Distributed Offline Policy Optimization Over Batch Data

Han Shen[⋆] **Songtao Lu**[†] **Xiaodong Cui**[†] **Tianyi Chen**[⋆]
[⋆] Rensselaer Polytechnic Institute
[†] IBM T. J. Watson Research Center

Abstract

Federated learning (FL) has received increasing interests during the past years, However, most of the existing works focus on supervised learning, and federated learning for sequential decision making has not been fully explored. Part of the reason is that learning a policy for sequential decision making typically requires repeated interaction with the environments, which is costly in many FL applications. To overcome this issue, this work proposes a federated offline policy optimization method abbreviated as FedOPO that allows clients to jointly learn the optimal policy without interacting with environments during training. Albeit the nonconcave-convex-strongly concave nature of the resultant max-min-max problem, we establish both the local and global convergence of our FedOPO algorithm. Experiments on the OpenAI gym demonstrate that our algorithm is able to find a near-optimal policy while enjoying various merits brought by FL, including training speedup and improved asymptotic performance.

1 Introduction

Federated Learning (FL) is a machine learning setting where clients collaboratively train a model under the coordination of a central server while keeping their data private (McMahan et al., 2017). FL was motivated by the growing need of training with the massive amount of data generated at different local devices, while mitigating the privacy risks and costs resulting from centralized training.

In recent years, FL has achieved tremendous success in numerous applications such as healthcare (Chen et al., 2020), finance (Liu et al., 2020b), IoT (Zhang et al., 2021b), product personalization (Hard et al., 2019). Although FL has proved to be an effective paradigm for a wide range of real-world tasks, existing works have been largely focusing on supervised learning settings, while applying other machine learning techniques in FL paradigm, such as reinforcement learning (RL), has not been widely studied (et al., 2021). Part of the reason is that learning a policy in RL typically requires frequent deployment of new policies in the environment to acquire online experiences. However, in many potential applications of federated RL such as healthcare (Murphy et al., 2001), finance (Liu et al., 2020b) and energy management (Li et al., 2021), deployment of new policy is costly or even impossible (Matsushima et al., 2020). For example, in healthcare, an unreliable treatment policy may have side-effects on patients. Or in power systems, a sub-optimal energy management policy may result in severe economic loss, and is not favored by end users.

To overcome this challenge, we propose an offline federated policy optimization method named FedOPO, which allows multiple clients to jointly train a policy with data sets distributed over local clients. The data set of each client is collected by unknown client-customized behavior polices prior to the training phase, and no data collection is needed during training. Therefore, FedOPO requires no online sampling and thus mitigates the costs of online policy deployment. Moreover, we show that compared to the case where each client trains a policy locally, FL brings several theoretical merits including training speedup, improved asymptotic performance and better data coverage.

1.1 Related works

To put our work in context, we review prior art that we group in the following categories.

Federated learning. Ever since the introduction of FL in (Konečný et al., 2016; McMahan et al., 2017), it has been extensively studied in semi-supervised learning setting (Papernot et al., 2017, 2018); hierarchical setting (Liu et al., 2020a; Briggs et al., 2020); non-iid data setting (Zhao et al., 2018; Hsieh et al., 2020); continual learning(Yoon et al., 2021); multi-task setting (Smith et al., 2018). The major concerns in FL include the communication efficiency (Seide et al., 2014; Stich et al., 2018; Chen et al., 2018; Wang and Joshi, 2018; Yu et al., 2019) and data privacy (Huang et al., 2021; Jin et al., 2021). For a complete survey of FL, see e.g.,

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

(et al., 2021). Though most of the works on FL focus on the supervised learning settings, there are a few works that study a restrictive class of RL problems (see e.g., (Liu et al., 2019a; Lee and Choi, 2020)). However, they are restricted on specific tasks that allow online sampling. Recently, Jin et al. (2022) studied a general federated RL method in an *online* setting. A provably-convergent offline federated RL method that can be applied to a wider range of tasks is, however, missing in the literature.

Model-based offline RL. Offline RL has gained growing interests recently thanks to its importance in safety-critical applications. Offline RL can be tackled either through model-based or model-free approaches. The model-based methods leverage the techniques from supervised learning and uncertainty quantification to learn a reliable Markov decision process (MDP) model, and then utilize the planning algorithms to solve the problem. MOPO regularizes the learnt MDP by penalizing the reward function with an uncertainty term (Tu et al., 2020). MOReL strictly discourages the target policy from visiting the uncertain domains via reward shaping (Kidambi et al., 2020). While COMBO, inspired by (Kumar et al., 2020), takes another approach by conservatively updating the Q function in out-of-support domain without needing uncertainty quantification (Yu et al., 2021). Recently, representation learning-based approaches have also been developed in (Lee et al., 2021).

Model-free offline RL. A closer line of research to our work is the model-free offline RL methods that use conservative policy updates either via incorporating implicit regularization into objective function or imposing explicit constraints on the problem. Such methods include the conservative Q-learning (Kumar et al., 2020); Q-learning with uncertainty quantification (Kumar et al., 2019; Siegel et al., 2020; Wu et al., 2019) or explicit constraints on the state-action domain (Liu et al., 2020c); the importance weighted offline policy gradient (PG) methods (Nachum et al., 2019b; Liu et al., 2019b; Imani et al., 2018). Many offline PG methods originate from the logged actor-critic (Off-PAC) algorithm (Degris et al., 2012). The policy gradient in off-PAC is computed with logged samples, which renders Off-PAC not provably convergent to the optimal policy. To resolve this issue, a popular approach is to reweight the logged update with correction ratios, see e.g., (Imani et al., 2018; Gelada and Bellemare, 2019; Zhang et al., 2019b; Liu et al., 2019b, 2018). However, these works all require the information on the behavior policy, which is often unknown in practice.

In another line of work, (Nachum et al., 2019a; Zhang et al., 2020) considered the offline *behavior-agnostic* setting and proposed offline policy evaluation methods, which were unified as the DICE family (Yang et al., 2020). With similar techniques, Nachum et al. (2019b) proposed the first offline *behavior-agnostic* policy optimization termed AlgaeDICE. AlgaeDICE improves over previous works by solving the distribution mismatch issue in *behavior-agnostic* setting.

1.2 Main contributions

In this context, we propose an offline federated policy optimization method that we term FedOPO. Our contributions can be summarized as follows.

C1) A new federated learning framework. We broaden the application of FL from supervised learning to sequential decision making, where the goal is to learn a policy that repeatedly takes actions based on states. We are interested in settings where online sampling is prohibited, and multiple clients aim to learn the optimal policy with locally distributed logged data generated by behavior polices different from the updating policy. Our algorithm corrects the offline policy update with the so-called density ratio, which is learnt by optimizing a max-min-max objective function with clients' data. Due to the offline nature of our method, it can be applied to a general class of federated RL tasks, including those that require little to none online policy deployment.

C2) Quantifiable benefits of FedOPO. We show both theoretically and empirically that the marriage of FL and offline PG brings two major benefits: 1) *Run-time speedup*. The synchronized parallel computing architecture inherent to FL speeds up the optimization process, which is in dire need as the max-min-max objective is particularly hard to optimize in practice. Our analysis shows that linear speedup is achieved, i.e. the convergence rate increases at a rate of $\Theta(N)$, where N is the number of clients. 2) *Better asymptotic performance*. FL is able to utilize the combined information provided by all clients while protecting data privacy. The shared data collection reduces the statistical error at a rate of $\Theta(1/N)$, thus improves the asymptotic performance. Numerical experiments are provided to verify our theoretical results.

C3) Improved analysis of offline PG. By setting $N=1$, our analysis reduces to the offline PG method. In this case, our work improves over the analysis of offline PG with density-ratio correction in (Huang and Jiang, 2021) in the following aspects: i) The algorithm in (Huang and Jiang, 2021) runs a *double-loop* manner where the inner-loop performs density ratio estimation and the outer loop performs the PG update. While our analysis allows a practical *single-loop* structure where the PG update and density ratio update are performed simultaneously. Albeit the single-loop structure is more difficult to analyze and often has slower convergence rate, we give an improved analysis that achieves the state-of-art rate of $\mathcal{O}(\frac{1}{\sqrt{NK}})$. ii) The Assumption C in (Huang and Jiang, 2021) is stronger than Assumption 1 in this work. Specifically, we show that the second bounded ratio assumption adopted in (Huang and Jiang, 2021) is not needed, relaxing the requirement on the batch data.

2 Preliminaries

In this section, we will define basic notations regarding the Markov decision process and then give a general formulation of the RL problem.

RL problems are often modeled as an MDP described by $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma\}$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}(s'|s,a)$ is the probability of transitioning to $s' \in \mathcal{S}$ given current state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, and $r(s,a)$ is the reward associated with the state-action pair (s,a), and $\gamma \in [0,1)$ is a discount factor. Without loss of generality, we assume the reward $r(s,a) \in [0,1]$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is defined as a mapping from the state space $\mathcal{S}$ to the probability distribution over the action space $\mathcal{A}$.

Considering discrete time t in an infinite horizon, a policy π generates a trajectory $(s_0, a_0, s_1, a_1, \ldots)$ with $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$. Given a policy π, we define the state and state-action value functions as

$$V_\pi(s) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s\right];$$
$$Q_\pi(s,a) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a\right] \quad (1)$$

where $\mathbb{E}$ is taken over the trajectory $(s_0, a_0, s_1, a_1, \ldots)$ generated under policy π. With the above definitions, the advantage function is $A_\pi(s,a) \coloneqq Q_\pi(s,a) - V_\pi(s)$. With ρ denoting the initial state distribution, the discounted state visitation measure induced by policy π is defined as $d_\pi(s) \coloneqq (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s \mid s_0 \sim \rho, \pi)$. We overload the notation and define the discounted state action visitation measure as $d_\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}(s_t = s \mid s_0 \sim \rho, \pi)\pi(a|s)$. In the case where π is parametrized by θ, we use d_θ as shorthand notations for d_{π_θ}.

