# A Policy Optimization Approach to the Solution of Unregularized Mean Field Games

**Sihan Zeng** [1] **Sujay Bhatt** [1] **Alec Koppel** [1] **Sumitra Ganesh** [1]

## Abstract

We study the problem of finding the equilibrium of a mean field game (MFG) – a policy performing optimally in a Markov decision process (MDP) determined by the mean field, which is a distribution over a population of agents and a function of the policy. Prior solution techniques build upon fixed-point iteration and are only guaranteed to solve a *regularized* approximation of the problem, with a regularization constant large enough to ensure that the equilibrium is the unique fixed point of a contraction mapping. This leads to a regularized solution that can deviate arbitrarily from the original equilibrium. In this work, for the first time, we demonstrate how direct gradient-based policy optimization instead of fixed-point iteration, may solve the original, *unregularized* infinite-horizon average-reward MFG. In particular, we propose Accelerated Single-loop Actor Critic Algorithm for Mean Field Games (`ASAC-MFG`), which by its namesake, is completely data-driven, single-loop, and single-sample-path. We characterize the finite-time and finite-sample convergence of the `ASAC-MFG` algorithm to a mean field equilibrium building on a novel multi-time-scale analysis without regularization. We support the theoretical results with numerical simulations that illustrate the superior convergence of the proposed algorithm.

## 1. Introduction

The mean field game (MFG) framework, introduced in Huang et al. (2006); Lasry & 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 (Xu et al., 2018; 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 describes the notion of solution in an MFG, and is a pair of policy and mean field such that the policy performs optimally in a Markov decision process (MDP) determined by the mean field and 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 a mean field equilibrium using samples of the state transition and reward. A series of recent works have proposed finite-time convergent RL solutions to MFGs (Guo et al., 2019; Xie et al., 2021; Anahtarci et al., 2023; Mao et al., 2022; Zaman et al., 2023; Yardim et al., 2023), which all make an assumption on the contraction of a mean field optimality-consistency operator. The assumption guarantees the uniqueness of the mean field equilibrium and allows fixed-point-iteration-type algorithms to converge. However, as pointed out in Yardim et al. (2024), the assumption only holds if 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.

We summarize our main contributions and include a detailed literature comparison in Table 1.

### Main Contributions

- We design a finite-time convergent algorithm `ASAC-MFG` that provably finds a mean field equilibrium *without regularization* or imposing the aforementioned contraction assumption. However, it is shown in Yardim et al. (2024) that finding an equilibrium in a general MFGs (even with Lipschitz transition kernel and reward function) is a PPAD-complete problem conjectured to be computationally intractable (Daskalakis et al., 2009). We identify a subclass of MFGs satisfying a proposed "herding condition" (Assumption 4)

[1]JPMorgan AI Research, United States. Correspondence to: Sihan Zeng <sihan.zeng@jpmchase.com>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

with $\Delta = 0$, where ASAC−MFG converges to the exact mean field equilibrium. For MFG instances not within this subclass, our algorithm converges to a $\Delta-$neighborhood around the mean field equilibrium. In this sense, this work complements and expands on the finding of Yardim et al. (2024).

• ASAC−MFG is single-loop, single-sample path policy optimization algorithm that finds the equilibrium in the tabular infinite-horizon average-reward MFG, and has finite-time complexity. Specifically, for a subclass of MFGs satisfying the herding condition with $\Delta = 0$, ASAC−MFG finds a global mean field equilibrium with a convergence rate of $\widetilde{\mathcal{O}}(k^{-1/2})$; for $\Delta > 0$, it converges to a $\sqrt{\Delta}-$approximate MFE with the same rate. To our knowledge, this work is the first to study a finite-time convergent algorithm for MFGs without the contraction assumption, and is also the first to propose a completely sample-based single-loop single-sample-path algorithm for MFGs. Single-loop single-sample-path RL algorithms are widely used in practice due to convenience and simplicity but their theoretical understanding is not as complete as their nested-loop counterparts. Our work fills in this important gap in the context of MFGs.

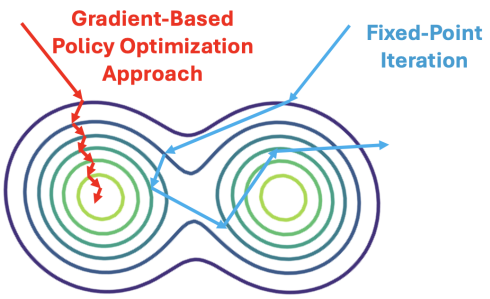


*Figure 1.* Possible trajectories of gradient-based versus fixed-point iteration methods in the landscape of MFG cumulative return without the contraction assumption. Fixed-point iteration may diverge on the unregularized problem, while gradient-based method converges.

• Our proof is based on a novel multi-time-scale analysis. We extend the techniques of analyzing two-time-scale actor-critic algorithms (Wu et al., 2020; Chen & Zhao, 2024) to the three-time-scale case where the additional time scale is introduced to carry out the mean field updates. The additional time scale may prevent the selection of the most suitable step sizes and result in convergence rate degradation if not treated properly[1]. We overcome the challenge by incorporating the latest innovation in convergence acceleration through smoothed gradient estimators (Zeng & Doan, 2024). Our multi-time-scale algorithm design methodology and analysis can be of independent interest and potentially applicable to other problems where the goal is to solve a coupled system of optimization problems.

[1]The restriction in step size selection when moving from a single time scale to two time scales is discussed in Zeng et al. (2024).

### 1.1. 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 & 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) build upon policy optimization and Anahtarcı et al. (2020); Angiuli et al. (2022; 2023); Gu et al. (2023); An et al. (2024) 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.

The aim of our paper is to design a finite-time convergent algorithm for finding the equilibrium of an MFG. Compared to the literature on this subject (Guo et al., 2019; Xie et al., 2021; Anahtarci et al., 2023; Mao et al., 2022; Zaman et al., 2023; Yardim et al., 2023), we base our algorithm on gradient-based policy optimization instead of fixed-point iteration, which allows us to remove the contraction assumption on a mean field optimality-consistency operator. Without the assumption, algorithms designed in the existing works, which leverage fixed-point iteration at the core, lose convergence/stability guarantees and may in theory exhibit arbitrary behaviors even when close to an equilibrium, as illustrated in Figure. 1. In contrast, a gradient-based algorithm can move more stably in the optimization landscape of the MFG objective due to the Lipschitz continuity.

It is worth pointing out the relevant works (Carmona et al., 2019; Fu et al., 2020; Zaman et al., 2020; Wang, 2024; Zaman et al., 2024) on linear-quadrtic MFGs (i.e. the state and action are continuous, the cost is a quadratic function of state and action, and the state transition is linear), which can be regarded as an extension of the single-agent linear-quadratic regulator. The linear-quadratic structure makes this class of problems more convenient to study and efficient to solve.

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.

The rest of the paper is organized as follows. Sec.2

presents the MFG formulation. Sec.3 develops the proposed ASAC-MFG algorithm. In Sec.4 we introduce the technical assumptions and state our main theoretical results. Simulation results are presented in Sec.5.

## 2. Formulation

We study MFGs in the *stationary infinite-horizon average-reward* setting, in which we denote the *finite* state and action spaces by $\mathcal{S}$ and $\mathcal{A}$. 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. Mathematically, we describe this aggregate behavior by the mean field $\mu \in \Delta_{\mathcal{S}}$[2], which conceptually 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 $\mathcal{P}(s' \mid s, a, \mu)$ describes the probability that the state of the representative agent transitions from $s$ to $s'$ when it takes action $a$ and mean field is $\mu$. 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 $\mu$. Not directly observing the mean field, the agent takes actions according to policy $\pi : \mathcal{S} \to \mathcal{A}$, which is represented as $\Delta_{\mathcal{A}}^{\mathcal{S}} \subset \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$.

Under a given policy $\pi$ and mean field $\mu$, the states sequentially generated form a Markov chain with transition matrix $P^{\pi,\mu} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$, where $P^{\pi,\mu}_{s',s} = \sum_{a\in\mathcal{A}} \mathcal{P}(s' \mid s, a, \mu)\pi(a \mid s)$. 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 $\mu$ and the agent generates actions according to $\pi$, the agent can expect to collect the cumulative reward $J(\pi, \mu)$

$$
\begin{aligned}
&J(\pi, \mu) \\
&\triangleq \lim_{K\to\infty} \frac{1}{K} \mathbb{E}_{a_k\sim\pi(\cdot|s_k), s_{k+1}\sim\mathcal{P}(\cdot|s_k,a_k,\mu)}[\textstyle\sum_{k=0}^{K-1} r(s_k, a_k, \mu)] \\
&= \mathbb{E}_{s\sim\nu^{\pi,\mu},\, a\sim\pi(\cdot|s)}[r(s, a, \mu)]. \qquad (1)
\end{aligned}
$$

As $J$ is independent of the initial state $s_0$, we use the differential value function $V^{\pi,\mu} \in \mathbb{R}^{|\mathcal{S}|}$ to quantify the relative value of each initial state

$$
\begin{aligned}
V^{\pi,\mu}(s) \triangleq\ & \mathbb{E}_{a_k\sim\pi(\cdot|s_k), s_{k+1}\sim\mathcal{P}(\cdot|s_k,a_k,\mu)} \\
& \Big[\textstyle\sum_{k=0}^{\infty} \big(r(s_k, a_k, \mu) - J(\pi, \mu)\big) \mid s_0 = s\Big].
\end{aligned}
$$

If the mean field were fixed to a given $\mu$, the goal of the agent would be to find a policy $\pi$ 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 $\pi$. 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 $\pi$. The following *consistency equation* needs to be satisfied by $\mu^\star(\pi)$

$$
\mu^\star(\pi) = \nu^{\pi,\mu^\star(\pi)} = (P^{\pi,\mu^\star(\pi)})^\top \mu^\star(\pi). \qquad (2)
$$

The goal of the representative agent in an MFG is to find a policy optimal under the mean field induced by the policy. Mathematically, the objective is to find a pair of policy and mean field $(\bar{\pi}, \bar{\mu})$, known to exist under mild regularity assumptions (Saldi et al., 2018), as the solution to the system

$$
\begin{cases}
J(\bar{\pi}, \bar{\mu}) \geq J(\pi, \bar{\mu}), \quad \forall\pi & (3) \\
\bar{\mu} = \mu^\star(\bar{\pi}). & (4)
\end{cases}
$$

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

**Definition 1.** *The pair of policy and mean field $(\pi, \mu)$ is an $\epsilon$-mean field equilibrium if*

$$
J(\pi', \mu) - J(\pi, \mu) \leq \epsilon, \forall\pi', \text{ and } \|\mu - \mu^\star(\pi)\| \leq \epsilon. \quad (5)
$$

We usually cannot hope to find the exact equilibrium. Definition 1 quantifies the distance between an exact equilibrium and any solution pair $(\pi, \mu)$ that we may find in finite time. It says that $(\pi, \mu)$ is an approximate mean field equilibrium if $\pi$ approximately optimizes the cumulative return in the MDP determined by $\mu$ and $\mu$ is close to the mean field induced by policy $\pi$. If a given solution $(\pi, \mu)$ satisfies (5) with $\epsilon = 0$, it is obviously an exact mean field equilibrium as a solution to (3)-(4).

## 3. Algorithm

Our algorithm departs from the existing literature in that we approach MFGs from the perspective of direct policy optimization rather than fixed-point iteration. As we do not directly deal with the mean field optimality-consistency operator, we bypass the need to impose strong and unrealistic assumptions. It is obvious from (3) 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 $\bar{\mu}$ and that the optimal policy under $\bar{\mu}$ may not be unique. However, inspired by the discussion above we take the approach of simultaneous learning. We maintain a parameter $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ that encodes the policy $\pi_\theta$ via the softmax function i.e. $\pi_\theta(a \mid s) = \exp(\theta(s, a))/\sum_{a'\in\mathcal{A}} \exp(\theta(s, a'))$, and a mean field iterate $\hat{\mu}$ to estimate the mean field induced by

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

| | *Contraction Assumption* | *Single Sample Path* | *Single Loop* | *Convergence Rate* |
|---|---|---|---|---|
| *Guo et al. (2019)* | *Required* | *No* | *No* | *Regularization Dependent* |
| *Anahtarcı et al. (2020)* | *Required* | *No* | *No* | - |
| *Xie et al. (2021)* | *Required* | *Yes** | *Yes** | $\widetilde{\mathcal{O}}(k^{-1/5})$§, *regularized solution* |
| *Mao et al. (2022)* | *Required* | *No* | *No* | $\widetilde{\mathcal{O}}(k^{-1/5})$§, *regularized solution* |
| *Zaman et al. (2023)* | *Required* | *No* | *Yes* | $\widetilde{\mathcal{O}}(k^{-1/4})$§, *regularized solution* |
| *Yardim et al. (2023)* | *Required* | *No* | *No* | $\widetilde{\mathcal{O}}(k^{-1/2})$§, *regularized solution* |
| ***Our Work (on MFG subclass $\Delta = 0^{\dagger}$)*** | ***No*** | ***Yes*** | ***Yes*** | $\widetilde{\mathcal{O}}(k^{-1/4})$***, original solution*** |
| ***Our Work (on other MFG instances $^{\dagger}$)*** | ***No*** | ***Yes*** | ***Yes*** | $\widetilde{\mathcal{O}}(k^{-1/4})$***, $\sqrt{\Delta}$-optimal solution*** |

*Table 1. Existing algorithms and their assumption, structure, and complexity. * 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 $\pi, \mu$. Mao et al. (2022) also relies on such an oracle. Our work, in comparison, is oracle-free. $^{\dagger}$ We introduce $\Delta$ to characterize the difficulty of a MFG in Assumption 4. § If these works could choose the regularization weight freely (note that they actually cannot since the contraction operator assumption only holds when the weight is sufficiently large), the algorithms can be used to solve the original unregularized game by making the weight small enough. The complexities, however, at least double, i.e. become $\widetilde{\mathcal{O}}(k^{-1/10})$, $\widetilde{\mathcal{O}}(k^{-1/8})$, $\widetilde{\mathcal{O}}(k^{-1/4})$ to the original solution.*

the current policy. We improve $\theta$ and $\hat{\mu}$ with respect to each other by iteratively taking the steps below

$$\theta_{k+1} = \theta_k + \alpha_k \nabla_\theta J(\pi_{\theta_k}, \hat{\mu}_k), \quad \hat{\mu}_{k+1} = \mu^\star(\pi_{\theta_k}) \quad (6)$$

where $k$ is the iteration index and $\alpha_k$ is a properly selected step sizes.

