangle \\
&\quad + \langle \Pi_{\mathcal{E}_\perp}(V - V^{\pi_\theta, \mu}), \mathbb{E}_{s \sim \nu^{\pi_\theta, \mu}}[(J(\pi_\theta, \mu) - J) e_s] \rangle - c_J (J - J(\pi_\theta, \mu))^2 \\
&\leq \left(\Pi_{\mathcal{E}_\perp}(V - V^{\pi_\theta, \mu}) \right)^\top \Pi_{\mathcal{E}_\perp} \mathbb{E}_{s \sim \nu^{\pi_\theta, \mu}, a \sim \pi_\theta(\cdot|s), s' \sim \mathcal{P}^\mu(\cdot|s,a)} \left[e_s (e_{s'} - e_s)^\top \right] \Pi_{\mathcal{E}_\perp} \left(V - V^{\pi_\theta, \mu} \right) \\
&\quad + \frac{\gamma}{2} \|\Pi_{\mathcal{E}_\perp}(V - V^{\pi_\theta, \mu})\|^2 + \frac{1}{2\gamma} \|\mathbb{E}_{s \sim \nu^{\pi_\theta, \mu}}[(J(\pi_\theta, \mu) - J) e_s]\|^2 - c_J (J - J(\pi_\theta, \mu))^2 \\
&= \left(\Pi_{\mathcal{E}_\perp}(V - V^{\pi_\theta, \mu}) \right)^\top \mathbb{E}_{s \sim \nu^{\pi_\theta, \mu}, a \sim \pi_\theta(\cdot|s), s' \sim \mathcal{P}^\mu(\cdot|s,a)} \left[e_s (e_{s'} - e_s)^\top \right] \Pi_{\mathcal{E}_\perp} \left(V - V^{\pi_\theta, \mu} \right) \\
&\quad + \frac{\gamma}{2} \|\Pi_{\mathcal{E}_\perp}(V - V^{\pi_\theta, \mu})\|^2 + \frac{1}{2\gamma} \|\mathbb{E}_{s \sim \nu^{\pi_\theta, \mu}}[(J(\pi_\theta, \mu) - J) e_s]\|^2 - c_J (J - J(\pi_\theta, \mu))^2 \\
&\leq -\frac{\gamma}{2} \|\Pi_{\mathcal{E}_\perp}(V - V^{\pi_\theta, \mu})\|^2 - \frac{1}{2\gamma} (J - J(\pi_\theta, \mu))^2,
\end{aligned}
$$

where the second inequality follows from the fact that $\langle \vec{a}, \vec{b} \rangle \leq \frac{c}{2}\|\vec{a}\|^2 + \frac{1}{2c}\|\vec{b}\|^2$ for any vectors $\vec{a}, \vec{b}$ and scalar $c > 0$, the third inequality applies Lemma 5 and the condition $c_J \geq 1/\gamma$, the third equation uses the property of the projection matrix $\Pi_{\mathcal{E}_\perp}^2 = \Pi_{\mathcal{E}_\perp} = \Pi_{\mathcal{E}_\perp}^\top$, and the second equation is a result of the equation below

$$
\mathbb{E}_{s \sim \nu^{\pi_\theta, \mu}, a \sim \pi_\theta(\cdot|s), s' \sim \mathcal{P}^\mu(\cdot|s,a)} \left[\left(r(s,a,\mu) - J(\pi_\theta, \mu) + (e_{s'} - e_s)^\top \Pi_{\mathcal{E}_\perp} V^{\pi_\theta, \mu} \right) e_s \right] = 0.
$$

Since $\gamma \in (0,1)$, we have $\frac{1}{2\gamma} \geq \frac{\gamma}{2}$. This leads to the claimed result.