The goal of RL is to find an optimal policy π^* defined as $\pi^* \in \arg\max_\pi J(\pi) \coloneqq (1-\gamma)\mathbb{E}_{s\sim\rho}[V_\pi(s)]$. We define the optimal return as $J^* \coloneqq \max_\pi J(\pi)$. When the state and action spaces are large, finding the optimal policy π becomes computationally intractable. To overcome the inherent difficulty of learning a function, the policy gradient methods search the best performing policy over a class of parametrized policies. We parametrize the policy with $\theta \in \mathbb{R}^d$, and solve the following problem

$$\max_{\theta\in\mathbb{R}^d} J(\theta) \coloneqq (1-\gamma)\mathbb{E}_{s\sim\rho}[V_{\pi_\theta}(s)] = \mathbb{E}_{s,a\sim d_{\pi_\theta}}[r(s,a)]. \quad (2)$$

To penalize degenerate policies or utilize prior knowledge, it is common to augment the objective function with a regularization, given by

$$J_\tau(\theta) \coloneqq J(\theta) + \tau \underbrace{\mathbb{E}_{s\sim\eta_p}\big[-D_{KL}(\pi_p(\cdot|s)|\pi_\theta(\cdot|s))\big]}_{R(\theta):=} \quad (3)$$

where $\tau \geq 0$ is a regularization constant, η_p is a prior state distribution, and π_p is a prior policy. The regularization $R(\theta)$ encourages π_θ to imitate π_p within the support of η_p, incorporating prior knowledge into training. When π_p and η_p are set as uniform distributions, the regularization term is reduced to the relative-entropy regularization widely analyzed in the literature (Agarwal et al., 2020; Bhandari and Russo, 2019; Zhang et al., 2021a). Moreover, the regularization prevents degenerate solutions that can lead to the pitfall of certain policy parametrization (Bhandari and Russo, 2019).

3 FedOPO: A Federated Offline Policy Optimization Algorithm

In this section, we will first derive a tractable objective function for federated offline PG, and then introduce our algorithm FedOPO.

3.1 Federated offline policy optimization

In an offline federated RL setting, we have N clients aiming to learn the optimal policy of the same MDP. In data collection phase prior to training process, each client n uses a possibly *unknown* behavior policy π_β^n to collect batch data. Following the convention in offline RL literature, we abbreviate the behavior visitation distribution $d_{\pi_\beta^n}(s,a)$ as $d_D^n(s,a)$, and define the averaged distribution as $\bar{d}_D(s,a) \coloneqq \frac{1}{N}\sum_{n=1}^N d_D^n(s,a)$.

To find the optimal policy that maximizes $J(\theta)$, θ is updated using the policy gradient given by (Sutton et al., 2000)

$$\nabla J(\theta) = \mathbb{E}_{s,a\sim d_{\pi_\theta}}\left[Q_{\pi_\theta}(s,a)\psi_\theta(s,a)\right], \quad (4)$$

where the score function is defined as $\psi_\theta(s,a) \coloneqq \nabla \log \pi_\theta(a|s)$. For a given target policy π_θ, running the policy gradient (4) is challenging in the offline RL setting, because samples from the distribution d_{π_θ} in (4) cannot be obtained without interacting with the environment via π_θ. A natural thought is to encourage d_{π_θ} to stay close to the more accessible distribution $\bar{d}_D$ which can be estimated by the batch data collected by clients. To this end, we augment $J_\tau(\theta)$ with a regularizer $D_{\mathcal{X}^2}\big(d_{\pi_\theta}||\bar{d}_D\big)$, that is

$$\max_{\theta\in\mathbb{R}^d} F_\lambda(\theta) \coloneqq J_\tau(\theta) - \lambda D_{\mathcal{X}^2}\big(d_{\pi_\theta}||\bar{d}_D\big), \quad (5)$$

where $D_{\mathcal{X}^2}$ is the $\mathcal{X}^2-$ divergence and $\lambda > 0$ is a regularization constant. It is worth noting that the regularizer in (5) serves different purpose than that in (3). The regularizer in (3) prevents degenerate solutions or incorporates prior knowledge, while the one in (5) encourages the on-policy visitation distribution d_{π_θ} to stay close to the averaged logged visitation distribution $\frac{1}{N}\sum_{n=1}^N d_D^n$, and thus encourages conservative policy updates. The relative temperature between the two regularizers is controlled by τ, λ.

To ensure the tractability of our problem, we make the following assumption which is common in previous works on offline RL (Zhang et al., 2020; Nachum et al., 2019a).

Assumption 1 (exploratory federated RL data)**.** *For all eligible θ, if $d_{\pi_\theta}(s,a) > 0$, there exists a client n whose $d_D^n(s,a) > 0$ for this particular pair (s,a). There exists a constant C_d such that $\|\frac{d_{\pi_\theta}}{\bar{d}_D}\|_\infty \le C_d$.*

As we show in the remark below, our assumption is weaker than those in (Nachum et al., 2019b; Huang and Jiang, 2021; Zhang et al., 2020; Nachum et al., 2019a).

Remark 1. *Suppose π_β^n is chosen randomly with at least probability $p > 0$ such that $\|\frac{d_{\pi_\theta}}{d_D^n}\| < \infty$, then it is immediate that Assumption 1 holds with probability at least $1-(1-p)^N$. Then for a federated system with large enough number of workers N, Assumption 1 holds with a probability sufficiently close to 1. Moreover, by definition of d_D^n, we have $d_D^n(s,a) \ge (1-\gamma)\rho(s)\pi_\beta^n(a|s)$. If $\rho(s) > 0$, and we select π_β^n within the subset of stochastic policies, i.e. $\pi_\beta^n(a|s) > 0$ for any (s,a), then we have $d_D^n > 0$ with probability $p=1$. In this case, Assumption 1 is guaranteed to hold. In addition, it is worth noting that our assumption is weaker than that of Huang and Jiang (2021) when $N = 1$, as we do not need the second inequality assumption in (Huang and Jiang, 2021, Assumption C).*

3.2 Federated max-min-max reformulations

Accessing d_{π_θ} in (5) requires sampling via π_θ, which is changing during learning process. Hence, the objective function $F_\lambda(\theta)$ is still hard to optimize in the federated RL setting. Inspired by the change of variable trick and the fenchel duality trick in (Nachum et al., 2019b,a), we will consider an equivalent form of (5) in the following lemma.

Lemma 1. *Under Assumption 1, the problem defined in* (5) *is equivalent to*

$$\max_{\theta\in\mathbb{R}^d}\min_{v\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}}\max_{\mu\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}} F_\lambda(\theta,v,\mu) := \frac{1}{N}\sum_{n=1}^N F_\lambda^n(\theta,v,\mu) \quad (6)$$
$$\text{with}\quad F_\lambda^n(\theta,v,\mu) := (1-\gamma)\mathbb{E}_{\substack{s_0\sim\rho\\ a_0\sim\pi_\theta}}\big[v(s_0,a_0)\big]$$
$$+\mathbb{E}_{s,a\sim d_D^n}\Big[\big(\mathcal{B}_{\pi_\theta}v - v\big)(s,a)\mu(s,a) - \frac{\lambda}{2}\mu(s,a)^2\Big] + \tau R(\theta)$$

where the Bellman operator is defined as $\mathcal{B}_{\pi_\theta}v(s,a) := r(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a),a'\sim\pi_\theta(\cdot|s')}[v(s',a')]$.

We term $F_\lambda(\theta,v,\mu)$ as the **population-level** offline federated RL objective function. By introducing v and μ in (6), our new objective function no longer depends on the inaccessible distribution d_{π_θ}, instead it depends on $\{d_D^n\}_{n=1}^N$ and ρ which can be estimated by clients' logged data. Given θ, define $(v_\theta^*,\mu_\theta^*) := \arg\min_v \max_\mu F_\lambda(\theta,v,\mu)$. Using the optimality condition of $F_\lambda(\theta,v,\mu)$ with respect to $\mu(s,a)$, we have

$$\mu_\theta^*(s,a) = \frac{1}{\lambda}\big(\mathcal{B}_{\pi_\theta}v_\theta^* - v_\theta^*\big)(s,a) = \frac{d_{\pi_\theta}(s,a)}{\frac{1}{N}\sum_{n=1}^N d_D^n(s,a)}. \quad (7)$$

Given $\bar{d}_D$, we can use the so-called density ratio μ_θ^* to obtain the distribution d_{π_θ}. Recalling the definition of $\mathcal{B}_{\pi_\theta}$, we can solve (7) with respect to v_θ^* and obtain

$$v_\theta^*(s,a) = \quad (8)$$
$$\mathbb{E}_{\pi_\theta}\Big[\sum_{t=0}^\infty \gamma^t\Big(r(s_t,a_t) - \lambda\frac{d_{\pi_\theta}}{\bar{d}_D}(s_t,a_t)\Big)\Big|s_0=s,a_0=a\Big].$$

The definition of v_θ^* is akin to the Q-function of a virtual MDP $\widetilde{\mathcal{M}} = \{\mathcal{S},\mathcal{A},\mathcal{P},\tilde{r},\gamma\}$ with reward function $\tilde{r}(s,a) := (r - \lambda\frac{d_{\pi_\theta}}{\bar{d}_D})(s,a)$. Hence, solving v and μ in (6) can be viewed as the *critic* problem for estimating the *actor* gradient. We formalize this intuition next.

Lemma 2. *Suppose the critic variables are optimized, i.e. $v = v_\theta^*$ and $\mu = \mu_\theta^*$, then it holds that*

$$\nabla_\theta F_\lambda(\theta,v_\theta^*,\mu_\theta^*) = \mathbb{E}_{d_{\pi_\theta}}[v_\theta^*(s,a)\psi_\theta(s,a)] + \tau\nabla R(\theta).$$
$$= \mathbb{E}_{d_{\pi_\theta}}[(\mu_\theta^*\cdot v_\theta^*)(s,a)\psi_\theta(s,a)] + \tau\nabla R(\theta) \quad (9)$$

Compared with the policy gradient in (4), (9) is the regularized policy gradient corresponding to the virtual MDP $\widetilde{\mathcal{M}}$. The virtual reward function $\tilde{r}(s,a)$ penalizes π_θ when it visits the less supported state-action space, thus encourages conservative policy updates. Moreover, when λ is chosen properly small, we have that $v_\theta^* \approx Q_{\pi_\theta}$, and therefore $\nabla_\theta F_\lambda(\theta,v_\theta^*,\mu_\theta^*) \approx \nabla J_\tau(\theta)$. In this sense, $\nabla_\theta F_\lambda(\theta,v,\mu)$ is an estimate of the regularized policy gradient given by the critic variable v,μ.