By the policy gradient theorem (Sutton et al., 1999), a closed-form expression for $\nabla_\theta J(\pi_\theta, \mu)$ is

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

In large and/or unknown environments in the real life, performing (6) poses computational challenges. The updates require the knowledge of $\mu^\star(\pi_{\theta_k})$ and value function $V^{\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})}$, neither of which can be exactly determined instantaneously. We propose learning $\mu^\star(\pi_{\theta_k})$ and $V^{\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})}$ simultaneously with the policy update using the same path of samples. We recognize that

$$\mu^\star(\pi_{\theta_k}) = \lim_{T\to\infty} \tfrac{1}{T}\mathbb{E}_{a_t\sim\pi_{\theta_k}(\cdot|s_t), s_{t+1}\sim\mathcal{P}(\cdot|s_t,a_t,\mu^\star(\pi_{\theta_k}))}[e_{s_k}], \quad (7)$$

where $e_s \in \mathbb{R}^{|\mathcal{S}|}$ is the indicator vector whose entry $s'$ is 1 if $s' = s$ and 0 otherwise. Solving Eq. (7) with multi-time-scale stochastic approximation, we iteratively perform

$$\hat{\mu}_{k+1} = \hat{\mu}_k + \xi_k(e_{s_k} - \hat{\mu}_k) \quad (8)$$

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.

It is well-known that $V^{\pi_{\theta_k}, \hat{\mu}_k}$ satisfies the Bellman equation

$$V^{\pi_{\theta_k}, \hat{\mu}_k} = \textstyle\sum_a \pi_{\theta_k}(a \mid \cdot) r(\cdot, a, \hat{\mu}_k) + J(\pi_{\theta_k}, \hat{\mu}_k)\mathbf{1}_{|\mathcal{S}|} + (P^{\pi_{\theta_k}, \hat{\mu}_k})^\top V^{\pi_{\theta_k}, \hat{\mu}_k}. \quad (9)$$

Here $\mathbf{1}_{|\mathcal{S}|}$ denotes the all-one vector of length $|\mathcal{S}|$. We introduce an auxiliary variable $\hat{V}$ to track $V^{\pi_{\theta_k}, \hat{\mu}_k}$ also by stochastic approximation. The following update solves (9)

$$\hat{V}_{k+1}(s_k) = \hat{V}_k(s_k) + \beta_k\big(r(s_k, a_k, \hat{\mu}_k) - \hat{J}_k + \hat{V}_k(s_{k+1}) - \hat{V}_k(s_k)\big), \quad (10)$$

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

$$\hat{J}_{k+1} = \hat{J}_k + \beta_k(r(s_k, a_k, \hat{\mu}_k) - \hat{J}_k). \quad (11)$$

Here we make the step size $\beta_k$ much larger than $\xi_k$ for $\hat{V}_{k+1}$ and $\hat{J}_{k+1}$ to chase the targets $V^{\pi_{\theta_k}, \hat{\mu}_k}$ and $J(\pi_{\theta_k}, \hat{\mu}_k)$ which evolve with the step size $\xi_k$.

Combining Eqs. (8), (10), and (11) with the $\theta$ update in (6) results in a single-loop single-sample-path algorithm where in the slowest time scale we ascend the policy parameter $\theta_k$ along the gradient direction and the fast time scales are used to compute the quantities necessary for the gradient evaluation. While such an algorithm can be shown to converge to a mean field equilibrium (under proper assumptions), the convergence does not occur at the best possible rate due to the coupling between iterates – $\theta_k$, $\hat{\mu}_k$, $\hat{V}_k$, and $\hat{J}_k$ – directly affect each other's update, causing potential noise in any variable to be immediately propagated to others. Zeng & Doan (2024) details the degradation in algorithm complexity resulting from such coupling effect when two variables are simultaneously updated. In this work we need to deal with three time scales ($\alpha_k$, $\beta_k$, $\xi_k$), which makes coupling worse. To alleviate the issue, Zeng & Doan (2024) proposes an improved algorithm that accelerates convergence by introducing a denoising step. We adopt this technique and extend it to handle the three-time-scale updates. The idea behind the acceleration 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 (15).

In (14), $\Pi_{B_V} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ denotes the projection to the $\ell_2$-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 (14) and are a frequently used tool in the analysis of actor-critic algorithms in the literature (Wu et al., 2020; Chen & Zhao, 2024; Panda & Bhatnagar, 2024).

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

1: **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 $f_0 = 0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, g_0^V = 0 \in \mathbb{R}^{|\mathcal{S}|}, g_0^J = 0 \in \mathbb{R}, h_0 = 0 \in \mathbb{R}^{|\mathcal{S}|}$

2: **Sample:** initial state $s_0 \in \mathcal{S}$ randomly

3: **for** iteration $k = 0, 1, 2, ...$ **do**

4: Take action $a_k \sim \pi_{\theta_k}(\cdot \mid s_k)$. Observe $r(s_k, a_k, \hat{\mu}_k)$ and $s_{k+1} \sim \mathcal{P}(\cdot \mid s_k, a_k, \hat{\mu}_k)$

5: Policy (actor) update:

$$\theta_{k+1} = \theta_k + \alpha_k f_k. \tag{12}$$

6: Mean field update:

$$\hat{\mu}_{k+1} = \hat{\mu}_k + \xi_k h_k. \tag{13}$$

7: Value function (critic) update:

$$\begin{aligned} \hat{V}_{k+1} &= \Pi_{B_V}(\hat{V}_k + \beta_k g_k^V), \\ \hat{J}_{k+1} &= \Pi_{[0,1]}(\hat{J}_k + \beta_k g_k^J). \end{aligned} \tag{14}$$

8: Gradient/Operator estimate update:

$$\begin{aligned} f_{k+1} &= (1-\lambda_k) f_k + \lambda_k \Big( r(s_k, a_k, \hat{\mu}_k) + \hat{V}_k(s_{k+1}) \\ &\quad - \hat{V}_k(s_k) \Big) \nabla \log \pi_{\theta_k}(a_k \mid s_k) \\ g_{k+1}^V &= (1-\lambda_k) g_k^V + \lambda_k \Big( r(s_k, a_k, \hat{\mu}_k) - \hat{J}_k \\ &\quad + \hat{V}_k(s_{k+1}) - \hat{V}_k(s_k) \Big) e_{s_k} \\ g_{k+1}^J &= (1-\lambda_k) g_k^J + \lambda_k c_J (r(s_k, a_k, \hat{\mu}_k) - \hat{J}_k) \\ h_{k+1} &= (1-\lambda_k) h_k + \lambda_k (e_{s_k} - \hat{\mu}_k) \end{aligned} \tag{15}$$

9: **end for**

## 4. Main Results

This section presents the finite-time convergence of Algorithm 1 to a mean field equilibrium. We start by introducing the technical assumptions made in this paper, most of which are standard.

**Assumption 1.** *Given any $\pi, \mu$, 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*

$$\sup_s d_{TV}\big(\mathbb{P}(s_k = \cdot \mid s_0 = s), \nu^{\pi,\mu}(\cdot)\big) \leq C_0 C_1^k, \forall k \geq 0, \tag{16}$$

*where $d_{TV}$ denotes the total variation (TV) distance*[3].

Eq. (16) states that the $k_{\text{th}}$ 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 $\pi, \mu$. This assumption is important and common among the papers that study the complexity of sample-based single-loop RL algorithms (Zou et al., 2019; Wu et al., 2020; Zeng et al., 2022; Chen & Zhao, 2024).

**Assumption 2.** *Given two distributions $d_1, d_2$ over $\mathcal{S}$, policies $\pi_1, \pi_2$, and mean fields $\mu_1, \mu_2$, 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*

$$\begin{aligned} d_{TV}(\mathbb{P}(s' = \cdot), \mathbb{P}(\hat{s}' = \cdot)) &\leq d_{TV}(d_1, d_2) \\ &\quad + L\|\pi_1 - \pi_2\| + L\|\mu_1 - \mu_2\|, \\ d_{TV}(\nu^{\pi_1,\mu_1}, \nu^{\pi_2,\mu_2}) &\leq L\|\pi_1 - \pi_2\| + L\|\mu_1 - \mu_2\|, \\ |r(s, a, \mu_1) - r(s, a, \mu_2)| &\leq L|\mu_1 - \mu_2\|, \\ \|\mu^\star(\pi_1) - \mu^\star(\pi_2)\| &\leq L\|\pi_1 - \pi_2\|. \end{aligned} \tag{18}$$

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

Eq. (18) amounts to a regularity condition on the transition probability matrix $P^{\pi,\mu}$ as a function of $\pi$ and $\mu$ and can be shown to hold if the transition kernel $\mathcal{P}(\cdot \mid \cdot, \cdot, \mu)$ is Lipschitz in $\mu$ (using an argument similar to Wu et al. (2020)[Lemma B.2]). The rest of Assumption 2 imposes the Lipschitz continuity of the stationary distribution, reward function, and induced mean field, as well as the boundedness of the differential value function. Importantly, Assumption 2 guarantees the Lipschitz continuity of the cumulative reward and differential value function, which we show in Lemma 1. All conditions in this assumption are common in the literature of MFGs and RL (Yardim et al., 2023; Anahtarci et al., 2023; Wu et al., 2020; Zeng et al., 2024).

**Assumption 3.** *There is a constant $\delta \in (0, 1)$ such that $\|\nu^{\pi,\mu_1} - \nu^{\pi,\mu_2}\| \leq \delta\|\mu_1 - \mu_2\|, \forall \pi, \mu_1, \mu_2$.*

[3] Given two probability distributions $\phi_1$ and $\phi_2$ over space $\mathcal{X}$, their TV distance is defined as

$$d_{\text{TV}}(\phi_1, \phi_2) = \frac{1}{2} \sup_{\psi:\mathcal{X}\to[-1,1]} \left| \int \psi d\phi_1 - \int \psi d\phi_2 \right|. \tag{17}$$

Assumption 3 states that for any $\pi$ the stationary distribution $\nu^{\pi,\mu}$ is a contractive mapping in $\mu$. The assumption allows us to estimate the induced mean field $\mu^\star(\pi)$ by measuring the stationary distribution of the Markov chain formed under the control of $\pi$. The validity of this assumption only depends on the transition kernel $\mathcal{P}$. To contrast, the common assumption in the existing literature on MFGs amounts to requiring the mapping $\nu^{\pi^\star(\mu),\mu}$ to be contractive (Xie et al., 2021; Zaman et al., 2023; Yardim et al., 2023), where $\pi^\star(\mu)$ is the (assumed unique) optimal policy under mean field $\mu$, i.e. $\pi^\star(\mu) = \operatorname{argmax}_\pi J(\pi,\mu)$. Specifically, they assume the existence of $\delta \in (0,1)$ such that

$$\|\nu^{\pi^\star(\mu_1),\mu_1} - \nu^{\pi^\star(\mu_2),\mu_2}\| \leq \delta\|\mu_1 - \mu_2\|. \tag{19}$$

It is pointed out in Yardim et al. (2024) that (19) is a strong assumption whose validity depends on both the transition kernel and reward function, and does not hold in MFGs unless a large regularization is added. Assumption 3 made in this paper is much milder.

The approach we take in this paper is to iteratively refine the policy parameter $\theta$ along a direction that may improve the cumulative reward under the induced mean field $\mu^\star(\pi_\theta)$. If we take $\theta' = \theta + \alpha\nabla_\theta J(\pi_\theta,\mu)\mid_{\mu=\mu^\star(\pi_\theta)}$ with a sufficiently small step size $\alpha$, we can approximately guarantee

$$J(\pi_{\theta'},\mu^\star(\pi_\theta)) \gtrsim J(\pi_\theta,\mu^\star(\pi_\theta)).$$

However, the induced mean field shifts from $\mu^\star(\pi_\theta)$ to $\mu^\star(\pi_{\theta'})$ as the policy changes. Due to the lack of strong structure on $\mu^\star(\pi_\theta)$ besides the Lipschitz condition, predicting/controlling whether $J(\pi_{\theta'},\mu^\star(\pi_{\theta'}))$ improves over $J(\pi_\theta,\mu^\star(\pi_\theta))$ is difficult. In this work, we characterize the difficulty of an MFG by the mean field shift error $\Delta$ introduced in the following assumption. A problem with a small or zero $\Delta$ is considered easier to solve. In fact, we later show in our analysis that the `ASAC-MFG` algorithm solves a MFG up to a sub-optimality gap proportional to $\Delta$.

**Assumption 4** (Herding Condition)**.** *There exists bounded constants $\rho, \Delta \geq 0$ such that $\forall \pi, \pi'$*

$$\begin{aligned} &J(\pi',\mu^\star(\pi)) - J(\pi',\mu^\star(\pi')) \\ &\leq \rho\Big(J(\pi,\mu^\star(\pi)) - J(\pi',\mu^\star(\pi))\Big) + \Delta\|\pi - \pi'\|. \end{aligned} \tag{20}$$

Conceptually, the MFGs with a small or zero $\Delta$ are those in which the reward is higher when the representative agent "follows the crowd" or displays a "herding" behavior. We discuss more on the interpretation, implication, and structure of the condition in Sec.4.2.

We denote by $\mathbb{F}(\theta)$ the Fisher information matrix at policy parameter $\theta$

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

**Assumption 5.** *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 $\forall\theta$.*

Our final assumption on Fisher non-degenerate policy implies a "gradient domination" condition – for any policy $\pi$, every stationary point of the cumulative reward $J(\pi,\mu^\star(\pi))$ is globally optimal. This is again a standard assumption in the existing literature on policy optimization (Liu et al., 2020; Fatkhullin et al., 2023; Ganesh et al., 2024).

### 4.1. Finite-Time Analysis

Each variable in Algorithm 1 has a target to chase. The target of $\theta_k$ 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 $\pi_{\theta_k}$ 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 at a sublinear rate.

$$\begin{aligned} &\varepsilon_k^\pi \triangleq \|\nabla_\theta J(\pi_{\theta_k},\mu)\mid_{\mu=\mu^\star(\pi_{\theta_k})}\|^2,\ \varepsilon_k^\mu \triangleq \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\|^2, \\ &\varepsilon_k^V \triangleq \|\Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k},\hat{\mu}_k})\|^2,\ \varepsilon_k^J \triangleq (\hat{J}_k - J(\pi_{\theta_k},\hat{\mu}_k))^2. \end{aligned} \tag{21}$$

We would like $\hat{V}_k$ to converge to $V^{\pi_{\theta_k},\hat{\mu}_k}$ which solves the Bellman equation (9). However, the solution is not unique. If $V \in \mathbb{R}^{|\mathcal{S}|}$ solves (9), 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 $\varepsilon_k^V$ in (21) 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 with the step sizes satisfying*

$$\begin{aligned} \lambda_k &= \frac{\lambda_0}{\sqrt{k+1}}, \quad \alpha_k = \frac{\alpha_0}{\sqrt{k+1}}, \\ \beta_k &= \frac{\beta_0}{\sqrt{k+1}}, \quad \xi_k = \frac{\xi_0}{\sqrt{k+1}}, \end{aligned} \tag{22}$$

*with constants $\lambda_0, \alpha_0, \beta_0, \xi_0$ and a sufficiently large $c_J$ specified in Appendix B.2. Under Assumptions 1-4, we have for all $k \geq \tau_k$*

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

*where $\tau_k$ denotes the mixing time, which is a linear 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 $\Delta$, under a single trajectory of Markovian samples. Since Algorithm 1 draws exactly 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 $\alpha_0 \leq \xi_0 \leq \beta_0 \leq \lambda_0$. Such a requirement makes intuitive sense: 1) the learning targets of $\hat{\mu}_k, \hat{V}_k, \hat{J}_k$ are depend on $\theta_k$, which requires $\theta_k$ to be relatively stable and hence updated with the smallest step size; 2) similarly, the learning target of $\hat{V}_k, \hat{J}_k$ is a function of $\hat{\mu}_k$, so $\hat{\mu}_k$ has to move slower; 3) we need the auxiliary variables $f_k, h_k, g_k^V, g_k^J$ to be updated the fastest to track the moving gradients/operators.