■

E.11 Proof of Lemma 11

The proof of this lemma proceeds in a manner similar to that of Lemma 7. We note that the samples generated in the algorithm follow the time-varying Markov chain

$$s_{k-\tau_k} \xrightarrow{\theta_{k-\tau_k}} a_{k-\tau_k} \xrightarrow{\hat{\mu}_{k-\tau_k}} s_{k-\tau_k+1} \xrightarrow{\theta_{k-\tau_k+1}} a_{k-\tau_k+1} \xrightarrow{\hat{\mu}_{k-\tau_k+1}} \cdots s_{k-1} \xrightarrow{\theta_{k-1}} a_{k-1} \xrightarrow{\hat{\mu}_{k-1}} s_k. \tag{71}$$

We construct an auxiliary Markov chain generated under a constant control

$$s_{k-\tau_k} \xrightarrow{\theta_{k-\tau_k}} a_{k-\tau_k} \xrightarrow{\hat{\mu}_{k-\tau_k}} \widetilde{s}_{k-\tau_k+1} \xrightarrow{\theta_{k-\tau_k}} \widetilde{a}_{k-\tau_k+1} \xrightarrow{\hat{\mu}_{k-\tau_k}} \cdots \widetilde{s}_{k-1} \xrightarrow{\theta_{k-\tau_k}} \widetilde{a}_{k-1} \xrightarrow{\hat{\mu}_{k-\tau_k}} \widetilde{s}_k \tag{72}$$

Let $\widetilde{\mu}$ denote the stationary distribution of state, action, and next state under (72). We denote $p_k(s,a,s') = \mathbb{P}(s_k = s, a_k = a, s_{k+1} = s')$ and $\widetilde{p}_k(s,a,s') = \mathbb{P}(\widetilde{s}_k = s, \widetilde{a}_k = a, \widetilde{s}_{k+1} = s')$ and define

$$\begin{aligned}
T_1 &\triangleq \mathbb{E}[\langle \Delta g_k - \Delta g_{k-\tau_k}, G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\rangle],\\
T_2 &\triangleq \mathbb{E}[\langle \Delta g_{k-\tau_k}, G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, \widetilde{s}_k, \widetilde{a}_k, \widetilde{s}_{k+1})\rangle],\\
T_3 &\triangleq \mathbb{E}[\langle \Delta g_{k-\tau_k}, G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, \widetilde{s}_k, \widetilde{a}_k, \widetilde{s}_{k+1}) - \mathbb{E}_{(s,a,s')\sim\widetilde{\mu}}[G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s, a, s')]]\rangle]\\
T_4 &\triangleq \mathbb{E}[\langle \Delta g_{k-\tau_k}, \mathbb{E}_{(s,a,s')\sim\widetilde{\mu}}[G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s, a, s')] - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\rangle].
\end{aligned}$$

It is obvious to see

$$\mathbb{E}[\langle \Delta g_k, G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\rangle\rangle] = T_1 + T_2 + T_3 + T_4. \tag{73}$$

We bound the terms individually. First, we treat T_1

$$\begin{aligned}
T_1 &= \mathbb{E}[\langle \Delta g_k - \Delta g_{k-\tau_k}, G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\rangle]\\
&\leq \mathbb{E}[\|g_k - g_{k-\tau_k}\|\|G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\|]\\
&\quad + \mathbb{E}\Big[\|\bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) - \bar{G}(\theta_{k-\tau_k}, \hat{V}_{k-\tau_k}, \hat{J}_{k-\tau_k}, \hat{\mu}_{k-\tau_k})\|\\
&\qquad\qquad \cdot \|G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\|\Big]\\
&\leq 2B_G \sum_{t=0}^{\tau_k-1} \|g_{k-t} - g_{k-t-1}\|\\
&\quad + 2L_G B_G \sum_{t=0}^{\tau_k-1} \Big(\|\theta_{k-t} - \theta_{k-t-1}\| + \|\hat{V}_{k-t} - \hat{V}_{k-t-1}\| + |\hat{J}_{k-t} - \hat{J}_{k-t-1}| + \|\hat{\mu}_{k-t} - \hat{\mu}_{k-t-1}\|\Big)\\
&\leq 4B_G^2 \tau_k \lambda_{k-\tau_k} + 2L_G B_G \tau_k (B_F \alpha_{k-\tau_k} + B_G \beta_{k-\tau_k} + B_G \beta_{k-\tau_k} + B_H \xi_{k-\tau_k})\\
&\leq 12 L_G B_F B_G^2 B_H \tau_k \lambda_{k-\tau_k},
\end{aligned}$$