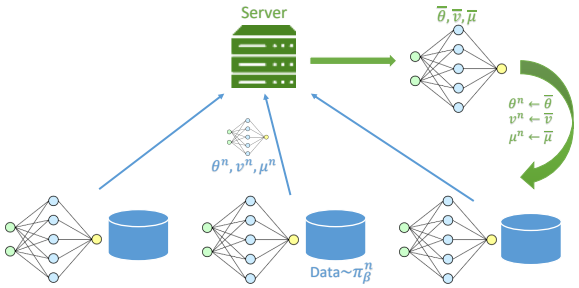

Figure 1: The framework of FedOPO.

3.3 Algorithm development

To optimize $F_\lambda(\theta,v,\mu)$, one needs to sample from $\{d_D^n\}_{n=1}^N$ and ρ. This sampling process can be approximated by sampling from local data sets $\mathcal{D}^n$ and $\mathcal{D}_0^n$, where $\mathcal{D}^n$ contains transition tuples $\{(s_i,a_i,s_i')\}_{i=1}^M$ with each tuple drawn from $d_D^n\otimes\mathcal{P}$, e.g., $(s_i,a_i)\sim d_D^n$ and

Algorithm 1 Federated Offline Policy Optimization

1: **global initialize:** Step sizes α and β. Communication interval I.
2: **client initialize:** Initial policy $\pi_{\theta_k^n}$. Initial v_k^n and μ_k^n parametrized by $\omega_{v,k}^n$ and $\omega_{\mu,k}^n$.
3: **for** $k = 1, 2, \ldots, K$, **do**:
4: **each client** $n \in \{1, 2, \ldots, N\}$ **do**:
5: Obtain samples following (11) and (13).
6: Update the local model via (12) and (15).
7: **if** $k \mod I$:
8: Clients upload local parameters.
9: Server averages local parameters via (16).
10: Broadcast averaged parameters to clients.
11: **end if**
12: **end for**

$s_i' \sim \mathcal{P}(\cdot|s_i, a_i)$, and $\mathcal{D}_0^n$ contains initial state samples $\{s_0^i\}_{i=1}^M \sim \rho$. For simplicity of notations, we assume the data set of every client contains M samples. Denote the empirical distribution of initial state samples in $\mathcal{D}_0^n$ as $\hat{\rho}^n$, and the empirical distribution of state-action pairs in data set $\mathcal{D}^n$ as $\hat{d}_D^n$. Then we can define the **empirical version** of $F_\lambda(\theta, v, \mu)$ as

$$\hat{F}_\lambda(\theta, v, \mu) \coloneqq \frac{1}{N}\sum_{n=1}^N \hat{F}_\lambda^n(\theta, v, \mu) \tag{10}$$

with $\hat{F}_\lambda^n(\theta, v, \mu) \coloneqq (1-\gamma)\mathbb{E}_{\substack{s_0\sim\hat{\rho}^n \\ a_0\sim\pi_\theta}}\big[v(s_0, a_0)\big] + \tau R(\theta) + \mathbb{E}_{\substack{s,a,s'\sim\hat{d}_D^n \\ a'\sim\pi_\theta}}\big[\big(r(s,a) + \gamma v(s', a') - v(s,a)\big)\mu(s,a) - \frac{\lambda}{2}\mu(s,a)^2\big]$.

Due to privacy concerns in FL, clients cannot access each other's data, and thus cannot directly optimize the global objective function $\hat{F}_\lambda$. We introduce FedOPO in Algorihtm 1, and summarize it next.

Local offline critic update. Consider the setting where each client n stores local models θ^n, v^n and μ^n, with v^n and μ^n being parametrized by ω_v^n and ω_μ^n respectively. At each iteration k, client n first uniformly samples transitions from $\mathcal{D}^n$ and initial states from $\mathcal{D}_0^n$. This sampling process can be written as

$$s_0 \sim \hat{\rho}^n, a_0 \sim \pi_{\theta_k^n}(\cdot|s_0); (s, a, s') \sim \hat{d}_D^n, a' \sim \pi_{\theta_k^n}(\cdot|s') \tag{11}$$

where $\pi_{\theta_k^n}$ is its local policy model, and the subscription k, n on samples are omitted for notation simplicity.

With the samples, the client can compute the stochastic gradient of local objective function $\hat{F}_\lambda^n(\theta_k^n, v_k^n, \mu_k^n)$ w.r.t. $\omega_{v,k}^n$ and $\omega_{\mu,k}^n$. We denote the gradients as $\hat{\nabla}_{\omega_v^n}\hat{F}_\lambda^n(\theta_k^n, v_k^n, \mu_k^n)$ and $\hat{\nabla}_{\omega_\mu^n}\hat{F}_\lambda^n(\theta_k^n, v_k^n, \mu_k^n)$, then given step size β, the update of v^n and μ^n can be written as

$$\begin{aligned}\omega_{v,k+1}^n &= \omega_{v,k}^n - \beta\hat{\nabla}_{\omega_v^n}\hat{F}_\lambda^n(\theta_k^n, v_k^n, \mu_k^n),\\ \omega_{\mu,k+1}^n &= \omega_{\mu,k}^n + \beta\hat{\nabla}_{\omega_\mu^n}\hat{F}_\lambda^n(\theta_k^n, v_k^n, \mu_k^n).\end{aligned} \tag{12}$$

Local offline actor update. Before the policy update, we first do the actor sampling via

$$(\tilde{s}, \tilde{a}) \sim \hat{d}_D^n;\ s_p \sim \eta_p, a_p \sim \pi_p(\cdot|s_p). \tag{13}$$

To update policy, a natural thought is to do gradient ascent with an unbiased estimator of $\nabla_{\theta_k^n}\hat{F}_\lambda^n(\theta_k^n, v_{k+1}^n, \mu_{k+1}^n)$. Instead of this natural choice of gradient, we use an estimator based on Lemma 2:

$$\hat{p}_k^n \coloneqq \mu_{k+1}^n(\tilde{s}, \tilde{a})v_{k+1}^n(\tilde{s}, \tilde{a})\psi_{\theta_k^n}(\tilde{s}, \tilde{a}) + \tau\psi_{\theta_k^n}(s_p, a_p) \tag{14}$$

The actor gradient defined in (14) can be viewed as a local estimation of $\nabla_\theta F_\lambda$ defined in (9). By Lemma 2, the expected global average of the two candidates $\nabla_{\theta_k^n}\hat{F}_\lambda^n, \hat{p}_k^n$ behave similarly when the critic is close to the optimal regime. Compared to $\nabla_{\theta_k^n}\hat{F}_\lambda^n$, $\hat{p}_k^n$ requires less sampling and computation is thus preferred. After the client obtains $\hat{p}_k^n$, it then updates its local policy model with step size α:

$$\theta_{k+1}^n = \theta_k^n + \alpha\hat{p}_k^n. \tag{15}$$

Periodic global average. For every I iterations, the server averages clients' model parameters via

$$\bar{\omega}_{v,k+1} = \Pi_{R_v}\Big(\frac{1}{N}\sum_{n=1}^N \omega_{v,k+1}^n\Big) \tag{16}$$

$$\bar{\omega}_{\mu,k+1} = \Pi_{R_\mu}\Big(\frac{1}{N}\sum_{n=1}^N \omega_{\mu,k+1}^n\Big),\ \bar{\theta}_{k+1} = \frac{1}{N}\sum_{n=1}^N \theta_{k+1}^n$$

where Π_R projects a vector to a L_2 ball with radius R. The projection guarantees the stability of our algorithm and has also been adopted in (Nachum et al., 2019a; Huang and Jiang, 2021). The averaging operation intrinsically allows knowledge sharing among clients, making it possible to optimize $\hat{F}_\lambda$.

4 Theoretical Results

In this section, we will first establish the local convergence of FedOPO with the general policy parametrization and its global convergence with the softmax policy parametrization.

4.1 General local convergence result

Due to space limitation, we will provide an error decomposition then present the theoretical results. We defer presentation of the full proof in the supplementary document.

Matrix-vector reformulation. Denote $\phi(\cdot) : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}^{d_1}$ as the feature vector. Suppose v and μ are parametrized linearly, e.g., $v(s,a) = \phi(s,a)^\top\omega_v$ and $\mu(s,a) = \phi(s,a)^\top\omega_\mu$, and define

$$\begin{aligned}A_\theta^n &\coloneqq \mathbb{E}_{s,a,s'\sim\hat{d}_D^n, a'\sim\pi_\theta}\big[(\gamma\phi(s',a') - \phi(s,a))\phi(s,a)^\top\big],\\ b_\theta^n &\coloneqq (1-\gamma)\mathbb{E}_{s_0\sim\hat{\rho}^n, a_0\sim\pi_\theta}[\phi(s_0,a_0)],\\ C^n &\coloneqq \mathbb{E}_{s,a\sim\hat{d}_D^n}[\phi(s,a)\phi(s,a)^\top],\\ h^n &\coloneqq \mathbb{E}_{s,a\sim\hat{d}_D^n}[r(s,a)\phi(s,a)].\end{aligned} \tag{17}$$

Let $A_\theta := \frac{1}{N}\sum_{n=1}^N A_\theta^n$, $b_\theta := \frac{1}{N}\sum_{n=1}^N b_\theta^n$, $C := \frac{1}{N}\sum_{n=1}^N C^n$ and $h := \frac{1}{N}\sum_{n=1}^N h^n$. Then the empirical objective defined in (10) can be written as

$$\hat{F}_\lambda(\theta,\omega_v,\omega_\mu)=b_\theta^\top\omega_v+\omega_v^\top A_\theta\omega_\mu+\omega_\mu^\top h-\frac{\lambda}{2}\omega_\mu^\top C\omega_\mu. \quad (18)$$

We also denote the stationary point of $\hat{F}_\lambda(\theta,\omega_v,\omega_\mu)$ with respect to ω_v and ω_μ as $\hat{\omega}_v^*(\theta)$ and $\hat{\omega}_\mu^*(\theta)$. By the first-order optimality condition, we have

$$A_\theta\hat{\omega}_\mu^*(\theta)+b_\theta=0, -\lambda C\hat{\omega}_\mu^*(\theta)+A_\theta{}^\top\hat{\omega}_v^*(\theta)+h=0. \quad (19)$$

To present our convergence result, we first make the following standard assumptions:

Assumption 2. *For any $(s,a)\in\mathcal{S}\times\mathcal{A}$, the feature vector $\|\phi(s,a)\|_2\le 1$. For any eligible θ, the smallest singular value of the matrix A_θ is lower bounded by $\sigma_{\inf}>0$. Matrix C is positive-definite and its smallest eigenvalue is lower bounded by $\eta>0$.*

This assumption is standard in the literature (Nachum et al., 2019a; Huang and Jiang, 2021). Under Assumption 2, we know from (19) that $\hat{\omega}_v^*(\theta)$ and $\hat{\omega}_\mu^*(\theta)$ are unique and there exist constants R_v, R_μ such that $\|\hat{\omega}_v^*(\theta)\|_2\le R_v$ and $\|\hat{\omega}_\mu^*(\theta)\|_2\le R_\mu$, which also justifies the projection chosen in Algorithm 1.