Our ultimate goal is to find an $\epsilon$-mean field equilibrium in the sense of Definition 1. This requires us to connect the convergence of $\varepsilon_k^\pi$ to the optimality gap below

$$\max_\pi J(\pi, \mu^\star(\pi_{\theta_k})) - J(\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})). \tag{23}$$

Under Assumption 5 a "gradient domination" condition holds, which upper bounds (23) by $\sqrt{\varepsilon_k^\pi}$. We take advantage of the gradient domination property to establish the convergence of Algorithm 1 to an approximate mean field equilibrium, as a corollary of Theorem 1.

**Corollary 1.** *Consider the policy $\pi_{\theta_k}$ generated by Algorithm 1 under any initialization with the step sizes satisfying (22). Under Assumptions 1-5, we have for all $k \geq \tau_k$*

$$\min_{\tau_k \leq t < k} \mathbb{E}\left[\max_\pi J(\pi, \mu^\star(\pi_{\theta_t})) - J(\pi_{\theta_t}, \mu^\star(\pi_{\theta_t}))\right] \leq \widetilde{\mathcal{O}}((k+1)^{-1/4}) + \mathcal{O}(\sqrt{\Delta}),$$

$$\min_{\tau_k \leq t < k} \mathbb{E}[\|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\|] \leq \widetilde{\mathcal{O}}((k+1)^{-1/4}) + \mathcal{O}(\sqrt{\Delta}).$$

Corollary 1 guarantees that Algorithm 1 finds an $(\epsilon + \mathcal{O}(\sqrt{\Delta}))$-mean field equilibrium in the sense of Definition 1 within at most $\widetilde{\mathcal{O}}(\epsilon^{-4})$ iterations. This is the first result showing that an algorithm provably (approximately) solves the MFG without regularization in finite time.

### 4.2. More On the Herding Condition

It can be shown that due to the Lipschitz continuity of $J$ and $\mu^\star$, Assumption 4 always holds in the worst case with $\rho = 0$ and $\Delta = LL_V$, where $L$ is from Assumption 2 and $L_V$ is the Lipschitz constant of $V$ and $J$ introduced in Lemma 1. However, specific MFG problems may be so structured that it satisfies (20) with a smaller $\Delta$ (or even $\Delta = 0$). The algorithm we propose solves MFGs to a precision proportional to $\Delta$, i.e., we have convergence to an exact mean field equilibrium for MFGs with $\Delta = 0$, and to a neighborhood around an equilibrium when $\Delta > 0$. In Example 1 we present a subclass of MFGs satisfying Assumption 4 but not (19), for which our algorithm finds an equilibrium but prior algorithms proposed in Xie et al. (2021); Anahtarci et al. (2023); Mao et al. (2022); Zaman et al. (2023); Yardim et al. (2023) theoretically fail.

**Example 1.** *Consider MFGs in which the transition probability kernel independent of the mean field and the reward function is $r(s, a, \mu) = \mu(s)$. This subclass of MFGs satisfies Assumption 4 with $\rho = 1$ and $\Delta = 0$, which we justify in Appendix F. However, (19) does not have to hold. Take a simple example with $|\mathcal{S}| = |\mathcal{A}| = 2$, 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 $\bar{\pi}_1$, $\bar{\pi}_2$*

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

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

*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 $[1/2, 1/2]^\top$ as the mean field (such as $\bar{\pi}_3(a \mid s) = 1/2$ for all $s, a$). The contraction assumption (19) does not hold as the equilibrium is not unique. The detailed derivation can be found in Appendix F.*

## 5. Numerical Simulations

We numerically verify the convergence of the proposed algorithm through simulations on small-scale synthetic MFGs. We consider two environments, first of dimension $|\mathcal{S}| = |\mathcal{A}| = 10$ and second $|\mathcal{S}| = |\mathcal{A}| = 20$, both of which have a randomly generated transition kernel and reward function.[4] Due to the unknown equilibria, we measure the convergence of the policy by $\|\nabla_\theta J(\pi_{\theta_k}, \hat{\mu}_k)\|$ and the convergence of the mean field 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 and enables a fair comparison. We consider two variations of their algorithm: 1) with regularization large enough that the contraction assumption holds, and 2) with regularization set to 0 which breaks the assumption. The environments do not satisfy Assumption 4 with $\Delta = 0$, so the theoretical result in Sec.4.1 guarantees the convergence of `ASAC-MFG` up to an error proportional to $\Delta$. As shown in Figure 2, 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. For the considered examples, `ASAC-MFG` and Zaman et al. (2023) with no regularization exhibit convergence to a global MFE. However, `ASAC-MFG` converges at a faster rate, which we believe can be attributed to the single-loop updates as well as the fact

[4] More discussion of the experimental setup can be found in Appendix H. The implementation code is also submitted as a part of the supplementary material.

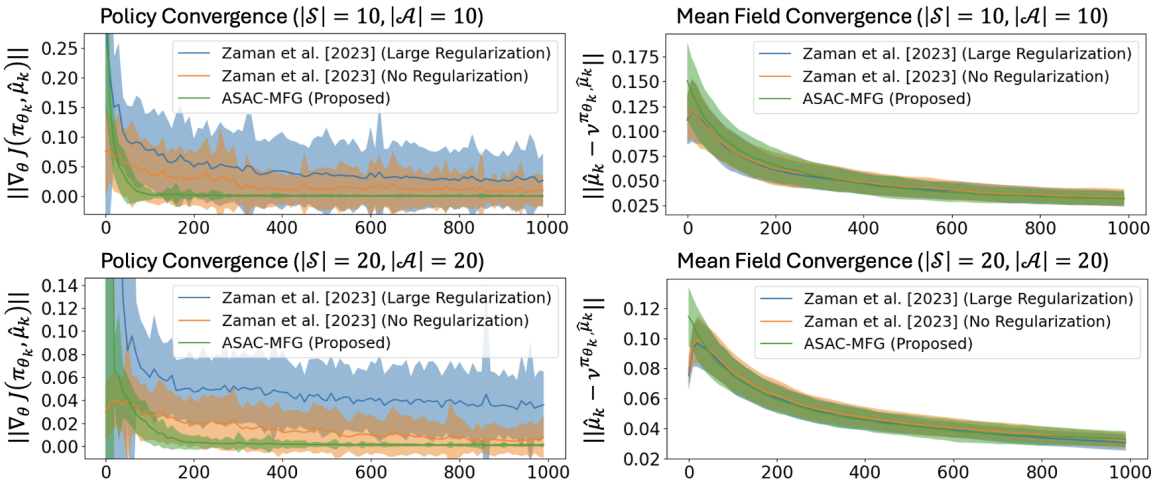


*Figure 2.* Algorithm Performance in Synthetic Games. Averaged over 100 trials. First columns shows sub-optimality gap of policy under latest mean field estimate. Second column shows convergence of mean field estimate to mean field induced by latest policy iterate. Large regularization is required for theoretical analyses by Zaman et al. (2023), which manifests in persistent bias.

that our work still enjoys convergence guarantees on this problem (though not to the exactly optimal solution) while Zaman et al. (2023) under no regularization loses any guarantee. `ASAC-MFG` is also superior in that the convergence path has a much smaller variance. The blue curve in Figure 2 shows that while Zaman et al. (2023) with sufficiently large regularization may converge to a solution of the regularized problem, the bias caused by the large regularization prevents it from finding an equilibrium of the original game.

## Disclaimer

This paper was prepared for informational purposes in part by the Artificial Intelligence Research group of JP Morgan 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

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

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

An, J., Lu, J., Wu, Y., and Xiang, Y. Why does the two-timescale q-learning converge to different mean field solutions? a unified convergence analysis. *arXiv preprint arXiv:2404.04357*, 2024.

Anahtarcı, B., Karıksız, C. D., and Saldi, N. Value iteration algorithm for mean-field games. *Systems & Control Letters*, 143:104744, 2020.

Anahtarci, B., Kariksiz, C. D., and Saldi, N. Q-learning in regularized mean-field games. *Dynamic Games and Applications*, 13(1):89–117, 2023.

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

Angiuli, A., Fouque, J.-P., Laurière, M., and Zhang, M. Convergence of multi-scale reinforcement q-learning algorithms for mean field game and control problems. *arXiv preprint arXiv:2312.06659*, 2023.

Carmona, R., Laurière, M., and Tan, Z. Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. *arXiv preprint arXiv:1910.04295*, 2019.

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

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

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

Fatkhullin, I., Barakat, A., Kireeva, A., and He, N. Stochastic policy gradient methods: Improved sample complexity for fisher-non-degenerate policies. In *International Conference on Machine Learning*, pp. 9827–9869. PMLR, 2023.

Fu, Z., Yang, Z., Chen, Y., and Wang, Z. Actor-critic provably finds nash equilibria of linear-quadratic mean-field games. In *International Conference on Learning Representations*, 2020.

Ganesh, S., Mondal, W. U., and Aggarwal, V. Variance-reduced policy gradient approaches for infinite horizon average reward markov decision processes. *arXiv preprint arXiv:2404.02108*, 2024.

Gu, H., Guo, X., Wei, X., and Xu, R. Dynamic programming principles for mean-field controls with learning. *Operations Research*, 71(4):1040–1054, 2023.

Guo, X., Hu, A., Xu, R., and Zhang, J. Learning mean-field games. *Advances in neural information processing systems*, 32, 2019.

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

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

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

Jiang, Y., Hu, Y., Bennis, M., Zheng, F.-C., and You, X. A mean field game-based distributed edge caching in fog radio access networks. *IEEE Transactions on Communications*, 68(3):1567–1580, 2019.

Kumar, N., Murthy, Y., Shufaro, I., Levy, K. Y., Srikant, R., and Mannor, S. On the global convergence of policy gradient in average reward markov decision processes. *arXiv preprint arXiv:2403.06806*, 2024.

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

Li, L., Cheng, Q., Tang, X., Bai, T., Chen, W., Ding, Z., and Han, Z. 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.

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

Mandal, D., Triantafyllou, S., and Radanovic, G. Performative reinforcement learning. In *International Conference on Machine Learning*, pp. 23642–23680. PMLR, 2023.

Mao, W., Qiu, H., Wang, C., Franke, H., Kalbarczyk, Z., Iyer, R., and Basar, T. A mean-field game approach to cloud resource management with function approximation. *Advances in Neural Information Processing Systems*, 35: 36243–36258, 2022.

Narasimha, D., Shakkottai, S., and Ying, L. 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*, pp. 1–10, 2019.

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

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

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

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

Wang, B.-C. Leader–follower mean field lq games: A direct method. *Asian Journal of Control*, 26(2):617–625, 2024.

Wu, Y. F., Zhang, W., Xu, P., and Gu, Q. A finite-time analysis of two time-scale actor-critic methods. *Advances*

*in Neural Information Processing Systems*, 33:17617–17628, 2020.

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

Xu, Y., Li, L., Zhang, Z., Xue, K., and Han, Z. A discrete-time mean field game in multi-uav wireless communication systems. In *2018 IEEE/CIC International Conference on Communications in China (ICCC)*, pp. 714–718. IEEE, 2018.

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

Yardim, B., Cayci, S., Geist, M., and He, N. Policy mirror ascent for efficient and independent learning in mean field games. In *International Conference on Machine Learning*, pp. 39722–39754. PMLR, 2023.

Yardim, B., Goldman, A., and He, N. When is mean-field reinforcement learning tractable and relevant? *arXiv preprint arXiv:2402.05757*, 2024.

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

Zaman, M. A. u., Koppel, A., Bhatt, S., and Basar, T. Oracle-free reinforcement learning in mean-field games along a single sample path. In *International Conference on Artificial Intelligence and Statistics*, pp. 10178–10206. PMLR, 2023.

Zaman, M. A. u., Koppel, A., Laurière, M., and Başar, T. Independent rl for cooperative-competitive agents: A mean-field perspective. *arXiv preprint arXiv:2403.11345*, 2024.

Zeng, S. and Doan, T. T. Fast two-time-scale stochastic gradient method with applications in reinforcement learning. *arXiv preprint arXiv:2405.09660*, 2024.

Zeng, S., Doan, T. T., and Romberg, J. 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)*, pp. 4028–4033. IEEE, 2022.

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

Zhang, S., Zhang, Z., and Maguluri, S. T. Finite sample analysis of average-reward td learning and $q$-learning. *Advances in Neural Information Processing Systems*, 34: 1230–1242, 2021a.

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

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

## Contents

## A. Notations and Frequently Used Identities

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

$$\begin{aligned}
F(\theta, V, \mu, s, a, s') &\triangleq (r(s,a,\mu) + V(s') - V(s))\nabla_\theta \log \pi_\theta(a \mid s), \\
G^V(V, J, \mu, s, a, s') &\triangleq (r(s,a,\mu) - J + V(s') - V(s))e_s, \\
G^J(J, \mu, s, a) &\triangleq c_J(r(s,a,\mu) - J), \\
G(V, J, \mu, s, a, s') &\triangleq \begin{bmatrix} G^V(V, J, \mu, s, a, s') \\ G^J(J, \mu, s, a) \end{bmatrix} = \begin{bmatrix} (r(s,a,\mu) - J + V(s') - V(s))e_s \\ c_J(r(s,a,\mu) - J) \end{bmatrix}, \\
H(\mu, s) &\triangleq e_s - \mu.
\end{aligned} \tag{24}$$

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

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

We also define

$$\begin{aligned}
\bar{F}(\theta, V, \mu) &\triangleq \mathbb{E}_{s\sim\nu^{\pi_\theta,\mu}, a\sim\pi_\theta(\cdot|s), s'\sim\mathcal{P}(\cdot|s,a,\mu)}[F(\theta, V, \mu, s, a, s')], \\
\bar{G}^V(\theta, V, J, \mu) &\triangleq \mathbb{E}_{s\sim\nu^{\pi_\theta,\mu}, a\sim\pi_\theta(\cdot|s), s'\sim\mathcal{P}(\cdot|s,a,\mu)}[G^V(V, J, \mu, s, a, s')], \\
\bar{G}^J(\theta, J, \mu) &\triangleq \mathbb{E}_{s\sim\nu^{\pi_\theta,\mu}, a\sim\pi_\theta(\cdot|s)}[G^J(J, \mu, s, a)], \\
\bar{G}(\theta, V, J, \mu) &\triangleq \mathbb{E}_{s\sim\nu^{\pi_\theta,\mu}, a\sim\pi_\theta(\cdot|s), s'\sim\mathcal{P}(\cdot|s,a,\mu)}[G(V, J, \mu, s, a, s')] = \begin{bmatrix} \bar{G}^V(\theta, V, J, \mu) \\ \bar{G}^J(\theta, J, \mu) \end{bmatrix}, \\
\bar{H}(\theta, \mu) &\triangleq \mathbb{E}_{s\sim\nu^{\pi_\theta,\mu}}[H(\mu, s)] = \mathbb{E}_{s\sim\nu^{\pi_\theta,\mu}}[e_s - \mu].
\end{aligned} \tag{25}$$

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

$$\Delta g_k = \begin{bmatrix} \Delta g_k^V \\ \Delta g_k^J \end{bmatrix} = g_k - \bar{G}(\theta_k, \hat{V}_k, \hat{J}_k, \hat{\mu}_k).$$

We use $\ell(\pi)$ to denote the cumulative reward collected by policy $\pi$ 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 $k$.