where the second inequality bounds $\|G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\|$ by $2B_G$ and $\|\bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k) - \bar{G}(\theta_{k-\tau_k}, \hat{V}_{k-\tau_k}, \hat{J}_{k-\tau_k}, \hat{\mu}_{k-\tau_k})\|$ using the Lipschitz continuity established in Lemma 3. The last inequality follows from the step size condition $\alpha_k \leq \xi_k \leq \beta_k \leq \lambda_k$ for all k. The third inequality follows from the fact that $\|g_{k+1} - g_k\| = \lambda_k \|g_k - G(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1})\| \leq 2B_G \lambda_k$ for all k and that the per-iteration drift of θ_k, $\hat{V}_k$, and $\hat{\mu}_k$ can be similarly bounded due to Lemma 2

$$\|\theta_{k+1} - \theta_k\| \leq B_F \alpha_k,\ \|\hat{V}_{k+1} - \hat{V}_k\| \leq B_G \beta_k,\ |\hat{J}_{k+1} - \hat{J}_k| \leq B_G \beta_k,\ \|\hat{\mu}_{k+1} - \hat{\mu}_k\| \leq B_H \xi_k.$$

We next bound T_2

$$\begin{aligned}
T_2 &= \mathbb{E}[\langle \Delta g_{k-\tau_k}, G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, \widetilde{s}_k, \widetilde{a}_k, \widetilde{s}_{k+1})\rangle]\\
&\leq 2B_G \mathbb{E}_{\mathcal{F}_{k-\tau_k}}[\mathbb{E}[\|G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, \widetilde{s}_k, \widetilde{a}_k, \widetilde{s}_{k+1})\| \mid \mathcal{F}_{k-\tau_k}]]\\
&\leq 2B_G \mathbb{E}[\int_{\mathcal{S}} \int_{\mathcal{A}} \int_{\mathcal{S}} G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s, a, s')\,(p_k(s,a,s') - \widetilde{p}_k(s,a,s'))\,ds\,da\,ds']
\end{aligned}$$

$$\leq 2B_G^2 \mathbb{E}[d_{TV}(p_k, \widetilde{p}_k)].$$

where the last inequality follows from the definition of TV distance in (24).

Applying Lemma B.2 from Wu et al. [2020], we then have

$$\begin{aligned}
&T_2\\
&\leq 2B_G^2 \mathbb{E}[d_{TV}(p_k, \widetilde{p}_k)]\\
&\leq 2B_G^2 \mathbb{E}[d_{TV}(\mathbb{P}(s_k = \cdot), \mathbb{P}(\widetilde{s}_k = \cdot)) + \frac{|\mathcal{A}|}{2}\|\theta_{k-1} - \theta_{k-\tau_k}\|]\\
&\leq 2B_G^2 \mathbb{E}[d_{TV}(\mathbb{P}(s_{k-1} = \cdot), \mathbb{P}(\widetilde{s}_{k-1} = \cdot)) + L\|\theta_{k-1} - \theta_{k-\tau_k}\| + L\|\hat{\mu}_{k-1} - \hat{\mu}_{k-\tau_k}\| + \frac{|\mathcal{A}|}{2}\|\theta_{k-1} - \theta_{k-\tau_k}\|]\\
&\leq |\mathcal{A}| B_G^2 \mathbb{E}[\|\theta_{k-1} - \theta_{k-\tau_k}\|] + 2LB_G^2 \sum_{t=k-\tau_k}^{k-1} \mathbb{E}[\|\theta_t - \theta_{k-\tau_k}\| + \|\hat{\mu}_t - \hat{\mu}_{k-\tau_k}\|]\\
&\leq (2L + |\mathcal{A}|) B_G^2 \tau_k^2 (B_F \alpha_{k-\tau_k} + B_H \xi_{k-\tau_k})\\
&\leq (4L + 2|\mathcal{A}|) B_F B_G^2 B_H \tau_k^2 \lambda_{k-\tau_k},
\end{aligned}$$