Assumption 3. *For any $\theta,\theta'\in\mathbb{R}^d$, there exist constants C_ψ, L_ψ and L_π such that: i) $\|\psi_\theta(s,a)\|_2\le C_\psi$; ii) $\|\psi_\theta(s,a)-\psi_{\theta'}(s,a)\|_2\le L_\psi\|\theta-\theta'\|_2$; iii) $|\pi_\theta(a|s)-\pi_{\theta'}(a|s)|\le L_\pi\|\theta-\theta'\|_2$.*

Assumption 3 is common in analyzing policy gradient-type algorithms, which has also been made by e.g., (Zhang et al., 2019a; Agarwal et al., 2020). This assumption holds for many popular policy parametrization methods such as softmax policy (Agarwal et al., 2020), Gaussian policy (Doya, 2000) and Boltzmann policy (Konda and Borkar, 1999).

Error decomposition. To gain more insights into the convergence of FedOPO, we provide the error decomposition of the actor gradient below. With the short-hand notation $\hat{p}_k^n := p(\theta_k^n, v_{k+1}^n, \mu_{k+1}^n)$, we have

$$\begin{aligned}
&\Big\|\frac{1}{N}\sum_{n=1}^N\mathbb{E}[\hat{p}_k^n]-\nabla J_\tau(\bar{\theta}_k)\Big\| \\
&\le \frac{1}{N}\sum_{n=1}^N\mathbb{E}\Big\|p(\theta_k^n,v_{k+1}^n,\mu_{k+1}^n)-p(\bar{\theta}_k,\bar{v}_{k+1},\bar{\mu}_{k+1})\Big\| \\
&\quad+\frac{1}{N}\sum_{n=1}^N\mathbb{E}\Big\|p(\bar{\theta}_k,\bar{v}_{k+1},\bar{\mu}_{k+1})-p(\bar{\theta}_k,v_{\bar{\theta}_k}^*,\mu_{\bar{\theta}_k}^*)\Big\| \\
&\quad+\Big\|\frac{1}{N}\sum_{n=1}^N\mathbb{E}[p(\bar{\theta}_k,v_{\bar{\theta}_k}^*,\mu_{\bar{\theta}_k}^*)]-\nabla J_\tau(\bar{\theta}_k)\Big\| \qquad (20)
\end{aligned}$$

where $\bar{v}_k$ and $\bar{\mu}_k$ are the averaged critic models parametrized by $\bar{\omega}_{v,k}$ and $\bar{\omega}_{\mu,k}$ respectively, and the expectation is taken over the actor samples at iteration k. The first term in (20) is due to the difference between local models and the global model, which can be controlled by communication interval I. The last term is due to the difference between: i) $\hat{d}_D$ and $\bar{d}_D$ (statistical error); and, ii) $v_{\bar{\theta}_k}^*$ and $Q_{\pi_{\bar{\theta}_k}}$ (regularization error). The second term is the *critic error* of $\bar{v}$ and $\bar{\mu}$. Using the critic error of $\bar{v}_k$ as an example, we have

$$\|\bar{v}_{k+1}-v_{\bar{\theta}_k}^*\|\le\|\bar{v}_{k+1}-\hat{v}_{\bar{\theta}_k}^*\|+\|\hat{v}_{\bar{\theta}_k}^*-v_{\bar{\theta}_k}^*\|. \quad (21)$$

Given $\bar{\theta}_k$, $\hat{v}_{\bar{\theta}_k}^*$ is the optimal solution of the empirical objective $\hat{F}_\lambda$ defined in (18). The first term in (21) is the critic optimization error. Different from usual convergence analysis where the optimal solution is stationary, in our case the optimal solution $\hat{v}_{\bar{\theta}_k}^*$ is drifting with the change of $\{\theta_k^n\}_{n=1}^N$ at every iteration k. However, by exploiting the smoothness of $\hat{v}_\theta^*$, we are able to show the relative stationery of $\hat{v}_{\bar{\theta}_k}^*$ and prove the convergence of $\bar{v}$. The second term in (21) is introduced by the difference between (18) and (6), which results in both statistical and function approximation errors.

With the insights provided by the error decomposition above, we are ready to give our local convergence result.

Theorem 1. *Consider Algorithm 1 with linear parametrization of v^n and μ^n. Suppose Assumptions 1-3 hold. Choose $\alpha=\sqrt{N/K}$, and $\beta=\Theta(\alpha)$. Assume $I^4N^3=\mathcal{O}(K)$, then for large enough K, it holds with probability at least $1-\delta$ that*

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^K\mathbb{E}\big\|\nabla J_\tau(\bar{\theta}_k)\big\|_2^2 \qquad (22)\\
&=\mathcal{O}\Big(\frac{1}{\sqrt{NK}}\Big)+\tilde{\mathcal{O}}\Big(\sqrt{\frac{\log\frac{3}{\delta}}{NM}}\Big)+\mathcal{O}\big(\epsilon_{\rm app}\big)+\mathcal{O}\big(\epsilon_\lambda\big)
\end{aligned}$$

where $\epsilon_\lambda := \Big(\lambda\frac{C_\psi C_d}{1-\gamma}\Big)^2$ is the regularization error and $\epsilon_{\rm app}$ is the function approximation error of v^n and μ^n.

The first term in the right hand side of (22) is the error of optimizing the empirical problem (10); the second term corresponds to the statistical error induced by the finite number of samples in the empirical problem (22); the third term $\epsilon_{\rm app}$ is the function approximation error introduced by the limited expressive power of the parametrization of v^n and μ^n; and the last term is the error introduced by the regularization term in (5). We can achieve $\epsilon_{\rm app}=0$ if the optimal solution $\{v_\theta^*,\mu_\theta^*\}$ falls in the span of the features.

In Theorem 1, we consider the optimality of $J_\tau(\theta)$ instead of $F_\lambda(\theta,v,\mu)$ since $F_\lambda(\theta,v,\mu)$ is essentially formulated to facilitate maximizing $J_\tau(\theta)$ in an offline setting. In addition, observe that $\hat{F}_\lambda(\theta,\omega_v,\omega_\mu)$ is generally non-concave with respect to θ, and is convex-strongly concave with respect to ω_v and ω_μ. To our best knowledge, the convergence of

such an objective has not been established in the literature. Therefore, our optimization result is of independent interest.

Remark 2 (linear speedup & reduced statistical error)**.** *Theorem 1 implies a convergence rate of $\mathcal{O}(\frac{1}{\sqrt{NK}})$. As number of clients N increases, the convergence rate grows at a rate of $\Theta(N)$, implying the speedup grows linearly w.r.t. N. Moreover, the statistical error decreases at a rate of $\mathcal{O}(N^{-\frac{1}{2}}M^{-\frac{1}{2}})$, indicating that federated learning also improves asymptotic performance.*

4.2 Global convergence result

Since $J(\theta)$ is a generally non-concave w.r.t. θ, the gradient ascent type algorithm FedOPO can only guarantee local convergence. However, inspired by recent advances on the global convergence of PG (Agarwal et al., 2020; Bhandari and Russo, 2019; Zhang et al., 2019a; Mei et al., 2020), under special policy parametrizations, we are able to show the global convergence of FedOPO along with the benefits like speedup and improved asymptotic performance.

Specifically, we consider the class of MDP which has finite state space and action space. Suppose the policy is parametrized by the a natural softmax policy $\pi_\theta(a|s) = \frac{\exp\theta_{s,a}}{\sum_{s,a}\exp\theta_{s,a}}$, where $\theta_{s,a}$ is the element corresponding to (s,a) pair of the parameter vector $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. It is known that the softmax policy class cannot represent deterministic policies with finite θ. To avoid driving θ to infinity, it is crucial to penalize the deterministic policies with the regularization term $R(\theta)$. To do so, we set $\tau > 0$, and choose the priors η_p and π_p as uniform distribution on the state-action space, i.e. $\eta_p = \frac{1}{|\mathcal{S}|}$ and $\pi_p = \frac{1}{|\mathcal{A}|}$.

Define the feature matrix $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times d_1}$:

$$\Phi := [\phi(s^1,a^1), \phi(s^2,a^2), \ldots, \phi(s^{|\mathcal{S}|},a^{|\mathcal{A}|})]^\top. \quad (23)$$

In the following assumption, we view the optimal solutions μ_θ^* and v_θ^* as vectors in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ space.

Assumption 4 (Linear realizable case)**.** *For any eligible θ, there exist $\tilde{\omega}_\mu^* \in \mathbb{R}^{d_1}$ and $\tilde{\omega}_v^* \in \mathbb{R}^{d_1}$ such that $\Phi\tilde{\omega}_\mu^*(\theta) = \mu_\theta^*$ and $\Phi\tilde{\omega}_v^*(\theta) = v_\theta^*$.*

Assumption 4 ensures the optimal solutions of $F_\lambda(\theta, v, \mu)$ w.r.t. v, μ can be accurately approximated by linear functions. For the assumption to hold, it suffices to select a squared full-rank feature matrix Φ. It is worth noting that when this assumption does not hold, our result in Theorem 2 holds with an extra error term, which is the function approximation error ϵ_{app}.

Theorem 2. *Consider Algorithm 1 with linear parametrization of v^n, μ^n and natural softmax parametrization of policies π_{θ^n}. Suppose Assumptions 1, 2 and 4 hold. Choose $\alpha = \sqrt{N/K}$, and $\beta = \Theta(\alpha)$. Assume $I^4N^3 = \mathcal{O}(K)$, then for large enough K, it holds with probability greater than $1-\delta$ that*

$$J^* - \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[J(\bar{\theta}_k)] = \mathcal{O}\Big(\frac{1}{\sqrt{NK}}\Big) + \tilde{\mathcal{O}}\Big(\sqrt{\frac{\log\frac{3}{\delta}}{NM}}\Big) + \mathcal{O}(\epsilon_\tau) \quad (24)$$

where $\epsilon_\tau := \frac{\tau C_d^2}{(1-\gamma)^2} + \tau\left\|\frac{d_{\pi^*}}{\rho}\right\|_\infty$ is the regularization error.