We use the notation $\mathcal{P}_\mu(s' \mid s, a) = \mathcal{P}(s' \mid s, a, \mu)$. Under Assumptions 1 and 2, 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 $\pi_1, \pi_2, \mu_1, \mu_2$

$$d_{TV}(\nu^{\pi_1,\mu_1} \otimes \pi_1 \otimes \mathcal{P}_{\mu_1}, \nu^{\pi_2,\mu_2} \otimes \pi_2 \otimes \mathcal{P}_{\mu_2}) \leq L_{TV}(\|\pi_1 - \pi_2\| + \|\mu_1 - \mu_2\|). \tag{26}$$

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 2.** *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 $\alpha_k$, where $\alpha_k$ 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 $\theta$ and $\mu$, as shown in the lemma below.

**Lemma 1.** *Under Assumption 2, there exist a bounded constant $L_V \geq 1$ such that for any policy parameter $\theta_1, \theta_2$ and mean field $\mu_1, \mu_2$, 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 these 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 $J_1, J_2 \in \mathbb{R}$*

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

*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 results that bound the energy of the auxiliary variables $f_k$, $g_k$, and $h_k$.

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

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

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 & 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 $\theta, \mu$ and $V \in \mathcal{E}_\perp$*

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

## B. Proof of Main Theorem

### B.1. Intermediate Results

The proof of Theorem 1 relies critically on the iteration-wise convergence of policy iterate $\theta_k$, 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 1-2, we have*

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

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

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

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

$$\mathbb{E}[\|\Delta h_{k+1}\|^2]$$

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

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 1-2,*

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

**Proposition 6.** *Under Assumptions 1-2, we have for all $k \geq \tau_k$*

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

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

$$\alpha_0 \leq \min\left\{\frac{1}{192(L_F^2 + L_G^2 + L_H^2 + L_V^2 + L^2/(1-\delta) + \rho + 1)}\lambda_0,\ C_\beta\beta_0,\ C_\xi\xi_0\right\},$$
$$\xi_0 \leq \min\left\{\frac{\lambda_0}{64(L_H^2L_F^2 + L_G^2 + L_V^2/\gamma + 1/(1-\delta))},\ \frac{(1-\delta)\gamma\beta_0}{6912(L_F^4L_V^2 + L_F^2L_G^2L_H^2L_V^2 + L_F^2L_H^4L_V^2 + L_V^2)}\right\}, \tag{27}$$
$$\beta_0 \leq \min\left\{\frac{\lambda_0}{72|\mathcal{S}|L_G^2 + 36L_F^2 + 8/\gamma},\ \frac{\gamma}{4L_G^2},\ \frac{1-\delta}{2L_H^2}\right\}, \quad \lambda_0 \leq \frac{1}{4},$$

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

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

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 $k \geq \tau_k$

$\mathcal{L}_{k+1}$

$$
\begin{aligned}
&= \mathbb{E}[\|\Delta f_{k+1}\|^2 + \|\Delta g_{k+1}\|^2 + \|\Delta h_{k+1}\|^2 - \ell(\pi_{\theta_{k+1}}) + \varepsilon^V_{k+1} + \varepsilon^J_{k+1} + \varepsilon^\mu_{k+1}] \\
&\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{(1+\rho)\alpha_k}{2}\mathbb{E}[\varepsilon_k^\pi] + (1+\rho)\alpha_k\mathbb{E}[\|\Delta f_k\|^2] \\
&\quad + (1+\rho)L_F^2\alpha_k\mathbb{E}[\varepsilon_k^V + \varepsilon_k^\mu] + \frac{(1+\rho)L_VB_F^2\alpha_k^2}{2} + B_F\alpha_k\Delta \\
&\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 \\
&\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{(1+\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\Delta \\
&\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{(1+\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 \\
&\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} + (1+\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^2L_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{(1+\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]
\end{aligned}
$$

$$+\underbrace{(-\frac{\gamma\beta_k}{8}+\frac{96L_F^4L_G^2\beta_k^2}{\lambda_k}+\frac{120|\mathcal{S}|L_F^2L_G^4\beta_k^2}{\lambda_k}+\frac{32L_F^2L_H^2\alpha_k^2}{\lambda_k}+(1+\rho)L_F^2\alpha_k+\frac{128L_V^2\alpha_k^2}{\gamma\beta_k}+\frac{32L^2L_F^2\alpha_k^2}{(1-\delta)\lambda_k})}_{A_5}\mathbb{E}[\varepsilon_k^V+\varepsilon_k^J]$$

$$+\underbrace{(-\frac{(1-\delta)\xi_k}{16}+\frac{216L_F^4L_V^2\xi_k^2}{\lambda_k}+\frac{216L_F^2L_G^2L_H^2L_V^2\xi_k^2}{\lambda_k}+\frac{144L_F^2L_H^4L_V^2\xi_k^2}{\lambda_k}+(1+\rho)L_F^2\alpha_k+\frac{192L_V^2\xi_k^2}{\gamma\beta_k})}_{A_6}\mathbb{E}[\varepsilon_k^\mu]. \tag{28}$$

We show that the terms $A_1$-$A_6$ are all non-positive under the step size conditions in (27). First, under the step size condition $\alpha_k \leq \frac{\gamma}{4}\beta_k$, $\lambda_k \leq 1/4$, and $\alpha_k \leq (192(L_F^2+L_G^2+L_H^2+L_V^2+L^2/(1-\delta)+\rho+1))^{-1}\lambda_k$

$$\begin{aligned}
A_1 &= -\frac{\lambda_k}{2}+\lambda_k^2+\frac{48L_F^2\alpha_k^2}{\lambda_k}+\frac{48L_G^2\alpha_k^2}{\lambda_k}+\frac{32L_H^2\alpha_k^2}{\lambda_k}+(1+\rho)\alpha_k+\frac{128L_V^2\alpha_k^2}{\gamma\beta_k}+\frac{32L^2\alpha_k^2}{(1-\delta)\lambda_k}\\
&\leq -\frac{\lambda_k}{4}+\frac{48(L_F^2+L_G^2+L_H^2+L^2/(1-\delta))\alpha_k^2}{\lambda_k}+(1+\rho)\alpha_k+32L_V^2\alpha_k\\
&\leq -\frac{\lambda_k}{4}+48(L_F^2+L_G^2+L_H^2+L_V^2+L^2/(1-\delta)+\rho+1)\alpha_k\\
&\leq 0.
\end{aligned} \tag{29}$$

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

$$\begin{aligned}
A_2 &= -\frac{\lambda_k}{2}+\lambda_k^2+\frac{72|\mathcal{S}|L_G^2\beta_k^2}{\lambda_k}+\frac{36L_F^2\beta_k^2}{\lambda_k}+\frac{8\beta_k}{\gamma}\\
&\leq -\frac{\lambda_k}{4}+(72|\mathcal{S}|L_G^2+36L_F^2+8/\gamma)\beta_k\\
&\leq 0.
\end{aligned} \tag{30}$$

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

$$\begin{aligned}
A_3 &= -\frac{\lambda_k}{2}+\lambda_k^2+\frac{16L_H^2\xi_k^2}{\lambda_k}+\frac{24L_F^2L_H^2\xi_k^2}{\lambda_k}+\frac{24L_G^2\xi_k^2}{\lambda_k}+\frac{64L_V^2\xi_k^2}{\gamma\lambda_k}+\frac{8\xi_k}{1-\delta}\\
&\leq -\frac{\lambda_k}{4}+64(L_H^2L_F^2+L_G^2+L_V^2/\gamma+1/(1-\delta))\xi_k\\
&\leq 0.
\end{aligned} \tag{31}$$

Next, we have

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

under the step size condition

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

Then,

$$A_5 = -\frac{\gamma\beta_k}{8}+\frac{96L_F^4L_G^2\beta_k^2}{\lambda_k}+\frac{120|\mathcal{S}|L_F^2L_G^4\beta_k^2}{\lambda_k}+\frac{32L_F^2L_H^2\alpha_k^2}{\lambda_k}$$

$$
\begin{aligned}
&+ (1+\rho)L_F^2\alpha_k + \frac{128L_V^2\alpha_k^2}{\gamma\beta_k} + \frac{32L^2L_F^2\alpha_k^2}{(1-\delta)\lambda_k} \\
&\leq -\frac{\gamma\beta_k}{8} + 432|\mathcal{S}|(L_F^4L_G^4 + L_F^2L_H^2 + (\rho+1)L_F^2 + L_V^2/\gamma + L^2L_F^2(1-\delta))\frac{\alpha_k^2}{\beta_k} \\
&\leq 0, \qquad (33)
\end{aligned}
$$

due to the condition

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

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

$$
\begin{aligned}
A_6 &= -\frac{(1-\delta)\xi_k}{16} + \frac{216L_F^4L_V^2\xi_k^2}{\lambda_k} + \frac{216L_F^2L_G^2L_H^2L_V^2\xi_k^2}{\lambda_k} \\
&\quad + \frac{144L_F^2L_H^4L_V^2\xi_k^2}{\lambda_k} + (1+\rho)L_F^2\alpha_k + \frac{192L_V^2\xi_k^2}{\gamma\beta_k} \\
&\leq -\frac{(1-\delta)\xi_k}{32} + \frac{216}{\gamma}(L_F^4L_V^2 + L_F^2L_G^2L_H^2L_V^2 + L_F^2L_H^4L_V^2 + L_V^2)\frac{\xi_k^2}{\beta_k} \\
&\leq 0. \qquad (34)
\end{aligned}
$$

Plugging (29)-(34) into (28), we have for all $k \geq \tau_k$

$$
\begin{aligned}
&\mathcal{L}_{k+1} \\
&\leq (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2 + \|\Delta g_k\|^2 + \|\Delta h_k\|^2] - \mathbb{E}[\ell(\pi_{\theta_k})] - \frac{(1+\rho)\alpha_k}{4}\mathbb{E}[\varepsilon_k^\pi] \\
&\quad + (1-\frac{\gamma\beta_k}{8})\mathbb{E}[\varepsilon_k^V + \varepsilon_k^J] + (1-\frac{(1-\delta)\xi_k}{16})\mathbb{E}[\varepsilon_k^\mu] + B_F\alpha_k\Delta \\
&\quad + (28L+2|\mathcal{A}|)L_FL_{TV}B_F^3B_GB_H^2\tau_k^2\lambda_k\lambda_{k-\tau_k} + (30L+2|\mathcal{A}|)L_FL_{TV}B_FB_G^2B_H\tau_k^2\lambda_k\lambda_{k-\tau_k} \\
&\quad + 24LB_FB_H^2\tau_k^2\lambda_k\lambda_{k-\tau_k} + \frac{(1+\rho)L_VB_F^2\alpha_k^2}{2} + 28L_V^2B_F^2B_G^2B_H^2\beta_k^2 + 9L^2B_F^2B_H^2\xi_k^2 \\
&\leq \mathcal{L}_k - \min\left\{\frac{(1+\rho)\alpha_k}{4}, \frac{\gamma\beta_k}{8}, \frac{(1-\delta)\xi_k}{16}\right\}\mathbb{E}[\varepsilon_k^\pi + \varepsilon_k^\mu + \varepsilon_k^V + \varepsilon_k^J] + B_F\alpha_k\Delta + \mathcal{O}(\frac{\log^2(k+1)}{k+1}) \\
&\leq \mathcal{L}_k - \frac{(1+\rho)\alpha_k}{4}\mathbb{E}[\varepsilon_k^\pi + \varepsilon_k^\mu + \varepsilon_k^V + \varepsilon_k^J] + B_F\alpha_k\Delta + \mathcal{O}(\frac{\log^2(k+1)}{k+1}), \qquad (35)
\end{aligned}
$$

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

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

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

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

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

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

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

□

## C. Proof of Corollaries

### C.1. Proof of Corollary 1

As a result of Assumption 5, we have the following gradient domination condition, which is adapted from Lemma 19 of Ganesh et al. (2024).

**Lemma 6.** *Under Assumption 5, we have the following gradient domination condition for any policy parameter $\theta$ and mean field $\mu$*

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

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

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

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,

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

$$\leq \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{k+1}}\right) + \mathcal{O}(\Delta).$$

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

Similarly, we have

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

□

# D. Proof of Propositions

## D.1. Proof of Proposition 1

By the $L_V$-Lipschitz continuity of the function $J$

$$\begin{aligned}
&J(\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})) - J(\pi_{\theta_{k+1}}, \mu^\star(\pi_{\theta_k})) \\
&\leq -\langle \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})}, \theta_{k+1} - \theta_k \rangle + \frac{L_V}{2}\|\theta_{k+1} - \theta_k\|^2 \\
&= -\alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})}, f_k \rangle + \frac{L_V \alpha_k^2}{2}\|f_k\|^2 \\
&= -\alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu)|_{\mu=\mu^\star(\pi_{\theta_k})}, \Delta f_k \rangle - \alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu)|_{\mu=\mu^\star(\pi_{\theta_k})}, \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \rangle + \frac{L_V \alpha_k^2}{2}\|f_k\|^2 \\
&= -\alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})}, \Delta f_k \rangle - \alpha_k \|\nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 \\
&\quad + \alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu)|_{\mu=\mu^\star(\pi_{\theta_k})}, \bar{F}(\theta_k, V^{\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})}, \mu^\star(\pi_{\theta_k})) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \rangle + \frac{L_V \alpha_k^2}{2}\|f_k\|^2 \\
&\leq -\alpha_k \|\nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 - \alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})}, \Delta f_k \rangle \\
&\quad + \alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})}, \bar{F}(\theta_k, V^{\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})}, \mu^\star(\pi_{\theta_k})) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \rangle + \frac{L_V B_F^2 \alpha_k^2}{2},
\end{aligned} \tag{36}$$

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 $\theta$.