where the third inequality is a result of Assumption 3, and the fourth inequality recursively applies the inequality above it.

The term T_3 is proportional to the distance between the distribution of the auxiliary Markov chain (72) at time k and its stationary distribution. To bound T_3,

$$\begin{aligned}
T_3 &= \mathbb{E}[\langle \Delta g_{k-\tau_k}, G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, \widetilde{s}_k, \widetilde{a}_k, \widetilde{s}_{k+1}) - \mathbb{E}_{(s,a,s')\sim\widetilde{\mu}}[G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s, a, s')]\rangle]\\
&\leq 2B_G \mathbb{E}_{\mathcal{F}_{k-\tau_k}}[\mathbb{E}[\|G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, \widetilde{s}_k, \widetilde{a}_k, \widetilde{s}_{k+1}) - \mathbb{E}_{(s,a,s')\sim\widetilde{\mu}}[G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s, a, s')]\| \mid \mathcal{F}_{k-\tau_k}]]\\
&\leq 2B_G \mathbb{E}[\int_{\mathcal{S}}\int_{\mathcal{A}}\int_{\mathcal{S}} G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s, a, s')\,(\widetilde{p}_k(s) - \widetilde{\mu}(s))\, ds\, da\, ds']\\
&\leq 2B_G^2 \mathbb{E}[d_{TV}(\widetilde{p}_k, \widetilde{\mu})]\\
&\leq 2B_G^2 \alpha_k,
\end{aligned}$$

where the final inequality follows from the definition of the mixing time τ_k as the number of iterations for the TV distance between $\widetilde{p}_k$ and $\widetilde{\mu}$ to drop below α_k.

Finally, we bound the term T_4

$$\begin{aligned}
T_4 &= \mathbb{E}[\langle \Delta g_{k-\tau_k}, \mathbb{E}_{(s,a,s')\sim\widetilde{\mu}}[G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s, a, s')] - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\rangle]\\
&\leq 2B_G \mathbb{E}[\|\mathbb{E}_{(s,a,s')\sim\widetilde{\mu}}[G(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k, s, a, s')] - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k)\|]\\
&\leq 2B_G^2 \mathbb{E}[d_{TV}(\widetilde{\mu}, \nu^{\pi_{\theta_k}, \hat{\mu}_k} \otimes \pi_{\theta_k} \otimes \mathcal{P}^{\hat{\mu}_k})]\\
&\leq 2L_{TV} B_G^2 \left(\|\pi_{\theta_k} - \pi_{\theta_{k-\tau_k}}\| + \|\hat{\mu}_k - \hat{\mu}_{k-\tau_k}\|\right)\\
&\leq 2L_{TV} B_G^2 \tau_k (B_F \alpha_{k-\tau_k} + B_H \xi_{k-\tau_k})\\
&\leq 4L_{TV} B_F B_G^2 B_H \xi_{k-\tau_k},
\end{aligned}$$

where the third inequality applies the result in (25).