The first term in (24) corresponds to the optimization error which diminishes at a rate of $\mathcal{O}(\frac{1}{\sqrt{NK}})$. This implies linear speedup as discussed in Remark 2. When $N = 1$, the decay rate of optimization error matches the best-known rate $\tilde{\mathcal{O}}(\frac{1}{\sqrt{K}})$ for stochastic PG (Zhang et al., 2021a). The second term in (24) is the statistical error, which can be reduced by introducing more clients or using larger data sets. The last term is the regularization error.

5 Numerical Experiments

In this section, we provide numerical tests of FedOPO in the OpenAI Gym environments. We empirically demonstrate the advantage of training collaboratively over training locally, and also showcase the training speedup brought by FedOPO. All test results are generated by 5 Monte-carlo runs. The hyperparameters are decided by a grid search.

Generating logged RL data. In all the following tests, the data of client n is generated by its behavior policy: $\pi_\beta^n = \xi_n\pi_{\text{uniform}} + (1-\xi_n)\pi_{\text{expert}}$, where $\xi_n \in [0,1]$ is the mixing factor and π_{expert} is a greedy policy given by the online actor-critic algorithm (Mnih et al., 2016).

We first test the performance of FedOPO and compare it with the performance of its local version and a baseline optimal policy. The optimal policy is given by the online actor-critic algorithm. In these tests, client n's behavior policy is generated with a random mixing constant $\xi_n \sim U(0.1, 0.7)$ (navigation); $\xi_n \sim U(0.3, 0.6)$ (cartpole); $\xi_n \sim U(0, 0.2)$ (frozenlake) respectively. Observing from Figure 2, FedOPO is able to achieve near-optimal performance. Furthermore, training in a federated system has a clear advantage over training locally due to two reasons: 1) Federated learning intrinsically shares the information of each local data set, and therefore can better capture the dynamic of the MDP. This corresponds to the reduced statistical error in our theoretical analysis. 2) As argued in the discussion of Assumption 1, federated learning enables the usage of a collection of diverse behavior polices, and is therefore more likely to have better data coverage.

Then we test the training speedup of FedOPO with different numbers of clients. In these tests, a data set is generated with mixing factor ξ and is then equally distributed to each client. The mixing factor $\xi = 0.4, 0.45, 0.2$ for navigation, cartpole and frozenlake respectively. Observing from Figure 3, federated learning effectively speeds up the training process with a larger number of clients. The key enabler

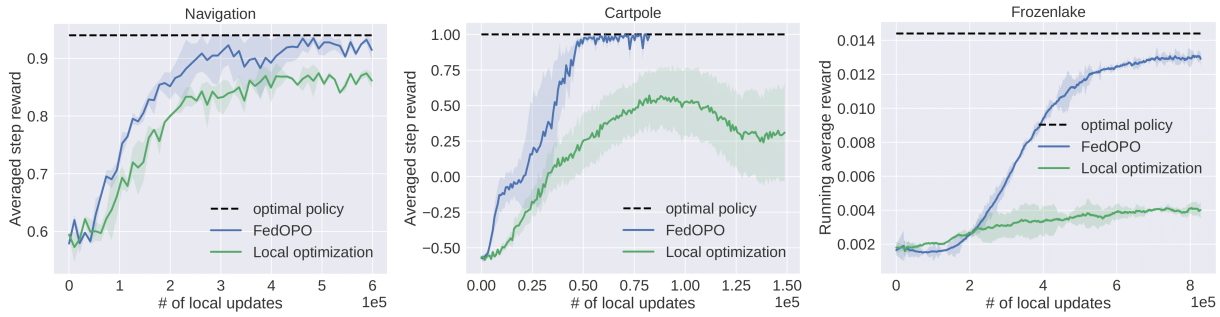

Figure 2: FedOPO compared with local optimization. The baseline *optimal policy* is given by the soft actor-critic method.

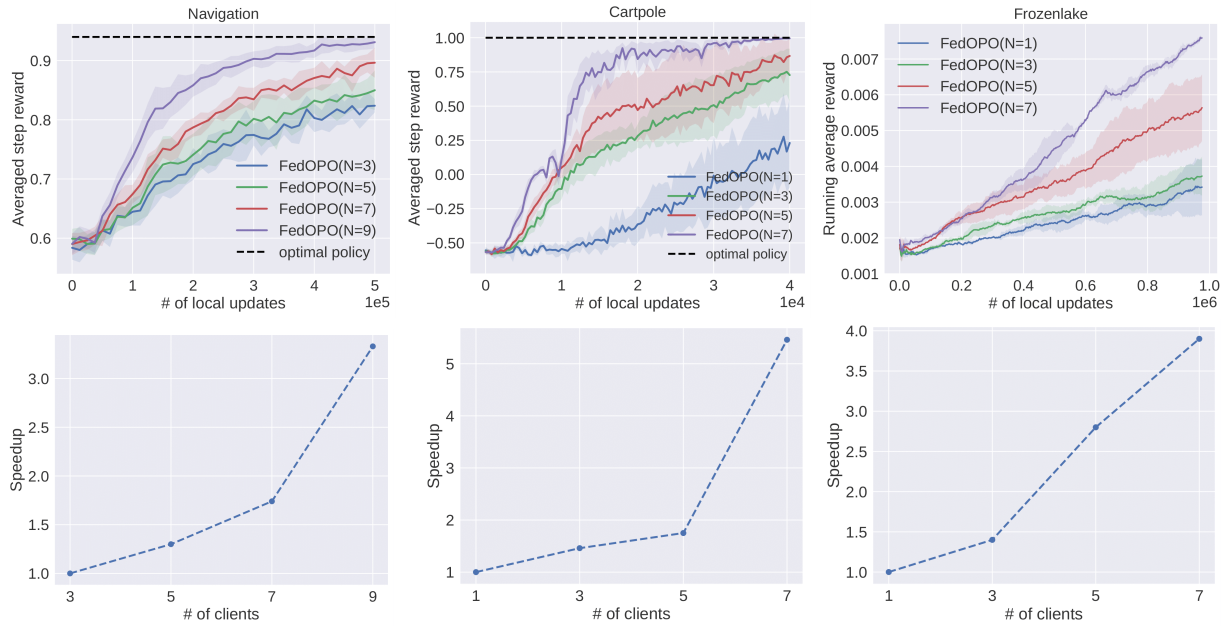

Figure 3: Training speedup test. The speedup plot is given by the iterations to reach a certain reward. The baseline *optimal policy* is given by the soft actor-critic method.

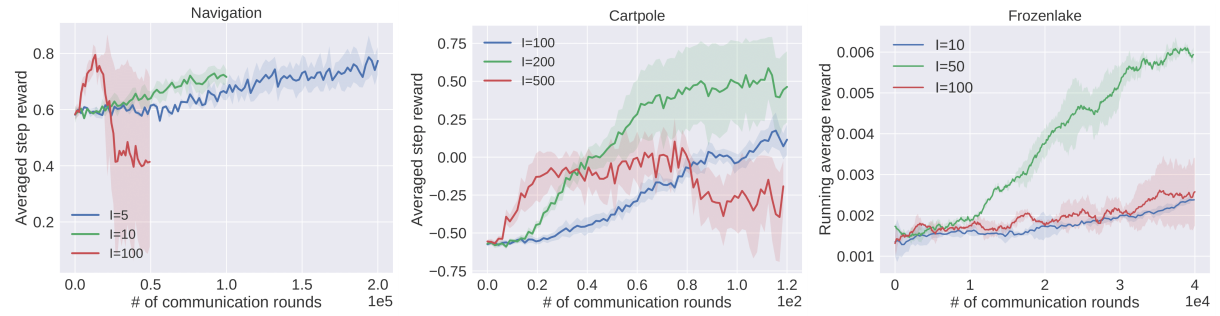

Figure 4: Tests on the communication interval I.

is that more clients result in enlarged effective batch size growing linearly with N, and thus we can safely choose more aggressive step size or gradient clipping bounds.

We also conduct tests on the communication efficiency of our method. Due to space limitation, we defer the results in the supplementary material.

5.1 Tests on communication efficiency

In this section, we provide an additional test on the communication efficiency of our algorithm. We change the communication interval I and observe how our algorithm performs under different I. As shown in Figure 4, an appropriately chosen communication interval can decrease the total number of communication rounds needed to reach

certain reward, while a too greedy choice will hinder the convergence too much, which results in additional communication rounds instead.

6 Conclusions

This paper considers the offline policy optimization in RL from behavior-agnostic batch data that are distributed over a set of clients. Leveraging recent advances on offline policy evaluation, we formulate the problem as a distributed max-min-max problem and propose a federated offline policy optimization algorithm that we term FedOPO. The proposed algorithm allows clients to jointly learn the optimal policy without sharing logged RL data. Albeit the nonconcave-strong convex-concave nature of the problem, we quantify the convergence rate of FedOPO, and establish its global convergence for a class of offline RL problems. We conduct numerical experiments on the OpenAI gym environment to verify its effectiveness.

Acknowledgment

The work of H. Shen and T. Chen was partially supported by and the Rensselaer-IBM AI Research Collaboration (`http://airc.rpi.edu`), part of the IBM AI Horizons Network (`http://ibm.biz/AIHorizons`) and National Science Foundation 2047177.

References

A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *Proc. of Thirty Third Conference on Learning Theory*, 2020.

J. Bhandari and D. Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint:1906.01786*, 2019.

C. Briggs, Z. Fan, and P. Andras. Federated learning with hierarchical clustering of local updates to improve training on non-iid data. In *International Joint Conference on Neural Networks*, 2020.

T. Chen, G. Giannakis, T. Sun, and W. Yin. LAG: Lazily aggregated gradient for communication-efficient distributed learning. In *Proc. of Advances in Neural Information Processing Systems*, 2018.

T. Chen, Y. Sun, and W. Yin. Tighter analysis of alternating stochastic gradient method for stochastic nested problems. In *Proc. of Advances in Neural Information Processing Systems*, 2021.

Y. Chen, J. Wang, C. Yu, W. Gao, and X. Qin. Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems*, 35(4):83–93, 2020.

T. Degris, M. White, and R.S. Sutton. Off-policy actor-critic. *arXiv preprint:1205.4839*, 2012.

K. Doya. Reinforcement learning in continuous time and space. *Neural Computation*, 12(1):219–245, 2000.

S. Du, J. Chen, L. Li, L. Xiao, and D. Zhou. Stochastic variance reduction methods for policy evaluation. *arXiv preprint:1702.07944*, 2017.