To bound the second term on the right hand side of (36), we use 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 $c > 0$

$$-\alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})}, \Delta f_k \rangle \leq \frac{\alpha_k}{4}\|\nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 + \alpha_k \|\Delta f_k\|^2. \tag{37}$$

Similarly, for the third term of (36), we have

$$\begin{aligned}
&\alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})}, \mu^\star(\pi_{\theta_k})) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k) \rangle \\
&\leq \frac{\alpha_k}{4}\|\nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 + \alpha_k \|\bar{F}(\theta_k, V^{\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})}, \mu^\star(\pi_{\theta_k})) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k)\|^2 \\
&\leq \frac{\alpha_k}{4}\|\nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 + L_F^2 \alpha_k \|\Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})})\|^2 + L_F^2 \alpha_k \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\|^2 \\
&= \frac{\alpha_k}{4}\|\nabla_\theta J(\pi_{\theta_k}, \mu) \mid_{\mu=\mu^\star(\pi_{\theta_k})} \|^2 + L_F^2 \alpha_k (\varepsilon_k^V + \varepsilon_k^\mu).
\end{aligned} \tag{38}$$

Plugging (37)-(38) into (36), we have

$$J(\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})) - J(\pi_{\theta_{k+1}}, \mu^\star(\pi_{\theta_k}))$$

$$
\begin{aligned}
&\leq -\alpha_k \|\nabla_\theta J(\pi_{\theta_k}, \mu)\mid_{\mu=\mu^\star(\pi_{\theta_k})}\|^2 - \alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu)\mid_{\mu=\mu^\star(\pi_{\theta_k})}, \Delta f_k\rangle \\
&\quad + \alpha_k \langle \nabla_\theta J(\pi_{\theta_k}, \mu)\mid_{\mu=\mu^\star(\pi_{\theta_k})}, \bar{F}(\theta_k, V^{\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})}, \mu^\star(\pi_{\theta_k})) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k)\rangle + \frac{L_V B_F^2 \alpha_k^2}{2} \\
&\leq -\alpha_k \|\nabla_\theta J(\pi_{\theta_k}, \mu)\mid_{\mu=\mu^\star(\pi_{\theta_k})}\|^2 + \frac{\alpha_k}{4}\|\nabla_\theta J(\pi_{\theta_k}, \mu)\mid_{\mu=\mu^\star(\pi_{\theta_k})}\|^2 + \alpha_k \|\Delta f_k\|^2 \\
&\quad + \frac{\alpha_k}{4}\|\nabla_\theta J(\pi_{\theta_k}, \mu)\mid_{\mu=\mu^\star(\pi_{\theta_k})}\|^2 + L_F^2 \alpha_k (\varepsilon_k^V + \varepsilon_k^\mu) + \frac{L_V B_F^2 \alpha_k^2}{2} \\
&\leq -\frac{\alpha_k}{2}\|\nabla_\theta J(\pi_{\theta_k}, \mu)\mid_{\mu=\mu^\star(\pi_{\theta_k})}\|^2 + \alpha_k \|\Delta f_k\|^2 + L_F^2 \alpha_k (\varepsilon_k^V + \varepsilon_k^\mu) + \frac{L_V B_F^2 \alpha_k^2}{2}.
\end{aligned} \tag{39}
$$

By Assumption 4, we have

$$
\begin{aligned}
&J(\pi_{\theta_{k+1}}, \mu^\star(\pi_{\theta_k})) - J(\pi_{\theta_{k+1}}, \mu^\star(\pi_{\theta_{k+1}})) \\
&\leq \rho\Big(J(\pi_{\theta_k}, \mu^\star(\pi_{\theta_k})) - J(\pi_{\theta_{k+1}}, \mu^\star(\pi_{\theta_k}))\Big) + \Delta\|\theta_{k+1} - \theta_k\| \\
&\leq \rho\big(-\frac{\alpha_k}{2}\|\nabla_\theta J(\pi_{\theta_k}, \mu)\mid_{\mu=\mu^\star(\pi_{\theta_k})}\|^2 + \alpha_k\|\Delta f_k\|^2 + L_F^2\alpha_k(\varepsilon_k^V + \varepsilon_k^\mu) + \frac{L_V B_F^2 \alpha_k^2}{2}\big) + B_F \alpha_k \Delta.
\end{aligned} \tag{40}
$$

Combining (39) and (40),

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

□

### 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 $k \geq \tau_k$*

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

By the update rule of $f_k$,

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

Taking the norm, we have

$$
\|\Delta f_{k+1}\|^2
$$

$$
\begin{aligned}
&= (1-\lambda_k)^2\|\Delta f_k\|^2 + \|\bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k) - \bar{F}(\theta_{k+1},\hat{V}_{k+1},\hat{\mu}_{k+1})\|^2 \\
&\quad + \lambda_k^2\|F(\theta_k,\hat{V}_k,\hat{\mu}_k,s_k,a_k,s_{k+1}) - \bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k)\|^2 \\
&\quad + (1-\lambda_k)\langle \Delta f_k, \bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k) - \bar{F}(\theta_{k+1},\hat{V}_{k+1},\hat{\mu}_{k+1})\rangle \\
&\quad + (1-\lambda_k)\lambda_k\langle \Delta f_k, F(\theta_k,\hat{V}_k,\hat{\mu}_k,s_k,a_k,s_{k+1}) - \bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k)\rangle \\
&\quad + \lambda_k\langle \bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k) - \bar{F}(\theta_{k+1},\hat{V}_{k+1},\hat{\mu}_{k+1}), F(\theta_k,\hat{V}_k,\hat{\mu}_k,s_k,a_k,s_{k+1}) - \bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k)\rangle \\
&\le (1-\lambda_k)^2\|\Delta f_k\|^2 + 2\|\bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k) - \bar{F}(\theta_{k+1},\hat{V}_{k+1},\hat{\mu}_{k+1})\|^2 \\
&\quad + 2\lambda_k^2\|F(\theta_k,\hat{V}_k,\hat{\mu}_k,s_k,a_k,s_{k+1}) - \bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k)\|^2 \\
&\quad + \frac{\lambda_k}{2}\|\Delta f_k\|^2 + \frac{2}{\lambda_k}\|\bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k) - \bar{F}(\theta_{k+1},\hat{V}_{k+1},\hat{\mu}_{k+1})\|^2 \\
&\quad + (1-\lambda_k)\lambda_k\langle \Delta f_k, F(\theta_k,\hat{V}_k,\hat{\mu}_k,s_k,a_k,s_{k+1}) - \bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k)\rangle \\
&\le (1-\lambda_k)\|\Delta f_k\|^2 + (-\frac{\lambda_k}{2}+\lambda_k^2)\|\Delta f_k\|^2 + \frac{4}{\lambda_k}\|\bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k) - \bar{F}(\theta_{k+1},\hat{V}_{k+1},\hat{\mu}_{k+1})\|^2 \\
&\quad + 8B_F^2\lambda_k^2 + (1-\lambda_k)\lambda_k\langle \Delta f_k, F(\theta_k,\hat{V}_k,\hat{\mu}_k,s_k,a_k,s_{k+1}) - \bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k)\rangle, \qquad (41)
\end{aligned}
$$

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

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

Taking the expectation, we can simplify (41) as

$$
\begin{aligned}
&\mathbb{E}[\|\Delta f_{k+1}\|^2] \\
&\le \mathbb{E}\Big[(1-\lambda_k)\|\Delta f_k\|^2 + (-\frac{\lambda_k}{2}+\lambda_k^2)\|\Delta f_k\|^2 + \frac{4}{\lambda_k}\|\bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k) - \bar{F}(\theta_{k+1},\hat{V}_{k+1},\hat{\mu}_{k+1})\|^2 \\
&\quad + 8B_F^2\lambda_k^2 + (1-\lambda_k)\lambda_k\langle \Delta f_k, F(\theta_k,\hat{V}_k,\hat{\mu}_k,s_k,a_k,s_{k+1}) - \bar{F}(\theta_k,\hat{V}_k,\hat{\mu}_k)\rangle\Big] \\
&\le (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] + (-\frac{\lambda_k}{2}+\lambda_k^2)\mathbb{E}[\|\Delta f_k\|^2] + 8B_F^2\lambda_k^2 \\
&\quad + \frac{4L_F^2}{\lambda_k}\mathbb{E}[\Big(\|\theta_k-\theta_{k+1}\| + \|\hat{V}_k-\hat{V}_{k+1}\| + \|\hat{\mu}_k-\hat{\mu}_{k+1}\|\Big)^2] \\
&\quad + (1-\lambda_k)\lambda_k\cdot(20L+2|\mathcal{A}|)L_F L_{TV} B_F^3 B_G B_H^2 \tau_k^2 \lambda_{k-\tau_k} \\
&\le (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] + (-\frac{\lambda_k}{2}+\lambda_k^2)\mathbb{E}[\|\Delta f_k\|^2] + (28L+2|\mathcal{A}|)L_F L_{TV} B_F^3 B_G B_H^2 \tau_k^2 \lambda_k\lambda_{k-\tau_k} \\
&\quad + \frac{4L_F^2}{\lambda_k}\mathbb{E}[(\alpha_k\|f_k\| + \beta_k\|g_k\| + \xi_k\|h_k\|)^2] \\
&\le (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] + (-\frac{\lambda_k}{2}+\lambda_k^2)\mathbb{E}[\|\Delta f_k\|^2] + (28L+2|\mathcal{A}|)L_F L_{TV} B_F^3 B_G B_H^2 \tau_k^2 \lambda_k\lambda_{k-\tau_k} \\
&\quad + \mathbb{E}\Big[\frac{12L_F^2\alpha_k}{\lambda_k}\left(\|\Delta f_k\| + L_F\sqrt{\varepsilon_k^V} + L_F(L_V+1)\sqrt{\varepsilon_k^\mu} + \sqrt{\varepsilon_k^\pi}\right)^2 \\
&\quad + \frac{12L_F^2\beta_k}{\lambda_k}\left(\|\Delta g_k\| + L_G\sqrt{\varepsilon_k^V} + L_G\sqrt{\varepsilon_k^J}\right)^2 + \frac{12L_F^2\xi_k}{\lambda_k}\left(L_H\|\Delta h_k\| + \sqrt{\varepsilon_k^\mu}\right)^2\Big], \qquad (42)
\end{aligned}
$$

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

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

$$
\begin{aligned}
&\leq \frac{48L_F^2\alpha_k^2}{\lambda_k}\|\Delta f_k\|^2 + \frac{48L_F^4\alpha_k^2}{\lambda_k}\varepsilon_k^V + \frac{192L_F^4L_V^2\alpha_k^2}{\lambda_k}\varepsilon_k^\mu + \frac{48L_F^2\alpha_k^2}{\lambda_k}\varepsilon_k^\pi \\
&\quad + \frac{36L_F^2\beta_k^2}{\lambda_k}\|\Delta g_k\|^2 + \frac{48L_F^2L_G^2\beta_k^2}{\lambda_k}\varepsilon_k^V + \frac{48L_F^2L_G^2\beta_k^2}{\lambda_k}\varepsilon_k^J \\
&\quad + \frac{24L_F^2L_H^2\xi_k^2}{\lambda_k}\|\Delta h_k\|^2 + \frac{24L_F^2\xi_k^2}{\lambda_k}\varepsilon_k^\mu \\
&\leq \frac{48L_F^2\alpha_k^2}{\lambda_k}\|\Delta f_k\|^2 + \frac{36L_F^2\beta_k^2}{\lambda_k}\|\Delta g_k\|^2 + \frac{24L_F^2L_H^2\xi_k^2}{\lambda_k}\|\Delta h_k\|^2 + \frac{48L_F^2\alpha_k^2}{\lambda_k}\varepsilon_k^\pi \\
&\quad + \frac{216L_F^4L_V^2\xi_k^2}{\lambda_k}\varepsilon_k^\mu + \frac{96L_F^4L_G^2\beta_k^2}{\lambda_k}\varepsilon_k^V + \frac{48L_F^2L_G^2\beta_k^2}{\lambda_k}\varepsilon_k^J. \qquad (43)
\end{aligned}
$$

Combining (42) and (43), we get

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

□

### 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 3, we have for any policy parameter $\theta$ and mean field $\mu$*

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

By the definition of $\varepsilon_k^\mu$,

$$
\begin{aligned}
\varepsilon_{k+1}^\mu &= \|\hat{\mu}_{k+1} - \mu^\star(\pi_{\theta_{k+1}})\|^2 \\
&= \|\hat{\mu}_k + \xi_k h_k - \mu^\star(\pi_{\theta_{k+1}})\|^2 \\
&= \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k \Delta h_k + \xi_k \bar{H}(\theta_k,\hat{\mu}_k) - \big(\mu^\star(\pi_{\theta_{k+1}}) - \mu^\star(\pi_{\theta_k})\big)\|^2 \\
&= \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k \bar{H}(\theta_k,\hat{\mu}_k)\|^2 + \xi_k^2\|\Delta h_k\|^2 + \|\mu^\star(\pi_{\theta_{k+1}}) - \mu^\star(\pi_{\theta_k})\|^2 \\
&\quad + 2\xi_k\langle \hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k\bar{H}(\theta_k,\hat{\mu}_k), \Delta h_k\rangle \\
&\quad + 2\langle \hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k\bar{H}(\theta_k,\hat{\mu}_k), \mu^\star(\pi_{\theta_{k+1}}) - \mu^\star(\pi_{\theta_k})\rangle \\
&\quad + 2\xi_k\langle \Delta h_k, \mu^\star(\pi_{\theta_{k+1}}) - \mu^\star(\pi_{\theta_k})\rangle. \qquad (44)
\end{aligned}
$$

To bound the first term of (44),

$$
\|\hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k\bar{H}(\theta_k,\hat{\mu}_k)\|^2
$$

$$
\begin{aligned}
&= \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k\big(\bar{H}(\theta_k, \hat{\mu}_k) - \bar{H}(\theta_k, \mu^\star(\pi_{\theta_k}))\big)\|^2 \\
&= \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\|^2 + \xi_k^2\|\bar{H}(\theta_k, \hat{\mu}_k) - \bar{H}(\theta_k, \mu^\star(\pi_{\theta_k}))\|^2 \\
&\quad + 2\xi_k\langle \hat{\mu}_k - \mu^\star(\pi_{\theta_k}), \bar{H}(\theta_k, \hat{\mu}_k) - \bar{H}(\theta_k, \mu^\star(\pi_{\theta_k}))\rangle \\
&\leq \|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\|^2 + L_H^2\xi_k^2\|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\|^2 - (1-\delta)\xi_k\|\hat{\mu}_k - \mu^\star(\pi_{\theta_k})\|^2 \\
&\leq (1 - \frac{(1-\delta)\xi_k}{2})\varepsilon_k^\mu,
\end{aligned} \tag{45}
$$

where the first equation uses $\bar{H}(\theta, \mu^\star(\pi_\theta)) = 0$ for any $\theta$, 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 (44) 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

$$
\begin{aligned}
\xi_k^2\|\Delta h_k\|^2 + \|\mu^\star(\pi_{\theta_{k+1}}) - \mu^\star(\pi_{\theta_k})\|^2 &\leq 2\xi_k^2\|h_k\|^2 + 2\xi_k^2\|\bar{H}(\theta_k, \hat{\mu}_k)\|^2 + L\|\pi_{\theta_{k+1}} - \pi_{\theta_k}\|^2 \\
&\leq 4B_H^2\xi_k^2 + L^2\|f_k\|^2 \\
&\leq 4B_H^2\xi_k^2 + L^2B_F^2\alpha_k^2.
\end{aligned} \tag{46}
$$