Collecting the bounds on T_1-T_4 and plugging them into (73), we get

$$\begin{aligned}
&\mathbb{E}[\langle \Delta g_k, G(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{\mu}_k)\rangle]\\
&= T_1 + T_2 + T_3 + T_4\\
&\leq 12 L_F B_F B_G^2 B_H \tau_k \lambda_{k-\tau_k} + (4L + 2|\mathcal{A}|) B_F B_G^2 B_H \tau_k^2 \lambda_{k-\tau_k} + 2B_G^2 \alpha_k + 4L_{TV} B_F B_G^2 B_H \xi_{k-\tau_k}\\
&\leq (22L + 2|\mathcal{A}|) L_F L_{TV} B_F B_G^2 B_H \tau_k^2 \lambda_{k-\tau_k}.
\end{aligned}$$

■

F Details for Example 1

We prove that the mentioned class of MFGs satisfies (5) with $\rho = 2$ and $\kappa = 0$ when $p = 1$ and note that a similar line of argument can be made for other p.

Equivalent to (5) with $\rho = 2, \kappa = 0, p = 0$ is

$$J(\pi', \mu^\star(\pi)) - J(\pi', \mu^\star(\pi')) \leq J(\pi, \mu^\star(\pi)) - J(\pi', \mu^\star(\pi)). \tag{74}$$

As the transition kernel does not depend on μ here, we use ν^π to denote the stationary distribution of states under policy π. Note in this case that $\mu^\star(\pi) = \nu^\pi$.

We first compute $J(\pi', \mu^\star(\pi))$

$$J(\pi', \mu^\star(\pi)) = \langle \nu^{\pi'}, q \sum_a \pi'(a \mid \cdot) r(\cdot, a, \nu^\pi) \rangle = q \sum_s \nu^{\pi'}(s) \nu^\pi(s). \tag{75}$$

Similarly, we have

$$J(\pi, \mu^\star(\pi)) = q \sum_s \Big(\nu^\pi(s)\Big)^2, \quad J(\pi', \mu^\star(\pi')) = q \sum_s (\nu^{\pi'}(s)\Big)^2$$

As a result,

$$\begin{aligned} J(\pi', \mu^\star(\pi) - J(\pi', \mu^\star(\pi')) &= q \sum_s \nu^{\pi'}(s) \Big(\nu^\pi(s) - \nu^{\pi'}(s)\Big), \\ J(\pi, \mu^\star(\pi) - J(\pi', \mu^\star(\pi)) &= q \sum_s \nu^{\pi}(s) \Big(\nu^\pi(s) - \nu^{\pi'}(s)\Big). \end{aligned}$$

This obvious leads to (74) as

$$\Big(J(\pi, \mu^\star(\pi) - J(\pi', \mu^\star(\pi))\Big) - \Big(J(\pi', \mu^\star(\pi) - J(\pi', \mu^\star(\pi'))\Big) = q \sum_s \Big(\nu^\pi(s) - \nu^{\pi'}(s)\Big)^2 \geq 0.$$

Next, we provide the detailed derivation on the equilibrium of the MFG in the special case $|\mathcal{S}| = |\mathcal{A}| = 2$ under the transition kernel such that in either state $s \in \{s_1, s_2\}$, the action a_1 (resp. a_2) leads the next state to s_1 (resp. s_2) with probability $p = 3/4$. A visualization of the transition kernel can be found in Figure. 5.

Under any policy π, the transition matrix is

$$P^\pi = \left[\begin{array}{cc} p\pi(a_1 \mid s_1) + (1-p)\pi(a_2 \mid s_1) & p\pi(a_1 \mid s_2) + (1-p)\pi(a_2 \mid s_2) \\ (1-p)\pi(a_1 \mid s_1) + p\pi(a_2 \mid s_1) & (1-p)\pi(a_1 \mid s_2) + p\pi(a_2 \mid s_2) \end{array} \right],$$

under which the stationary distribution (induced mean field) is

$$\nu^\pi \propto \left[\frac{\pi(a_2 \mid s_2) + p - 2p\pi(a_2 \mid s_2)}{\pi(a_1 \mid s_1) + p - 2p\pi(a_1 \mid s_1)}, 1 \right]^\top.$$