P. et al., Kairouz. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 4:1–210, 2021.

C. Gelada and M. G. Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proc. of AAAI Conference on Artificial Intelligence*, 2019.

A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint:1811.03604*, 2019.

D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69 (2):217–232, 1995.

K. Hsieh, A. Phanishayee, O. Mutlu, and P. B. Gibbons. The non-iid data quagmire of decentralized machine learning. In *Proc. of International Conference on Machine Learning*, 2020.

J. Huang and N. Jiang. On the convergence rate of off-policy policy optimization methods with density-ratio correction. *arXiv preprint:2106.00993*, 2021.

Y. Huang, S. Gupta, Z. Song, K. Li, and S. Arora. Evaluating gradient inversion attacks and defenses in federated learning. In *Proc. of Advances in Neural Information Processing Systems*, 2021.

E. Imani, E. Graves, and M. White. An off-policy policy gradient theorem using emphatic weightings. In *Proc. of Advances in Neural Information Processing Systems*, 2018.

H. Jin, Y. Peng, W. Yang, S. Wang, and Z. Zhang. Federated reinforcement learning with environment heterogeneity. In *Proc. of International Conference on Artificial Intelligence and Statistics*, 2022.

X. Jin, P. Chen, C. Hsu, C. Yu, and T. Chen. Cafe: Catastrophic data leakage in vertical federated learning. In *Proc. of Advances in Neural Information Processing Systems*, 2021.

R. Kidambi, A. Rajeswaran, P. Netrapalli, and T. Joachims. Morel: Model-based offline reinforcement learning. In *Proc. of Advances in Neural Information Processing Systems*, 2020.

V. Konda and V. Borkar. Actor-critic–type learning algorithms for markov decision processes. *SIAM Journal on Control and Optimization*, 38(1):94–123, 1999.

J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint:1610.02527*, 2016.

A. Kumar, J. Fu, G. Tucker, and S. Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Proc. of Advances in Neural Information Processing Systems*, 2019.

A. Kumar, A. Zhou, G. Tucker, and S. Levine. Conservative q-learning for offline reinforcement learning. In *Proc. of Advances in Neural Information Processing Systems*, 2020.

B. Lee, J. Lee, and K. Kim. Representation balancing offline model-based reinforcement learning. In *Proc. of International Conference on Learning Representations*, 2021.

S. Lee and D. Choi. Federated reinforcement learning for energy management of multiple smart homes with distributed energy resources. *IEEE Transactions on Industrial Informatics*, 18(1):488–497, 2020.

Y. Li, X. Li, G. Li, , and Z. Li. Privacy protection in prosumer energy management based on federated learning. *IEEE Access*, 9(1):16707–16715, 2021.

B. Liu, L. Wang, and M. Liu. Lifelong federated reinforcement learning: A learning architecture for navigation in cloud robotic systems. *IEEE Robotics and Automation Letters*, 4(4):4555–4562, 2019a.

L. Liu, J. Zhang, S. H. Song, and K. B. Letaief. Client-edge-cloud hierarchical federated learning. In *Proc. IEEE International Conference on Communications*, 2020a.

Q. Liu, L. Li, Z. Tang, and D. Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *arXiv preprint:1810.12429*, 2018.

Y. Liu, A. Swaminathan, A. Agarwal, and E. Brunskill. Off-policy policy gradient with stationary distribution correction. In *Proc. of Conference on Uncertainty in Artificial Intelligence*, 2019b.

Y. Liu, S. Sun, Z. Ai, S. Zhang, Z. Liu, and H. Yu. Fedcoin: A peer-to-peer payment system for federated learning. *arXiv preprint:2002.11711*, 2020b.

Y. Liu, A. Swaminathan, A. Agarwal, and E. Brunskill. Provably good batch reinforcement learning without great exploration. *arXiv preprint:2007.08202*, 2020c.

T. Matsushima, H. Furuta, Y. Matsuo, O. Nachum, and S. Gu. Deployment-efficient reinforcement learning via model-based offline optimization. *arXiv preprint:2006.03647*, 2020.

H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proc. of International Conference on Artificial Intelligence and Statistics*, 2017.

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

V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proc. of International Conference on Machine Learning*, 2016.

S. A. Murphy, Mark J van der Laan, J. M. Robins, and Conduct Problems Prevention ResearchGroup. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

O. Nachum, Y. Chow, B. Dai, and L. Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Proc. of Advances in Neural Information Processing Systems*, 2019a.

O. Nachum, B. Dai, I. Kostrikov, Y. Chow, L. Li, and D. Schuurmans. Algaedice: Policy gradient from arbitrary experience. *arXiv preprint:1912.02074*, 2019b.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. Estimating divergence functions and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5851, 2010.

N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar. Federated semi-supervised learning with inter-client consistency and disjoint learning. In *Proc. of International Conference on Learning Representations*, 2017.

N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and U. Erlingsson. Scalable private learning with pate. In *Proc. of International Conference on Learning Representations*, 2018.

D. Pollard. *Convergence of stochastic processes*. Springer-Verlag Berlin and Heidelberg GmbH and Co. KG, 1984.

F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Proc. Interspeech*, 2014.

N. Y. Siegel, J. T. Springenberg, F. Berkenkamp, A. Abdolmaleki, M. Neunert, T. Lampe, R. Hafner, N. Heess, and M. Riedmiller. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. In *Proc. of International Conference on Learning Representations*, 2020.

V. Smith, C. Chiang, M. Sanjabi, and A. Talwalkar. Federated multi-task learning. *arXiv preprint:1705.10467*, 2018.

S. U Stich, J. Cordonnier, and M. Jaggi. Sparsified sgd with memory. In *Proc. of Advances in Neural Information Processing Systems*, 2018.

R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with

function approximation. In *Proc. of Advances in Neural Information Processing Systems*, 2000.

T. Tu, G. Thomas, L. Yu, S. Ermon, J. Zou, S. Levine, C. Finn, and T. Ma. Mopo: Model-based offline policy optimization. In *Proc. of Advances in Neural Information Processing Systems*, 2020.

J. Wang and G. Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. *arXiv preprint:1806.00582*, 2018.

Y. Wu, G. Tucker, and O. Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint:1911.11361*, 2019.

M. Yang, O. Nachum, B. Dai, L. Li, and D. Schuurmans. Off-policy evaluation via the regularized lagrangian. In *Proc. of Advances in Neural Information Processing Systems*, 2020.

J. Yoon, W. Jeong, G. Lee, E. Yang, and S. J. Hwang. Federated continual learning with weighted inter-client transfer. In *Proc. of International Conference on Machine Learning*, 2021.

H. Yu, S. Yang, and S. Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proc. of AAAI Conference on Artificial Intelligence*, 2019.

T. Yu, A. Kumar, R. Rafailov, A. Rajeswaran, S. Levine, and C. Finn. Combo: Conservative offline model-based policy optimization. *arXiv preprint:2102.08363*, 2021.

J. Zhang, J. Kim, B. Donoghue, and S. Boyd. Sample efficient reinforcement learning with reinforce. In *Proc. of AAAI Conference on Artificial Intelligence*, 2021a.

K. Zhang, A. Koppel, H. Zhu, and T. Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Applied Mathematics*, 58(6):3586–3612, 2019a.

R. Zhang, B. Dai, L. Li, and D. Schuurmans. Gendice: Generalized offline estimation of stationary values. In *Proc. of International Conference on Learning Representations*, 2020.

S. Zhang, W. Boehmer, and S. Whiteson. Generalized off-policy actor-critic. In *Proc. of Advances in Neural Information Processing Systems*, 2019b.

W. Zhang, Q. Lu, Q. Yu, Z. Li, Y. Liu, S. K. Lo, S. Chen, and L. Xu, X.and Zhu. Blockchain-based federated learning fordevice failure detection in industrial iot. *IEEE Internet of Things Journal*, 8(7):5926 – 5937, 2021b.

Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra. Federated learning with non-iid data. *arXiv preprint:1806.00582*, 2018.

Supplementary Material for "Distributed Offline Policy Optimization Over Logged Data"

A Preliminary

A.1 Proof of Lemma 1

Proof. With Assumption 1, we can use the variational form of $D_{\mathcal{X}^2}\big(d_{\pi_\theta}||\bar{d}_D\big)$ (Nguyen et al., 2010) by introducing a dual variable $x \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and rewrite (5) as

$$\begin{aligned}\max_{\theta\in\mathbb{R}^d}\min_{x\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}} \tilde{F}_\lambda(\theta,x) \coloneqq \mathbb{E}_{s,a\sim d_{\pi_\theta}}[r(s,a)] + \tau R(\theta)\\ + \frac{\lambda}{2}\mathbb{E}_{s,a\sim\bar{d}_D}[x(s,a)^2] - \lambda\mathbb{E}_{s,a\sim d_{\pi_\theta}}[x(s,a)]\end{aligned} \tag{25}$$

where given π_θ, the optimal solution for x is given by $x^*_\theta(s,a) \coloneqq d_{\pi_\theta}(s,a)/\bar{d}_D(s,a)$.

By defining the Bellman operator $\mathcal{B}_{\pi_\theta}v(s,a) \coloneqq r(s,a)+\gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a),a'\sim\pi_\theta(\cdot|s')}[v(s',a')]$, we can use a change of variable $x(s,a) = \frac{1}{\lambda}(\mathcal{B}_{\pi_\theta}v - v)(s,a)$ to rewrite (25) as

$$\begin{aligned}\max_{\theta\in\mathbb{R}^d}\min_{v\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}} F_\lambda(\theta,v) \coloneqq (1-\gamma)\mathbb{E}_{s_0\sim\rho,a_0\sim\pi_\theta}\big[v(s_0,a_0)\big]\\ + \frac{\lambda}{2}\mathbb{E}_{s,a\sim\bar{d}_D}\big[\big((\mathcal{B}_{\pi_\theta}v - v)(s,a)/\lambda\big)^2\big] + \tau R(\theta).\end{aligned} \tag{26}$$

However, $F_\lambda(\theta,v)$ still is not easy to optimize due to the expectation inside the quadratic function, which will cause the so-called *double sampling* problem when trying to obtain the unbiased gradient of $F_\lambda(\theta,v)$.