The fourth term of (44) can be bounded leveraging the result in (45) as follows

$$
\begin{aligned}
&2\xi_k\langle \hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k\bar{H}(\theta_k, \hat{\mu}_k), \Delta h_k\rangle \\
&\leq \frac{(1-\delta)\xi_k}{8}\|\hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k\bar{H}(\theta_k, \hat{\mu}_k)\|^2 + \frac{8\xi_k}{1-\delta}\|\Delta h_k\|^2 \\
&\leq \frac{(1-\delta)\xi_k}{8}\cdot(1 - \frac{(1-\delta)\xi_k}{2})\varepsilon_k^\mu + \frac{8\xi_k}{1-\delta}\|\Delta h_k\|^2 \\
&\leq \frac{(1-\delta)\xi_k}{8}\varepsilon_k^\mu + \frac{8\xi_k}{1-\delta}\|\Delta h_k\|^2.
\end{aligned} \tag{47}
$$

Similarly, for the fifth term of (44), we have

$$
\begin{aligned}
&2\langle \hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k\bar{H}(\theta_k, \hat{\mu}_k), \mu^\star(\pi_{\theta_{k+1}}) - \mu^\star(\pi_{\theta_k})\rangle \\
&\leq \frac{(1-\delta)\xi_k}{8}\|\hat{\mu}_k - \mu^\star(\pi_{\theta_k}) + \xi_k\bar{H}(\theta_k, \hat{\mu}_k)\|^2 + \frac{8}{(1-\delta)\xi_k}\|\mu^\star(\pi_{\theta_{k+1}}) - \mu^\star(\pi_{\theta_k})\|^2 \\
&\leq \frac{(1-\delta)\xi_k}{8}\varepsilon_k^\mu + \frac{8L^2}{(1-\delta)\xi_k}\|\pi_{\theta_{k+1}} - \pi_{\theta_k}\|^2 \\
&\leq \frac{(1-\delta)\xi_k}{8}\varepsilon_k^\mu + \frac{8L^2\alpha_k^2}{(1-\delta)\xi_k}\|f_k\|^2 \\
&\leq \frac{(1-\delta)\xi_k}{8}\varepsilon_k^\mu + \frac{8L^2\alpha_k^2}{(1-\delta)\xi_k}\left(\|\Delta f_k\| + L_F\sqrt{\varepsilon_k^V} + L_F(L_V+1)\sqrt{\varepsilon_k^\mu} + \sqrt{\varepsilon_k^\pi}\right)^2 \\
&\leq \frac{(1-\delta)\xi_k}{8}\varepsilon_k^\mu + \frac{32L^2\alpha_k^2}{(1-\delta)\xi_k}\left(\|\Delta f_k\|^2 + L_F^2\varepsilon_k^V + 4L_F^2L_V^2\varepsilon_k^\mu + \varepsilon_k^\pi\right),
\end{aligned} \tag{48}
$$

where the fourth inequality follows from Lemma 4.

The final term of (44) can be bounded simply with the Cauchy-Schwarz inequality

$$
\begin{aligned}
2\xi_k\langle \Delta h_k, \mu^\star(\pi_{\theta_{k+1}}) - \mu^\star(\pi_{\theta_k})\rangle &\leq 2\xi_k\|\Delta h_k\|\|\mu^\star(\pi_{\theta_{k+1}}) - \mu^\star(\pi_{\theta_k})\| \\
&\leq 4B_H\xi_k\cdot L\|\pi_{\theta_{k+1}} - \pi_{\theta_k}\| \\
&\leq 4LB_FB_H\alpha_k\xi_k.
\end{aligned} \tag{49}
$$

Plugging (45)-(49) into (44), we get

$$
\varepsilon_{k+1}^\mu
$$

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

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

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

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

$$
\begin{aligned}
&\mathbb{E}[\|\Delta h_{k+1}\|^2] \\
&\leq (1-\lambda_k)\mathbb{E}[\|\Delta h_k\|^2] + (-\frac{\lambda_k}{2} + \lambda_k^2)\mathbb{E}[\|\Delta h_k\|^2] + \frac{4}{\lambda_k}\mathbb{E}[(L_H\|\theta_k - \theta_{k+1}\| + L_H\|\hat{\mu}_k - \hat{\mu}_{k+1}\|)^2]
\end{aligned}
$$

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

where the third inequality bounds $\|f_k\|$ and $\|h_k\|$ with Lemma 4. The step size condition $\alpha_k \le \xi_k$ 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 $\theta$, $\mu$, and $V$ that*

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

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

By the definition of $\varepsilon_k^V$,

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

$$
\begin{aligned}
&+ \left\| \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_{k+1}},\hat{\mu}_{k+1}} - V^{\pi_{\theta_k},\hat{\mu}_k}) \\ J(\pi_{\theta_{k+1}},\hat{\mu}_{k+1}) - J(\pi_{\theta_k},\hat{\mu}_k) \end{bmatrix} \right\|^2 \\
&+ 2\beta_k \left\langle \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k},\hat{\mu}_k}) \\ \hat{J}_k - J(\pi_{\theta_k},\hat{\mu}_k) \end{bmatrix} + \beta_k \begin{bmatrix} \Pi_{\mathcal{E}_\perp}\bar{G}^V(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k) \\ \bar{G}^J(\theta_k,\hat{J}_k,\hat{\mu}_k) \end{bmatrix}, \begin{bmatrix} \Pi_{\mathcal{E}_\perp}\Delta g_k^V \\ \Delta g_k^J \end{bmatrix} \right\rangle \\
&+ 2 \left\langle \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k},\hat{\mu}_k}) \\ \hat{J}_k - J(\pi_{\theta_k},\hat{\mu}_k) \end{bmatrix} + \beta_k \begin{bmatrix} \Pi_{\mathcal{E}_\perp}\bar{G}^V(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k) \\ \bar{G}^J(\theta_k,\hat{J}_k,\hat{\mu}_k) \end{bmatrix}, \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_{k+1}},\hat{\mu}_{k+1}} - V^{\pi_{\theta_k},\hat{\mu}_k}) \\ J(\pi_{\theta_{k+1}},\hat{\mu}_{k+1}) - J(\pi_{\theta_k},\hat{\mu}_k) \end{bmatrix} \right\rangle \\
&+ 2\beta_k \left\langle \Delta g_k, \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_{k+1}},\hat{\mu}_{k+1}} - V^{\pi_{\theta_k},\hat{\mu}_k}) \\ J(\pi_{\theta_{k+1}},\hat{\mu}_{k+1}) - J(\pi_{\theta_k},\hat{\mu}_k) \end{bmatrix} \right\rangle,
\end{aligned} \tag{50}
$$

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 (50),

$$
\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{51}
$$

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 $\theta,\mu$, 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 (50), we use the boundedness of $\Delta g_k$ and the Lipschitz continuity conditions from Lemma 1

$$
\begin{aligned}
&\beta_k^2\|\Delta g_k\|^2 + \left\| \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_{k+1}},\hat{\mu}_{k+1}} - V^{\pi_{\theta_k},\hat{\mu}_k}) \\ J(\pi_{\theta_{k+1}},\hat{\mu}_{k+1}) - J(\pi_{\theta_k},\hat{\mu}_k) \end{bmatrix} \right\|^2 \\
&\leq \beta_k^2\|\Delta g_k\|^2 + \|\Pi_{\mathcal{E}_\perp}(V^{\pi_{\theta_{k+1}},\hat{\mu}_{k+1}} - V^{\pi_{\theta_k},\hat{\mu}_k})\|^2 + (J(\pi_{\theta_{k+1}},\hat{\mu}_{k+1}) - J(\pi_{\theta_k},\hat{\mu}_k))^2 \\
&\leq 4B_G^2\beta_k^2 + (L_V\|\theta_{k+1}-\theta_k\| + L_V\|\hat{\mu}_{k+1}-\hat{\mu}_k\|)^2 + (L_V\|\theta_{k+1}-\theta_k\| + L_V\|\hat{\mu}_{k+1}-\hat{\mu}_k\|)^2 \\
&= 4B_G^2\beta_k^2 + 2L_V^2\left(\alpha_k\|f_k\| + \xi_k\|h_k\|\right)^2 \\
&= 4B_G^2\beta_k^2 + 2L_V^2\xi_k^2\left(B_F + B_H\right)^2 \\
&\leq 4L_V^2(B_F^2 + B_G^2 + B_H^2)\beta_k^2,
\end{aligned} \tag{52}
$$

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

The fourth term of (50) can be bounded leveraging the result in (51) as follows

$$
\begin{aligned}
&2\beta_k \left\langle \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k},\hat{\mu}_k}) \\ \hat{J}_k - J(\pi_{\theta_k},\hat{\mu}_k) \end{bmatrix} + \beta_k \begin{bmatrix} \Pi_{\mathcal{E}_\perp}\bar{G}^V(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k) \\ \bar{G}^J(\theta_k,\hat{J}_k,\hat{\mu}_k) \end{bmatrix}, \begin{bmatrix} \Pi_{\mathcal{E}_\perp}\Delta g_k^V \\ \Delta g_k^J \end{bmatrix} \right\rangle \\
&\leq \frac{\gamma\beta_k}{8} \left\| \begin{bmatrix} \Pi_{\mathcal{E}_\perp}(\hat{V}_k - V^{\pi_{\theta_k},\hat{\mu}_k}) \\ \hat{J}_k - J(\pi_{\theta_k},\hat{\mu}_k) \end{bmatrix} + \beta_k \begin{bmatrix} \Pi_{\mathcal{E}_\perp}\bar{G}^V(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k) \\ \bar{G}^J(\theta_k,\hat{J}_k,\hat{\mu}_k) \end{bmatrix} \right\|^2 + \frac{8\beta_k}{\gamma} \left\| \begin{bmatrix} \Pi_{\mathcal{E}_\perp}\Delta g_k^V \\ g_k^J \end{bmatrix} \right\|^2 \\
&\leq \frac{\gamma\beta_k}{8}(1-\frac{\gamma\beta_k}{2})(\varepsilon_k^V + \varepsilon_k^J) + \frac{8\beta_k}{\gamma}\|\Delta g_k\|^2 \\
&\leq \frac{\gamma\beta_k}{8}(\varepsilon_k^V + \varepsilon_k^J) + \frac{8\beta_k}{\gamma}\|\Delta g_k\|^2. \qquad (53)
\end{aligned}
$$

Similarly, for the fifth term of (50), 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), \qquad (54)
\end{aligned}
$$

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

The final term of (50) can be bounded simply with the Cauchy-Schwarz inequality

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

$$
\begin{aligned}
&\leq 4B_G\beta_k \cdot \big(L_V(\|\pi_{\theta_{k+1}} - \pi_{\theta_k}\| + \|\hat{\mu}_{k+1} - \hat{\mu}_k\|) + L_V(\|\pi_{\theta_{k+1}} - \pi_{\theta_k}\| + \|\hat{\mu}_{k+1} - \hat{\mu}_k\|)\big) \\
&\leq 8L_V B_G \beta_k (B_F\alpha_k + B_H\xi_k) \\
&\leq 16 L_V B_F B_G B_H \beta_k \xi_k. \qquad (55)
\end{aligned}
$$

Plugging (51)-(55) into (50), we get

$$
\begin{aligned}
&\varepsilon^V_{k+1} + \varepsilon^J_{k+1} \\
&\leq (1 - \frac{\gamma\beta_k}{2})(\varepsilon^V_k + \varepsilon^J_k) + 4L_V^2(B_F^2 + B_G^2 + B_H^2)\beta_k^2 + \frac{\gamma\beta_k}{8}(\varepsilon^V_k + \varepsilon^J_k) + \frac{8\beta_k}{\gamma}\|\Delta g_k\|^2 \\
&\quad + \frac{\gamma\beta_k}{8}(\varepsilon^V_k + \varepsilon^J_k) + \frac{128L_V^2\alpha_k^2}{\gamma\beta_k}\left(\|\Delta f_k\|^2 + L_F^2\varepsilon^V_k + 4L_F^2L_V^2\varepsilon^\mu_k + \varepsilon^\pi_k\right) \\
&\quad + \frac{64L_V^2\xi_k^2}{\gamma\beta_k}\left(\|\Delta h_k\|^2 + L_H^2\varepsilon^\mu_k\right) + 16L_V B_F B_G B_H \beta_k \xi_k \\
&\leq (1 - \frac{\gamma\beta_k}{4})(\varepsilon^V_k + \varepsilon^J_k) + \frac{128L_V^2\alpha_k^2}{\gamma\beta_k}\|\Delta f_k\|^2 + \frac{8\beta_k}{\gamma}\|\Delta g_k\|^2 + \frac{64L_V^2\xi_k^2}{\gamma\beta_k}\|\Delta h_k\|^2 \\
&\quad + \frac{128L_V^2\alpha_k^2}{\gamma\beta_k}(L_F^2\varepsilon^V_k + \varepsilon^\pi_k) + \frac{192L_V^2\xi_k^2}{\gamma\beta_k}\varepsilon^\mu_k + 28L_V^2B_F^2B_G^2B_H^2\beta_k^2,
\end{aligned}
$$

where we use the conditions $\xi_k \leq \beta_k$ and $\alpha_k \leq \frac{L_H}{2L_FL_V}\xi_k$ 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 $f_k$,

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

Taking the norm, we have

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

$$
\begin{aligned}
&\leq (1-\lambda_k)^2\|\Delta g_k\|^2 + 2\|\bar{G}(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k) - \bar{G}(\theta_{k+1},\hat{V}_{k+1},\hat{J}_{k+1},\hat{\mu}_{k+1})\|^2 \\
&\quad + 2\lambda_k^2\|G(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k,s_k,a_k,s_{k+1}) - \bar{G}(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k)\|^2 \\
&\quad + \frac{\lambda_k}{2}\|\Delta g_k\|^2 + \frac{2}{\lambda_k}\|\bar{G}(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k) - \bar{G}(\theta_{k+1},\hat{V}_{k+1},\hat{J}_{k+1},\hat{\mu}_{k+1})\|^2 \\
&\quad + (1-\lambda_k)\lambda_k\langle \Delta g_k, G(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k,s_k,a_k,s_{k+1}) - \bar{G}(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k)\rangle \\
&\leq (1-\lambda_k)\|\Delta g_k\|^2 + (-\frac{\lambda_k}{2}+\lambda_k^2)\|\Delta g_k\|^2 \\
&\quad + \frac{4}{\lambda_k}\|\bar{G}(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k) - \bar{G}(\theta_{k+1},\hat{V}_{k+1},\hat{J}_{k+1},\hat{\mu}_{k+1})\|^2 \\
&\quad + 8B_G^2\lambda_k^2 + (1-\lambda_k)\lambda_k\langle \Delta g_k, G(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k,s_k,a_k,s_{k+1}) - \bar{G}(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k)\rangle,
\end{aligned} \tag{56}
$$

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 (56) 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 \\
&\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\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 \\
&\quad + \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] \\
&\quad + (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} \\
&\quad + \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} \\
&\quad + \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} \\
&\quad + \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] \\
&\quad + \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],
\end{aligned} \tag{57}
$$

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

$$
\begin{aligned}
&\frac{12L_G^2\alpha_k^2}{\lambda_k}\mathbb{E}\left[\Bigg(\|\Delta f_k\| + L_F\sqrt{\varepsilon_k^V} + L_F(L_V+1)\sqrt{\varepsilon_k^\mu} + \sqrt{\varepsilon_k^\pi}\Bigg)^2\right] \\
&+ \frac{24|\mathcal{S}|L_G^2\beta_k^2}{\lambda_k}\mathbb{E}[\Bigg(\|\Delta g_k\| + L_G\sqrt{\varepsilon_k^V} + L_G\sqrt{\varepsilon_k^J}\Bigg)^2] + \frac{12L_G^2\xi_k^2}{\lambda_k}\mathbb{E}[\Bigg(\|\Delta h_k\| + L_H\sqrt{\varepsilon_k^\mu}\Bigg)^2] \\
&\leq \frac{48L_G^2\alpha_k^2}{\lambda_k}\mathbb{E}[\|\Delta f_k\|^2 + L_F^2\varepsilon_k^V + 4L_F^2L_V^2\varepsilon_k^\mu + \varepsilon_k^\pi] + \frac{72|\mathcal{S}|L_G^2\beta_k^2}{\lambda_k}\mathbb{E}[\|\Delta g_k\|^2 + L_G^2\varepsilon_k^V + L_G^2\varepsilon_k^J]
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{24L_G^2\xi_k^2}{\lambda_k}\mathbb{E}[\|\Delta h_k\|^2 + L_H^2\varepsilon_k^\mu] \\
&\leq \mathbb{E}\Big[\frac{48L_G^2\alpha_k^2}{\lambda_k}\|\Delta f_k\|^2 + \frac{72|\mathcal{S}|L_G^2\beta_k^2}{\lambda_k}\|\Delta g_k\|^2 + \frac{24L_G^2\xi_k^2}{\lambda_k}\|\Delta h_k\|^2 + \frac{48L_G^2\alpha_k^2}{\lambda_k}\varepsilon_k^\pi \\
&\quad + \frac{216L_F^2L_G^2L_H^2L_V^2\xi_k^2}{\lambda_k}\varepsilon_k^\mu + \frac{120|\mathcal{S}|L_F^2L_G^4\beta_k^2}{\lambda_k}(\varepsilon_k^V + \varepsilon_k^J)\Big]. \qquad (58)
\end{aligned}
$$

Combining (57) and (58), we have

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

□

## 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 2.

The Lipschitz gradient condition of $J$ in $\theta$ is proved in Lemma 4 of Kumar et al. (2024) and can be extended to the gradient of $J$ in $\mu$ by a similar argument.