In the case $p = 3/4$ we have

$$\mu^\star(\pi) = \nu^\pi = \frac{1}{1 + \frac{3/4 - \pi(a_2 \mid s_2)/2}{3/4 - \pi(a_1 \mid s_1)/2}} \left[\frac{3/4 - \pi(a_2 \mid s_2)/2}{3/4 - \pi(a_1 \mid s_1)/2}, 1 \right]^\top.$$

The fact that $\bar{\pi}_1$, $\bar{\pi}_2$, and any policy inducing $[1/2, 1/2]^\top$ as the mean field can be easily verified at this point.

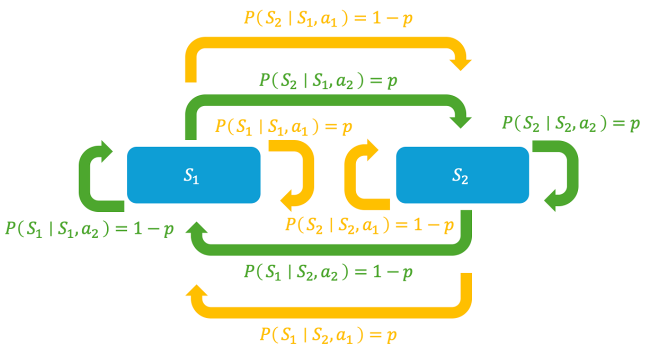

Figure 5: Example Mean Field Game Transition

G Average-Reward MDP – Detailed Formulation and Algorithm

Consider a standard average-reward MDP characterized by state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel $\mathcal{P}:\mathcal{S}\times\mathcal{A}\rightarrow\Delta_{\mathcal{S}}$, and reward function $r:\mathcal{S}\times\mathcal{A}\rightarrow[0,1]$. The cumulative reward collected by a policy $\pi:\mathcal{S}\rightarrow\Delta_{\mathcal{A}}$ is denoted by $J_{\text{MDP}}(\pi)$

$$J_{\text{MDP}}(\pi)\triangleq\lim_{T\rightarrow\infty}\tfrac{1}{T}\mathbb{E}_{a_t\sim\pi(\cdot|s_t),s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}[\textstyle\sum_{t=0}^{T-1}r(s_t,a_t)\mid s_0]. \tag{76}$$

The policy optimization objective under softmax parameterization is

$$\max_{\theta}\quad J_{\text{MDP}}(\pi_\theta). \tag{77}$$

The differential value function under policy π_θ is

$$V_{\text{MDP}}^{\pi_\theta}(s)=\mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t),s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t)}\Big[\sum_{t=0}^{\infty}\big(r(s_t,a_t)-J_{\text{MDP}}(\pi)\big)\mid s_0=s\Big].$$

We use P^π and ν^π to denote the transition probability matrix and the stationary distribution of states under the control of π. The policy gradient is

$$\nabla_\theta J_{\text{MDP}}(\pi_\theta)=\mathbb{E}_{s\sim\nu^{\pi_\theta},a\sim\pi_\theta(\cdot|s),s'\sim\mathcal{P}(\cdot|s,a)}\Big[(r(s,a)+V_{\text{MDP}}^{\pi_\theta}(s')-V_{\text{MDP}}^{\pi_\theta}(s))\nabla_\theta\log\pi_\theta(a\mid s)\Big], \tag{78}$$

and V_{MDP}^{π} satisfies the Bellman equation

$$V_{\text{MDP}}^{\pi_\theta}=\sum_{a}\pi_\theta(a\mid\cdot)r(\cdot,a)+J_{\text{MDP}}(\pi_\theta)\mathbf{1}_{|\mathcal{S}|}+(P^{\pi_\theta})^\top V_{\text{MDP}}^{\pi_\theta}. \tag{79}$$

The algorithm for optimizing J_{MDP} in an average-reward MDP, simplified from Algorithm 1, is presented in Algorithm 2. We have three main iterates in the algorithm, namely, policy parameter θ_k and value function estimates $\hat{V}_k$ and $\hat{V}_k$ which are used to track $V_{\text{MDP}}^{\pi_{\theta_k}}$ and $J_{\text{MDP}}(\pi_{\theta_k})$. The policy parameter is updated along the direction of an approximated policy gradient, while the value functions are updated to solve (79) and (77) using stochastic approximation.