To tackle this problem, we can use the convex conjugate technique: Any convex function $f(x)$ can be written as $f(x) = \max_\mu \mu x + f^*(\mu)$ where f^* is the convex conjugate of f. With the fact that the conjugate of $f(x) = \frac{1}{2}x^2$ is itself, we can rewrite the quadratic term in (26) and obtain an equivalent objective

$$\begin{aligned}&\max_{\theta\in\mathbb{R}^d}\min_{v\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}}\max_{\mu\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}} F_\lambda(\theta,v,\mu) \coloneqq \frac{1}{N}\sum_{n=1}^N F^n_\lambda(\theta,v,\mu) \\ &\text{with}\quad F^n_\lambda(\theta,v,\mu) \coloneqq (1-\gamma)\mathbb{E}_{\substack{s_0\sim\rho\\ a_0\sim\pi_\theta}}\big[v(s_0,a_0)\big] \\ &+ \mathbb{E}_{s,a\sim d^n_D}\Big[\big(\mathcal{B}_{\pi_\theta}v - v\big)(s,a)\mu(s,a) - \frac{\lambda}{2}\mu(s,a)^2\Big] + \tau R(\theta).\end{aligned} \tag{27}$$

Using the optimality condition of $F_\lambda(\theta,v,\mu)$ with respect to $\mu(s,a)$, we have

$$\mu^*_\theta(s,a) = \underbrace{\frac{1}{\lambda}\big(\mathcal{B}_{\pi_\theta}v^*_\theta - v^*_\theta\big)(s,a)}_{x^*_\theta(s,a)} = \frac{d_{\pi_\theta}(s,a)}{\frac{1}{N}\sum_{n=1}^N d^n_D(s,a)}. \tag{28}$$

Solving (28) with respect to v^*_θ gives

$$v^*_\theta(s,a) = \mathbb{E}_{\pi_\theta}\Big[\sum_{t=0}^\infty \gamma^t\Big(r(s_t,a_t) - \lambda\frac{d_{\pi_\theta}}{\bar{d}_D}(s_t,a_t)\Big)\Big|s_0 = s, a_0 = a\Big]. \tag{29}$$

This completes the proof. □

A.2 Proof of Lemma 2

To prove Lemma 2, we first give a useful lemma on the fixed point equation of the visitation distribution as follows.

Lemma 3. *Given policy π, suppose d_π is its visitation distribution under initial distribution ρ and transition kernel $\mathcal{P}$. For any $Z : \mathcal{S}\times\mathcal{A}\to\mathbb{R}^d$ s.t. $\|Z\| < \infty$, we have*

$$\mathbb{E}_{s,a\sim d_\pi}\Big[Z(s,a)\Big] = (1-\gamma)\mathbb{E}_{s_0\sim\rho,a_0\sim\pi(\cdot|s_0)}\Big[Z(s_0,a_0)\Big] + \gamma\mathbb{E}_{s,a\sim d_\pi,s'\sim\mathcal{P}(\cdot|s,a),a'\sim\pi(\cdot|s')}\Big[Z(s',a')\Big]. \tag{30}$$

Proof. Expanding the expectation in the second term in RHS of (30) gives

$$\begin{aligned}
&\gamma\mathbb{E}_{s,a\sim d_\pi,s'\sim\mathcal{P}(\cdot|s,a),a'\sim\pi(\cdot|s')}\Big[Z(ds',da')\Big]\\
&=(1-\gamma)\int_{s\in\mathcal{S}}\int_{a\in\mathcal{A}}\sum_{t=0}^{\infty}\gamma^{t+1}\mathbb{P}(s_t=ds,a_t=da)\int_{s'\in\mathcal{S}}\int_{a'\in\mathcal{A}}\mathcal{P}(ds'|ds,da)\pi(da'|ds')Z(ds',da')\\
&=(1-\gamma)\int_{s'\in\mathcal{S}}\int_{a'\in\mathcal{A}}\sum_{t=0}^{\infty}\gamma^{t+1}\int_{s\in\mathcal{S}}\int_{a\in\mathcal{A}}\mathbb{P}(s_{t+1}=ds',a_{t+1}=da',s_t=ds,a_t=da)Z(ds',da')\\
&=(1-\gamma)\int_{s'\in\mathcal{S}}\int_{a'\in\mathcal{A}}\sum_{t=1}^{\infty}\gamma^{t}\mathbb{P}(s_t=ds',a_t=da')Z(ds',da'),
\end{aligned}\tag{31}$$

where the first equality is due to the definition of d_π, and the second equality is due to the interchangeability of integral. Expanding the expectation in the first term in RHS of (30) gives

$$(1-\gamma)\mathbb{E}_{s_0\sim\rho,a_0\sim\pi(\cdot|s_0)}\Big[Z(s_0,a_0)\Big]=(1-\gamma)\int_{\mathcal{S}}\int_{\mathcal{A}}\gamma^0\mathbb{P}(s_0=ds,a_0=da)Z(ds,da).\tag{32}$$

Adding (31) and (32) together gives

$$\begin{aligned}
&(1-\gamma)\mathbb{E}_{s_0\sim\rho,a_0\sim\pi(\cdot|s_0)}\Big[Z(s_0,a_0)\Big]+\gamma\mathbb{E}_{s,a\sim d_\pi,s'\sim\mathcal{P}(\cdot|s,a),a'\sim\pi(\cdot|s')}\Big[Z(s',a')\Big]\\
&=\int_{\mathcal{S}}\int_{\mathcal{A}}(1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t=ds,a_t=da)Z(ds,da)=\mathbb{E}_{s,a\sim d_\pi}\big[Z(s,a)\big]
\end{aligned}\tag{33}$$

which completes the proof. □

Now we are ready to give the proof of Lemma 2.

Proof. By reversing the derivation of (27) to (26), we have $F_\lambda(\theta,v_\theta^*,\zeta_\theta^*)=F_\lambda(\theta,v_\theta^*)$. Then we have

$$\begin{aligned}
\nabla_\theta F_\lambda(\theta,v_\theta^*,\zeta_\theta^*)&=\nabla_\theta F_\lambda(\theta,v_\theta^*)\\
&=(1-\gamma)\nabla_\theta\mathbb{E}_{s_0\sim\rho,a_0\sim\pi_\theta}\big[v_\theta^*(s_0,a_0)\big]\\
&\quad+\nabla_\theta\mathbb{E}_{s,a\sim\bar{d}_D}\big[\frac{1}{\lambda}(\mathcal{B}_{\pi_\theta}v_\theta^*-v_\theta^*)(s,a)\nabla_\theta\mathcal{B}_{\pi_\theta}v_\theta^*(s,a)\big]+\tau\nabla_\theta R(\theta)\\
&=(1-\gamma)\nabla_\theta\mathbb{E}_{s_0\sim\rho,a_0\sim\pi_\theta}\big[v_\theta^*(s_0,a_0)\big]+\mathbb{E}_{s,a\sim d_{\pi_\theta}}\big[\nabla_\theta\mathcal{B}_{\pi_\theta}v_\theta^*(s,a)\big]+\tau\nabla_\theta R(\theta)\\
&=(1-\gamma)\mathbb{E}_{s_0\sim\rho,a_0\sim\pi_\theta}\big[v_\theta^*(s_0,a_0)\psi_\theta(s_0,a_0)\big]+\mathbb{E}_{s,a\sim d_{\pi_\theta},s'\sim\mathcal{P},a'\sim\pi_\theta}\big[v_\theta^*(s',a')\psi_\theta(s',a')\big]+\tau\nabla_\theta R(\theta)
\end{aligned}\tag{34}$$

where the third equality follows from (28), and the last equality is obtained by using the so-called log-trick:

$$\nabla_\theta\pi_\theta(a|s)=\pi_\theta(a|s)\psi_\theta(s,a).\tag{35}$$

Applying Lemma 3 to (34) gives

$$\nabla_\theta F_\lambda(\theta,v_\theta^*,\zeta_\theta^*)=\mathbb{E}_{s,a\sim d_{\pi_\theta}}\big[v_\theta^*(s,a)\psi_\theta(s,a)\big]+\tau\nabla R(\theta)\tag{36}$$

which completes the proof. □

A.3 Definitions

With the definition in (17), $\hat{F}_\lambda^n(\theta,v,\mu)$ with linear parametrization of v and μ can be written as

$$\hat{F}_\lambda^n(\theta,\omega_v,\omega_\mu)=b_\theta^n\omega_v+\omega_v^\top A_\theta^n\omega_\mu+\omega_\mu^\top h^n-\frac{\lambda}{2}\omega_\mu^\top C^n\omega_\mu-\tau R(\theta).\tag{37}$$

We then define G_θ^n which will later be used in proof of Theorem 4:

$$G_\theta^n \coloneqq \begin{bmatrix} 0 & \sqrt{\varphi}A_\theta^n \\ -\sqrt{\varphi}(A_\theta^n)^\top & \varphi\lambda C^n \end{bmatrix}. \tag{38}$$

We also define $\hat{d}_D \coloneqq \frac{1}{N}\sum_{n=1}^N \hat{d}_D^n$ and $\hat{\rho} = \frac{1}{N}\sum_{n=1}^N \hat{\rho}^n$. The distribution $\hat{d}_D$ and $\hat{\rho}$ can be respectively viewed as the empirical distribution of samples in $\{\mathcal{D}^n\}_{n=1}^N$ and $\{\mathcal{D}_0^n\}_{n=1}^N$, which aggregate the data from all clients. Then we have the empirical surrogation of the global objective function $F_\lambda(\theta, v, \mu)$ can be written as

$$\hat{F}_\lambda(\theta, \omega_v, \omega_\mu) = b_\theta\omega_v + \omega_v^\top A_\theta\omega_\mu + \omega_\mu^\top h - \frac{\lambda}{2}\omega_\mu^\top C\omega_\mu - \tau R(\theta), \tag{39}$$

where A_θ, b_θ, C and h are respectively the average of A_θ^n, b_θ^n, C^n and h^n for $n \in \{1, 2, \ldots, N\}$:

$$A_\theta \coloneqq \frac{1}{N}\sum_{n=1}^N A_\theta^n = \mathbb{E}_{s,a,s'\sim\hat{d}_D, a'\sim\pi_\theta}\left[(\gamma\phi(s', a') - \phi(s, a))\phi(s, a)^\top\right] \tag{40a}$$

$$b_\theta \coloneqq \frac{1}{N}\sum_{n=1}^N b_\theta^n = (1-\gamma)\mathbb{E}_{s_0\sim\hat{\rho}, a_0\sim\pi_\theta}[\phi(s_0, a_0)] \tag{40b}$$

$$C \coloneqq \frac{1}{N}\sum_{n=1}^N C^n = \mathbb{E}_{s,a\sim\hat{d}_D}[\phi(s, a)\phi(s, a)^\top] \tag{40c}$$

$$h \coloneqq \frac{1}{N}\sum_{n=1}^N h^n = \mathbb{E}_{s,a\sim\hat{d}_D}[r(s, a)\phi(s, a)]. \tag{40d}$$