□

### E.2. Proof of Lemma 2

First, by definition in (24),

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

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

Similarly, we have

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

and

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

which implies

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

Finally, we have

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

□

### E.3. Proof of Lemma 3

By the definition of $\bar{F}(\theta,V,\mu)$ in (25),

$$\begin{aligned}&\|\bar{F}(\theta_1,V_1,\mu_1)-\bar{F}(\theta_2,V_2,\mu_2)\| \\ &= \|\mathbb{E}_{s\sim\nu^{\pi_{\theta_1},\mu_1},a\sim\pi_{\theta_1}(\cdot|s),s'\sim\mathcal{P}(\cdot|s,a,\mu_1)}[F(\theta_1,V_1,\mu_1,s,a,s')] \\ &\qquad - \mathbb{E}_{s\sim\nu^{\pi_{\theta_2},\mu_2},a\sim\pi_{\theta_2}(\cdot|s),s'\sim\mathcal{P}(\cdot|s,a,\mu_2)}[F(\theta_2,V_2,\mu_2,s,a,s')]\| \\ &= \|\mathbb{E}_{s\sim\nu^{\pi_{\theta_1},\mu_1},a\sim\pi_{\theta_1}(\cdot|s),s'\sim\mathcal{P}(\cdot|s,a,\mu_1)}[F(\theta_1,\Pi_{\mathcal{E}_\perp}V_1,\mu_1,s,a,s')] \\ &\qquad - \mathbb{E}_{s\sim\nu^{\pi_{\theta_2},\mu_2},a\sim\pi_{\theta_2}(\cdot|s),s'\sim\mathcal{P}(\cdot|s,a,\mu_2)}[F(\theta_2,\Pi_{\mathcal{E}_\perp}V_2,\mu_2,s,a,s')]\| \\ &= \|\sum_{s,a,s'}(\nu^{\pi_{\theta_1},\mu_1}(s)\pi_{\theta_1}(a\mid s)\mathcal{P}_{\mu_1}(\cdot\mid s,a)-\nu^{\pi_{\theta_2},\mu_2}(s)\pi_{\theta_2}(a\mid s)\mathcal{P}_{\mu_2}(\cdot\mid s,a))F(\theta_2,\Pi_{\mathcal{E}_\perp}V_2,\mu_2,s,a,s') \\ &\qquad + \mathbb{E}_{s\sim\nu^{\pi_{\theta_1},\mu_1},a\sim\pi_{\theta_1}(\cdot|s),s'\sim\mathcal{P}(\cdot|s,a,\mu_1)}[F(\theta_1,\Pi_{\mathcal{E}_\perp}V_1,\mu_1,s,a,s')-F(\theta_2,\Pi_{\mathcal{E}_\perp}V_2,\mu_2,s,a,s')]\| \\ &\le \|\mathbb{E}_{s\sim\nu^{\pi_{\theta_1},\mu_1},a\sim\pi_{\theta_1}(\cdot|s),s'\sim\mathcal{P}_{\mu_1}(\cdot|s,a,\mu_1)}[F(\theta_1,\Pi_{\mathcal{E}_\perp}V_1,\mu_1,s,a,s')-F(\theta_2,\Pi_{\mathcal{E}_\perp}V_2,\mu_2,s,a,s')]\| \\ &\qquad + 2B_F d_{TV}(\nu^{\pi_{\theta_1},\mu_1}\otimes\pi_{\theta_1}\otimes\mathcal{P}_{\mu_1},\nu^{\pi_{\theta_2},\mu_2}\otimes\pi_{\theta_2}\otimes\mathcal{P}_{\mu_2}),\end{aligned} \tag{59}$$

where the inequality comes from the definition of TV distance in (17) and the second equation is a result of the fact that for any constant $c$

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

For any $s,a,s'$ we have from (24)

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

$$
\begin{aligned}
&\leq |r(s,a,\mu_1) - r(s,a,\mu_2)| \|\nabla_\theta \log \pi_{\theta_1}(a \mid s)\| \\
&\quad + |r(s,a,\mu_2)| \|\nabla_\theta \log \pi_{\theta_1}(a \mid s) - \nabla_\theta \log \pi_{\theta_2}(a \mid s)\| \\
&\quad + |\Pi_{\mathcal{E}_\perp} V_1(s') - \Pi_{\mathcal{E}_\perp} V_1(s) - \Pi_{\mathcal{E}_\perp} V_2(s') + \Pi_{\mathcal{E}_\perp} V_2(s)| \|\nabla_\theta \log \pi_{\theta_1}(a \mid s)\| \\
&\quad + |\Pi_{\mathcal{E}_\perp} V_2(s') - \Pi_{\mathcal{E}_\perp} V_2(s)| \|\nabla_\theta \log \pi_{\theta_1}(a \mid s) - \nabla_\theta \log \pi_{\theta_2}(a \mid s)\| \\
&\leq |r(s,a,\mu_1) - r(s,a,\mu_2)| + (1 + 2\|V\|) \|\nabla_\theta \log \pi_{\theta_1}(a \mid s) - \nabla_\theta \log \pi_{\theta_2}(a \mid s)\| \\
&\leq L\|\mu_1 - \mu_2\| + \|\nabla_\theta \log \pi_{\theta_1}(a \mid s) - \nabla_\theta \log \pi_{\theta_2}(a \mid s)\| \\
&\quad + 2\|\Pi_{\mathcal{E}_\perp} V_1 - \Pi_{\mathcal{E}_\perp} V_2\| + 2\|\Pi_{\mathcal{E}_\perp} V_2\| \|\nabla_\theta \log \pi_{\theta_1}(a \mid s) - \nabla_\theta \log \pi_{\theta_2}(a \mid s)\| \\
&\leq 5(2B_V + 1)\|\theta_1 - \theta_2\| + L\|\mu_1 - \mu_2\| + 2\|\Pi_{\mathcal{E}_\perp} V_1 - \Pi_{\mathcal{E}_\perp} V_2\|, \qquad (60)
\end{aligned}
$$

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 2, 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]).

Plugging (60) and the relation in (26) into (59), we have

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

Following a line of argument similar to (59),

$$
\begin{aligned}
&\|\bar{G}(\theta_1, V_1, J_1, \mu_1) - \bar{G}(\theta_2, V_2, J_2, \mu_2)\| \\
&\leq \|\mathbb{E}_{s \sim \nu^{\pi_{\theta_1}, \mu_1}, a \sim \pi_{\theta_1}(\cdot|s), s' \sim \mathcal{P}_{\mu_1}(\cdot|s,a,\mu_1)} [G(\Pi_{\mathcal{E}_\perp} V_1, J_1, \mu_1, s, a, s') - G(\Pi_{\mathcal{E}_\perp} V_2, J_2, \mu_2, s, a, s')]\| \\
&\quad + 2B_G d_{TV}(\nu^{\pi_{\theta_1}, \mu_1} \otimes \pi_{\theta_1} \otimes \mathcal{P}_{\mu_1}, \nu^{\pi_{\theta_2}, \mu_2} \otimes \pi_{\theta_2} \otimes \mathcal{P}_{\mu_2}). \qquad (61)
\end{aligned}
$$

The first term of (61) can be bounded in a manner similar to (60). For any $s, a, s'$, we have

$$
\begin{aligned}
&\|G(\Pi_{\mathcal{E}_\perp} V_1, J_1, \mu_1, s, a, s') - G(\Pi_{\mathcal{E}_\perp} V_2, J_2, \mu_2, s, a, s')\| \\
&\leq \|(r(s,a,\mu_1) - J_1 + \Pi_{\mathcal{E}_\perp} V_1(s') - \Pi_{\mathcal{E}_\perp} V_1(s))e_s \\
&\qquad\qquad - (r(s,a,\mu_2) - J_2 + \Pi_{\mathcal{E}_\perp} V_2(s') - \Pi_{\mathcal{E}_\perp} V_2(s))e_s\| \\
&\quad + c_J |r(s,a,\mu_1) - J_1 - r(s,a,\mu_2) + J_2| \\
&\leq |r(s,a,\mu_1) - r(s,a,\mu_2)| \|e_s\| + |J_1 - J_2| \|e_s\| + 2\|\Pi_{\mathcal{E}_\perp} V_1 - \Pi_{\mathcal{E}_\perp} V_2\| \|e_s\| \\
&\quad + c_J |r(s,a,\mu_1) - r(s,a,\mu_2)| + c_J |J_1 - J_2| \\
&\leq (c_J + 1)|r(s,a,\mu_1) - r(s,a,\mu_2)| + (c_J + 1)|J_1 - J_2| + 2\|\Pi_{\mathcal{E}_\perp} V_1 - \Pi_{\mathcal{E}_\perp} V_2\| \\
&\leq (c_J + 1)L|\mu_1 - \mu_2| + (c_J + 1)|J_1 - J_2| + 2\|\Pi_{\mathcal{E}_\perp} V_1 - \Pi_{\mathcal{E}_\perp} V_2\|. \qquad (62)
\end{aligned}
$$

Plugging (62) into (61), we get

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

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

Finally, again following steps similar to (59) we can show

$$\begin{aligned}&\|\bar{H}(\theta_1,\mu_1) - \bar{H}(\theta_2,\mu_2)\| \\ &\leq \|\mathbb{E}_{s\sim\nu^{\pi_{\theta_1},\mu_1}}[H(\mu_1,s) - H(\mu_2,s)]\| + 2B_H d_{TV}(\nu^{\pi_{\theta_1},\mu_1}, \nu^{\pi_{\theta_2},\mu_2}).\end{aligned} \tag{63}$$

From the definition of $H(\mu, s)$ in (24), we have for any $s$

$$\|H(\mu_1,s) - H(\mu_2,s)\| = \|(e_s - \mu_1) - (e_s - \mu_2)\| = \|\mu_1 - \mu_2\|. \tag{64}$$

By Assumption 2,

$$\begin{aligned}d_{TV}(\nu^{\pi_{\theta_1},\mu_1}, \nu^{\pi_{\theta_2},\mu_2}) &= \frac{1}{2}\|\nu^{\pi_{\theta_1},\mu_1} - \nu^{\pi_{\theta_2},\mu_2}\|_1 \\ &\leq L(\|\pi_{\theta_1} - \pi_{\theta_2}\| + \|\mu_1 - \mu_2\|) \\ &\leq L(\|\theta_1 - \theta_2\| + \|\mu_1 - \mu_2\|),\end{aligned} \tag{65}$$

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

Plugging (64) and (65) into (63), we have

$$\begin{aligned}\|\bar{H}(\theta_1,\mu_1) - \bar{H}(\theta_2,\mu_2)\| &\leq \|\mathbb{E}_{s\sim\nu^{\pi_{\theta_1},\mu_1}}[H(\mu_1,s) - H(\mu_2,s)]\| + 2B_H d_{TV}(\nu^{\pi_{\theta_1},\mu_1}, \nu^{\pi_{\theta_2},\mu_2}) \\ &\leq \|\mu_1 - \mu_2\| + L(\|\theta_1 - \theta_2\| + \|\mu_1 - \mu_2\|) \\ &\leq (L+1)(\|\theta_1 - \theta_2\| + \|\mu_1 - \mu_2\|).\end{aligned} \tag{66}$$

□

### E.4. Proof of Lemma 4

By the definition $\Delta f_k$,

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

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 relation

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

Similarly, by the definition of $\Delta g_k$, 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 $\Delta h_k$, 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 & 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

$$s_{k-\tau_k} \xrightarrow{\theta_{k-\tau_k}} a_{k-\tau_k} \xrightarrow{\hat{\mu}_{k-\tau_k}} s_{k-\tau_k+1} \xrightarrow{\theta_{k-\tau_k+1}} a_{k-\tau_k+1} \xrightarrow{\hat{\mu}_{k-\tau_k+1}} \cdots s_{k-1} \xrightarrow{\theta_{k-1}} a_{k-1} \xrightarrow{\hat{\mu}_{k-1}} s_k. \tag{67}$$

We construct an auxiliary Markov chain generated under a constant control

$$s_{k-\tau_k} \xrightarrow{\theta_{k-\tau_k}} a_{k-\tau_k} \xrightarrow{\hat{\mu}_{k-\tau_k}} \widetilde{s}_{k-\tau_k+1} \xrightarrow{\theta_{k-\tau_k}} \widetilde{a}_{k-\tau_k+1} \xrightarrow{\hat{\mu}_{k-\tau_k}} \cdots \widetilde{s}_{k-1} \xrightarrow{\theta_{k-\tau_k}} \widetilde{a}_{k-1} \xrightarrow{\hat{\mu}_{k-\tau_k}} \widetilde{s}_k \tag{68}$$

Let $\widetilde{\mu}$ denote the stationary distribution of state, action, and next state under (68). 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

$$\mathbb{E}[\langle \Delta f_k, F(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{F}(\theta_k, \hat{V}_k, \hat{\mu}_k)\rangle] = T_1 + T_2 + T_3 + T_4. \tag{69}$$

We bound the terms individually. First, we treat $T_1$

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

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

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 relation $\alpha_k \leq \xi_k \leq \beta_k \leq \lambda_k$ 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 $\theta_k$, $\hat{V}_k$, 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 $T_2$

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

where the last inequality follows from the definition of TV distance in (17).