H Simulation Details

We choose the reward function to be

$$r(s,a,\mu)=\mu(s)+\omega_r(s,a)*0.01,\quad\forall s,a,$$

where $\omega_r(s,a)\in\mathbb{R}$ is sampled from the standard normal distribution.

Algorithm 2 Online Actor Critic Algorithm for Average-Reward MDP

```
**Initialize:** policy parameter $\theta_0 \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, value function estimate $\hat{V}_0 \in \mathbb{R}^{|\mathcal{S}|}$, $\hat{J}_0 \in \mathbb{R}$, gradient/operator estimates $f_0 = 0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, g_0^V = 0 \in \mathbb{R}^{|\mathcal{S}|}, g_0^J = 0 \in \mathbb{R}$
**Sample:** initial state $s_0 \in \mathcal{S}$ randomly
**for** iteration $k = 0, 1, 2, ...$ **do**
    Take action $a_k \sim \pi_{\theta_k}(\cdot \mid s_k)$. Observe reward $r(s_k, a_k)$ and next state $s_{k+1} \sim \mathcal{P}(\cdot \mid s_k, a_k)$
    Policy (actor) update:
        $\theta_{k+1} = \theta_k + \alpha_k f_k.$
    Value function (critic) update:
        $\hat{V}_{k+1} = \Pi_{B_V}(\hat{V}_k + \beta_k g_k^V), \quad \hat{J}_{k+1} = \Pi_{[0,1]}(\hat{J}_k + \beta_k g_k^J).$
    Gradient/Operator estimate update:
        $f_{k+1} = (1-\lambda_k) f_k + \lambda_k (r(s_k, a_k) + \hat{V}_k(s_{k+1})) \nabla \log \pi_{\theta_k}(a_k \mid s_k),$
        $g_{k+1}^V = (1-\lambda_k) g_k^V + \lambda_k (r(s_k, a_k) - \hat{J}_k + \hat{V}_k(s_{k+1}) - \hat{V}_k(s_k)) e_{s_k}$
        $g_{k+1}^J = (1-\lambda_k) g_k^J + \lambda_k c_J (r(s_k, a_k) - \hat{J}_k).$
**end for**
```

For Environments 1 and 2, the transition kernel $\mathcal{P}$ is randomly generated such that for all s, a

$$\mathcal{P}^{\mu}(\cdot \mid s, a) \propto \omega_P(s, a),$$

where $\omega_P(s, a) \in \mathbb{R}^{|\mathcal{S}|}$ is drawn element-wise i.i.d. from the standard uniform distribution.

For Environment 3, the transition kernel $\mathcal{P}$ is also randomly generated such that for all s, a

$$\mathcal{P}^{\mu}(\cdot \mid s, a) \propto \omega_P(s, a) + \mu,$$

where $\omega_P(s, a) \in \mathbb{R}^{|\mathcal{S}|}$ is drawn element-wise i.i.d. from the standard uniform distribution.

For the proposed algorithm, we select the initial step size parameters to be $\alpha_0 = 10$, $\beta_0 = 0.1$, $\xi_0 = 0.02$, and $\lambda_0 = 1$. The step size parameters for the algorithm in Zaman et al. [2023] are taken from the paper in the Numerical Results section. We tried to adjust the parameters of their algorithm in an attempt to see whether we can get it to converge faster, and found out that the parameters prescribed in the paper are good enough and hard to improve at least locally.