Applying Lemma B.2 from Wu et al. (2020), we then have

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

where the third inequality is a result of (18), 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 (68) at time $k$ and its stationary distribution. To bound $T_3$,

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

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

Finally, we bound the term $T_4$

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

where the third inequality applies the result in (26).

Collecting the bounds on $T_1$-$T_4$ and plugging them into (69), we get

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

□

### 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 3.

□

### 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

$$s_{k-\tau_k} \xrightarrow{\theta_{k-\tau_k}, \hat{\mu}_{k-\tau_k}} s_{k-\tau_k+1} \xrightarrow{\theta_{k-\tau_k+1}, \hat{\mu}_{k-\tau_k+1}} \cdots s_{k-1} \xrightarrow{\theta_{k-1}, \hat{\mu}_{k-1}} s_k. \tag{70}$$

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 $s_{k-\tau_k}$.

$$s_{k-\tau_k} \xrightarrow{\theta_{k-\tau_k}, \hat{\mu}_{k-\tau_k}} \widetilde{s}_{k-\tau_k+1} \xrightarrow{\theta_{k-\tau_k}, \hat{\mu}_{k-\tau_k}} \cdots \widetilde{s}_{k-1} \xrightarrow{\theta_{k-\tau_k}, \hat{\mu}_{k-\tau_k}} \widetilde{s}_k. \tag{71}$$

Defining

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

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

we see that

$$\mathbb{E}[\langle \Delta h_k, e_{s_k} - \mathbb{E}_{s\sim\nu^{\pi_{\theta_k},\hat{\mu}_k}}[e_s]\rangle] = T_1 + T_2 + T_3 + T_4. \tag{72}$$

We bound the terms individually. First, we treat $T_1$

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

where the last inequality follows from the step size relation $\alpha_k \leq \xi_k \leq \lambda_k$ 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 $\theta_k$ 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 $\widetilde{p}_k(s) = \mathbb{P}(\widetilde{s}_k = s)$.

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

where the third inequality follows from the definition of TV distance in (17), and the fourth and fifth inequalities are a result of (18).

The term $T_3$ is proportional to the distance between the distribution of the auxiliary Markov chain (71) at time $k$ and its stationary distribution. Let $\widetilde{\mu}$ denote the stationary distribution of (71). We can bound this term as follows under Assumption 1

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

$$\leq 2B_H\alpha_k,$$

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

The term $T_4$ can be treated by the Lipschitz continuity of $\nu$

$$\begin{aligned}
T_4 &= \mathbb{E}[\langle \Delta h_{k-\tau_k}, \mathbb{E}_{s\sim\nu^{\pi_{\theta_{k-\tau_k}},\hat{\mu}_{k-\tau_k}}}[e_s] - \mathbb{E}_{s\sim\nu^{\pi_{\theta_k},\hat{\mu}_k}}[e_s]\rangle] \\
&\leq 2B_H\mathbb{E}[\|\nu^{\pi_{\theta_{k-\tau_k}},\hat{\mu}_{k-\tau_k}} - \nu^{\pi_{\theta_k},\hat{\mu}_k}\|] \\
&\leq 2B_HL\mathbb{E}[\|\pi_{\theta_k} - \pi_{\theta_{k-\tau_k}}\|] + 2B_H\delta\mathbb{E}[\|\hat{\mu}_{k-\tau_k} - \hat{\mu}_k\|] \\
&\leq 2B_HL\sum_{t=k-\tau_k}^{k}\mathbb{E}[\|\alpha_t f_t\|] + 2B_HL\sum_{t=k-\tau_k}^{k}\mathbb{E}[\|\xi_t h_t\|] \\
&\leq 2B_HL\tau_k\left(B_F\alpha_{k-\tau_k} + B_H\xi_{k-\tau_k}\right) \\
&\leq 2LB_FB_H^2\xi_{k-\tau_k}
\end{aligned}$$

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 (72), we get

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

□

### E.10. Proof of Lemma 10

By the definition of operators $G^V$ and $G^J$ in (24), 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}(\cdot|s,a,\mu)}[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}(\cdot|s,a,\mu)}\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}(\cdot|s,a,\mu)}\left[e_s(e_{s'} - e_s)^\top\right]\Pi_{\mathcal{E}_\perp}(V - V^{\pi_\theta,\mu})\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}(\cdot|s,a,\mu)}\left[e_s(e_{s'} - e_s)^\top\right]\Pi_{\mathcal{E}_\perp}(V - V^{\pi_\theta,\mu}) \\
&\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}(\cdot|s,a,\mu)}\left[e_s(e_{s'} - e_s)^\top\right]\Pi_{\mathcal{E}_\perp}(V - V^{\pi_\theta,\mu}) \\
&\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

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

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

$$s_{k-\tau_k}\xrightarrow{\theta_{k-\tau_k}}a_{k-\tau_k}\xrightarrow{\hat{\mu}_{k-\tau_k}}s_{k-\tau_k+1}\xrightarrow{\theta_{k-\tau_k+1}}a_{k-\tau_k+1}\xrightarrow{\hat{\mu}_{k-\tau_k+1}}\cdots s_{k-1}\xrightarrow{\theta_{k-1}}a_{k-1}\xrightarrow{\hat{\mu}_{k-1}}s_k. \tag{73}$$

We construct an auxiliary Markov chain generated under a constant control

$$s_{k-\tau_k}\xrightarrow{\theta_{k-\tau_k}}a_{k-\tau_k}\xrightarrow{\hat{\mu}_{k-\tau_k}}\widetilde{s}_{k-\tau_k+1}\xrightarrow{\theta_{k-\tau_k}}\widetilde{a}_{k-\tau_k+1}\xrightarrow{\hat{\mu}_{k-\tau_k}}\cdots\widetilde{s}_{k-1}\xrightarrow{\theta_{k-\tau_k}}\widetilde{a}_{k-1}\xrightarrow{\hat{\mu}_{k-\tau_k}}\widetilde{s}_k \tag{74}$$

Let $\widetilde{\mu}$ denote the stationary distribution of state, action, and next state under (74). 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

$$\mathbb{E}[\langle\Delta g_k,G(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k,s_k,a_k,s_{k+1})-\bar{G}(\theta_k,\hat{V}_k,\hat{J}_k,\hat{\mu}_k)\rangle\rangle]=T_1+T_2+T_3+T_4. \tag{75}$$

We bound the terms individually. First, we treat $T_1$

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

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 relation $\alpha_k\leq\xi_k\leq\beta_k\leq\lambda_k$ 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 $\theta_k$, $\hat{V}_k$, 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 $T_2$

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

where the last inequality follows from the definition of TV distance in (17).

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 2, 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 (74) at time $k$ and its stationary distribution. To bound $T_3$,

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

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

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 (26).

Collecting the bounds on $T_1$-$T_4$ and plugging them into (75), we get

$$\mathbb{E}[\langle \Delta g_k, G(\theta_k, \hat{V}_k, \hat{\mu}_k, s_k, a_k, s_{k+1}) - \bar{G}(\theta_k, \hat{V}_k, \hat{\mu}_k)\rangle]$$

$$
\begin{aligned}
&= T_1 + T_2 + T_3 + T_4 \\
&\leq 12 L_F B_F B_G^2 B_H \tau_k \lambda_{k-\tau_k} + (4L + 2|\mathcal{A}|) B_F B_G^2 B_H \tau_k^2 \lambda_{k-\tau_k} + 2 B_G^2 \alpha_k + 4 L_{TV} B_F B_G^2 B_H \xi_{k-\tau_k} \\
&\leq (22L + 2|\mathcal{A}|) L_F L_{TV} B_F B_G^2 B_H \tau_k^2 \lambda_{k-\tau_k}.
\end{aligned}
$$

□

## F. Details for Example 1

We first prove that the mentioned class of MFGs satisfies Assumption 4 with $\rho = 1$ and $\Delta = 0$. Specifically, we need to show

$$
J(\pi', \mu^\star(\pi)) - J(\pi', \mu^\star(\pi')) \leq J(\pi, \mu^\star(\pi)) - J(\pi', \mu^\star(\pi)). \tag{76}
$$

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

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

$$
J(\pi', \mu^\star(\pi)) = \langle \nu^{\pi'}, r(\cdot, \nu^\pi) \rangle = \sum_{s \in \{s_1, s_2\}} \nu^{\pi'}(s) \nu^\pi(s). \tag{77}
$$

Similarly, we have

$$
J(\pi, \mu^\star(\pi)) = \sum_{s \in \{s_1, s_2\}} \Big(\nu^\pi(s)\Big)^2, \quad J(\pi', \mu^\star(\pi')) = \sum_{s \in \{s_1, s_2\}} \big(\nu^{\pi'}(s)\Big)^2
$$

As a result,

$$
\begin{aligned}
J(\pi', \mu^\star(\pi) - J(\pi', \mu^\star(\pi')) &= \sum_{s \in \{s_1, s_2\}} \nu^{\pi'}(s) \Big(\nu^\pi(s) - \nu^{\pi'}(s)\Big), \\
J(\pi, \mu^\star(\pi) - J(\pi', \mu^\star(\pi)) &= \sum_{s \in \{s_1, s_2\}} \nu^{\pi}(s) \Big(\nu^\pi(s) - \nu^{\pi'}(s)\Big).
\end{aligned}
$$

This obvious leads to (76) as

$$
\Big(J(\pi, \mu^\star(\pi) - J(\pi', \mu^\star(\pi))\Big) - \Big(J(\pi', \mu^\star(\pi) - J(\pi', \mu^\star(\pi'))\Big) = \sum_{s \in \{s_1, s_2\}} \Big(\nu^\pi(s) - \nu^{\pi'}(s)\Big)^2 \geq 0.
$$

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. 3.

Under any policy $\pi$, the transition matrix is

$$
P^\pi = \begin{bmatrix} p\pi(a_1 \mid s_1) + (1-p)\pi(a_2 \mid s_1) & p\pi(a_1 \mid s_2) + (1-p)\pi(a_2 \mid s_2) \\ (1-p)\pi(a_1 \mid s_1) + p\pi(a_2 \mid s_1) & (1-p)\pi(a_1 \mid s_2) + p\pi(a_2 \mid s_2) \end{bmatrix},
$$

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

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

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.

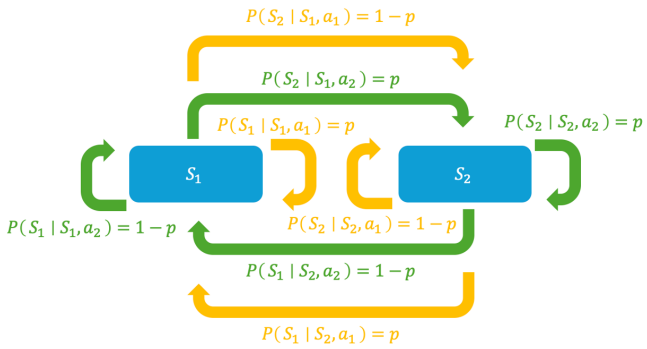


*Figure 3.* Example Mean Field Game Transition

## 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 $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$, and reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$. The cumulative reward collected by a policy $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ is denoted by $J_{\text{MDP}}(\pi)$

$$J_{\text{MDP}}(\pi) \triangleq \mathbb{E}_{a_k \sim \pi(\cdot|s_k), s_{k+1} \sim \mathcal{P}(\cdot|s_k, a_k)}[\textstyle\sum_{k=0}^{\infty} r(s_k, a_k) \mid s_0 = s]. \tag{78}$$

The policy optimization objective under softmax parameterization is

$$\max_{\theta} \quad J_{\text{MDP}}(\pi_\theta). \tag{79}$$

The differential value function under policy $\pi_\theta$ is

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

We use $P^\pi$ and $\nu^\pi$ to denote the transition probability matrix and the stationary distribution of states under the control of $\pi$. The policy gradient is

$$\nabla_\theta J_{\text{MDP}}(\pi_\theta) = \mathbb{E}_{s \sim \nu^{\pi_\theta}, a \sim \pi_\theta(\cdot|s), s' \sim \mathcal{P}(\cdot|s,a,\mu)} \Big[ (r(s, a) + V_{\text{MDP}}^{\pi_\theta}(s') - V_{\text{MDP}}^{\pi_\theta}(s)) \nabla_\theta \log \pi_\theta(a \mid s) \Big], \tag{80}$$

and $V_{\text{MDP}}^{\pi}$ satisfies the Bellman equation

$$V_{\text{MDP}}^{\pi_\theta} = \sum_a \pi_\theta(a \mid \cdot) r(\cdot, a) + J_{\text{MDP}}(\pi_\theta) \mathbf{1}_{|\mathcal{S}|} + (P^{\pi_\theta})^\top V_{\text{MDP}}^{\pi_\theta}. \tag{81}$$

The algorithm for optimizing $J_{\text{MDP}}$ 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 $\theta_k$ 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 (81) and (79) using stochastic approximation.

## H. Simulation Details

We choose the reward function to be

$$r(s, a, \mu) = \mu(s) + \omega_r(s, a) * 0.1, \quad \forall s, a,$$

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

The transition kernel $\mathcal{P}$ is also randomly generated such that for all $s, a$

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

**Algorithm 2** Online Actor Critic Algorithm for Average-Reward MDP

1: **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 $f_0 = 0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $g_0^V = 0 \in \mathbb{R}^{|\mathcal{S}|}$, $g_0^J = 0 \in \mathbb{R}$

2: **Sample:** initial state $s_0 \in \mathcal{S}$ randomly

3: **for** iteration $k = 0, 1, 2, ...$ **do**

4: Take action $a_k \sim \pi_{\theta_k}(\cdot \mid s_k)$. Observe reward $r(s_k, a_k)$ and next state $s_{k+1} \sim \mathcal{P}(\cdot \mid s_k, a_k)$

5: Policy (actor) update:

$$\theta_{k+1} = \theta_k + \alpha_k f_k.$$

6: Value function (critic) update:

$$\hat{V}_{k+1} = \Pi_{B_V}(\hat{V}_k + \beta_k g_k^V), \quad \hat{J}_{k+1} = \Pi_{[0,1]}(\hat{J}_k + \beta_k g_k^J).$$

7: Gradient/Operator estimate update:

$$\begin{aligned} f_{k+1} &= (1-\lambda_k) f_k + \lambda_k (r(s_k, a_k) + \hat{V}_k(s_{k+1})) \nabla \log \pi_{\theta_k}(a_k \mid s_k), \\ g_{k+1}^V &= (1-\lambda_k) g_k^V + \lambda_k (r(s_k, a_k) - \hat{J}_k + \hat{V}_k(s_{k+1}) - \hat{V}_k(s_k)) e_{s_k} \\ g_{k+1}^J &= (1-\lambda_k) g_k^J + \lambda_k c_J (r(s_k, a_k) - \hat{J}_k). \end{aligned}$$

8: **end for**

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 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.