Then we can also define the averaged matrix G_θ

$$G_\theta \coloneqq \frac{1}{N}\sum_{n=1}^N G_\theta^n = \begin{bmatrix} 0 & \sqrt{\varphi}A_\theta \\ -\sqrt{\varphi}(A_\theta)^\top & \varphi\lambda C \end{bmatrix}. \tag{41}$$

We define the linear function classes $\mathcal{F}_v \coloneqq \{\phi(\cdot)^\top\omega_v | \omega_v \in \mathbb{R}^{d_1}, \|\omega_v\|_2 \le R_v\}$ and $\mathcal{F}_\mu \coloneqq \{\phi(\cdot)^\top\omega_\mu | \omega_\mu \in \mathbb{R}^{d_1}, \|\omega_\mu\|_2 \le R_\mu\}$. Given π_θ, define $\hat{v}_\theta^*, \hat{\mu}_\theta^* \coloneqq \arg\min_{v\in\mathcal{F}_v}\max_{\mu\in\mathcal{F}_\mu}\hat{F}_\lambda(\theta, v, \mu)$. Under linear parametrization, $\hat{v}_\theta^*(s, a) = \phi(s, a)^\top\hat{\omega}_v^*(\theta)$ and $\hat{\mu}_\theta^*(s, a) = \phi(s, a)^\top\hat{\omega}_\mu^*(\theta)$. We also define $\tilde{v}_\theta^*, \tilde{\mu}_\theta^* \in \arg\min_{v\in\mathcal{F}_v}\max_{\mu\in\mathcal{F}_\mu} F_\lambda(\theta, v, \mu)$. Under linear parametrization, $\tilde{v}_\theta^*(s, a) = \phi(s, a)^\top\tilde{\omega}_v^*(\theta)$ and $\tilde{\mu}_\theta^*(s, a) = \phi(s, a)^\top\tilde{\omega}_\mu^*(\theta)$.

Lastly, we define the concatenated critic variable as $\omega \coloneqq [\omega_v, \frac{1}{\sqrt{\varphi}}\omega_\mu]^\top \in \mathbb{R}^{2d_1}$ and $\hat{\omega}^*(\theta) \coloneqq [\omega_v^*(\theta), \frac{1}{\sqrt{\varphi}}\omega_\mu^*(\theta)]^\top$ is the optimal ω of the empirical objective function $\hat{F}_\lambda(\theta, v, \mu)$. Similarly, define $\bar{\omega} \coloneqq [\bar{\omega}_v \ \ \frac{1}{\sqrt{\varphi}}\bar{\omega}_\mu]^\top$ as the concatenated client-averaged critic variable.

B Proof of Theorem 1

B.1 Main proof

We first give a proposition regarding the L_τ-Lipschitz of the policy gradient under Assumption 3, which has been shown by (Agarwal et al., 2020; Zhang et al., 2019a).

Proposition 1. *Suppose Assumption 3 holds. For any $\theta, \theta' \in \mathbb{R}^d$, we have $\|\nabla J_\tau(\theta) - \nabla J_\tau(\theta')\|_2 \le L_\tau\|\theta - \theta'\|_2$, where L_τ is a positive constant.*

Now we begin to consider Algorithm 1. Recall the policy update takes the following form

$$\theta_{k+1}^n = \theta_k^n + \alpha\hat{p}_k^n, \tag{42}$$

Recall the policy gradient estimation $\hat{p}_k^n$ is defined as

$$\hat{p}_k^n \coloneqq \mu_{k+1}^n(\tilde{s}, \tilde{a})v_{k+1}^n(\tilde{s}, \tilde{a})\psi_{\theta_k^n}(\tilde{s}, \tilde{a}) + \tau\psi_{\theta_k^n}(s_p, a_p), \tag{43}$$

where $(\tilde{s},\tilde{a})\sim\hat{d}_D^n$, $s_p\sim\eta_p$ and $a_p\sim\pi_p(\cdot|s_p)$. Note that the random samples in (43) is obtained by client n at iteration k, but we omit the subscriptions k,n on samples for ease of notation.

Define the client-averaged expected gradient as $p_k\coloneqq\mathbb{E}\Big[\frac{1}{N}\sum_{n=1}^N\hat{p}_k^n\big|\mathcal{F}_k\Big]$, where $\mathcal{F}_k\coloneqq\{\mu_{k+1}^n,v_{k+1}^n,\theta_k^n\}_{n=1}^N$. Then it is easy to verify that $\|\hat{p}_k^n\|_2\leq C_p\coloneqq C_\mu C_v C_\psi+\tau C_\psi$ and $\|p_k\|_2\leq C_p$, where C_v,C_μ are respectively the upper bound of $\|\omega_{v,k}^n\|_2$ and $\|\omega_{\mu,k}^n\|_2$ as shown in Lemma 7

Now we are ready to give the convergence proof.

Proof. By Proposition 1, we have

$$J_\tau(\bar{\theta}_{k+1})\geq J_\tau(\bar{\theta}_k)+\big\langle\nabla J_\tau(\bar{\theta}_k),\bar{\theta}_{k+1}-\bar{\theta}_k\big\rangle-\frac{L_\tau}{2}\big\|\bar{\theta}_{k+1}-\bar{\theta}_k\big\|_2^2. \tag{44}$$

The second term can be bounded as

$$\begin{aligned}\mathbb{E}[\big\langle\nabla J_\tau(\bar{\theta}_k),\bar{\theta}_{k+1}-\bar{\theta}_k\big\rangle]&=\alpha\mathbb{E}\Big\langle\nabla J_\tau(\bar{\theta}_k),\frac{1}{N}\sum_{n=1}^N\hat{p}_k^n\Big\rangle=\alpha\mathbb{E}\big\langle\nabla J_\tau(\bar{\theta}_k),p_k\big\rangle\\&=\frac{\alpha}{2}\mathbb{E}\Big[\big\|\nabla J_\tau(\bar{\theta}_k)\big\|_2^2+\big\|p_k\big\|_2^2-\big\|\nabla J_\tau(\bar{\theta}_k)-p_k\big\|_2^2\Big]\end{aligned} \tag{45}$$

where the second equality is due to the towering property of expectation. Taking expectation on both sides of (44) and substituting the above bound into it yield

$$\begin{aligned}\mathbb{E}\big[J_\tau(\bar{\theta}_{k+1})\big]\geq\mathbb{E}[J_\tau(\bar{\theta}_k)]&+\frac{\alpha}{2}\mathbb{E}\big\|\nabla J_\tau(\bar{\theta}_k)\big\|_2^2+\frac{\alpha}{2}\mathbb{E}\big\|p_k\big\|_2^2-\frac{\alpha}{2}\mathbb{E}\big\|\nabla J_\tau(\bar{\theta}_k)-p_k\big\|_2^2\\&-\frac{L_\tau}{2}\mathbb{E}\big\|\bar{\theta}_{k+1}-\bar{\theta}_k\big\|_2^2.\end{aligned} \tag{46}$$

Define a Lyapunov function $\mathcal{L}_k\coloneqq-J_\tau(\bar{\theta}_k)+\big\|\bar{\omega}_k-\hat{\omega}_k^*\big\|_2^2$, where $\hat{\omega}_k^*$ is the shorthand notation for $\hat{\omega}^*(\bar{\theta}_k)$. Then we have

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{k+1}-\mathcal{L}_k]&=-\mathbb{E}\big[J_\tau(\bar{\theta}_{k+1})-J_\tau(\bar{\theta}_k)\big]+\mathbb{E}\big\|\bar{\omega}_{k+1}-\hat{\omega}_{k+1}^*\big\|_2^2-\mathbb{E}\big\|\bar{\omega}_k-\hat{\omega}_k^*\big\|_2^2\\&\leq-\frac{\alpha}{2}\mathbb{E}\big\|\nabla J_\tau(\bar{\theta}_k)\big\|_2^2-\frac{\alpha}{2}\mathbb{E}\big\|p_k\big\|_2^2+\frac{\alpha}{2}\mathbb{E}\big\|\nabla J_\tau(\bar{\theta}_k)-p_k\big\|_2^2\\&\quad+\frac{L_\tau}{2}\mathbb{E}\big\|\bar{\theta}_{k+1}-\bar{\theta}_k\big\|_2^2+\mathbb{E}\big\|\bar{\omega}_{k+1}-\hat{\omega}_{k+1}^*\big\|_2^2-\mathbb{E}\big\|\bar{\omega}_k-\hat{\omega}_k^*\big\|_2^2\end{aligned} \tag{47}$$

where the last inequality follows from (46).

Applying Theorem 4 and Lemma 4 to the last inequality gives

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{k+1}-\mathcal{L}_k]&\leq-\frac{\alpha}{2}\mathbb{E}\big\|\nabla J_\tau(\bar{\theta}_k)\big\|_2^2-\frac{\alpha}{4}\mathbb{E}\big\|p_k\big\|_2^2+\big(1+(4L_\omega^2+C_\psi^2C_{v,\mu}^2+1)\alpha\big)\|\bar{\omega}_{k+1}-\hat{\omega}_k^*\|_2^2-\mathbb{E}\big\|\bar{\omega}_k-\hat{\omega}_k^*\big\|_2^2\\&\quad+\frac{L_\tau+2L_\omega^2}{2}\mathbb{E}\big\|\bar{\theta}_{k+1}-\bar{\theta}_k\big\|_2^2+\Big(\frac{C_1}{2}+\frac{C_p^4L_{\omega,2}^2}{2}\Big)(I-1)^2\alpha^3+\frac{C_2}{2}(I-1)^2\alpha\beta^2+\alpha\epsilon\\&\leq-\frac{\alpha}{2}\mathbb{E}\big\|\nabla J_\tau(\bar{\theta}_k)\big\|_2^2-\Big(\frac{\alpha}{4}-\frac{L_\tau+2L_\omega^2}{2}\alpha^2\Big)\mathbb{E}\big\|p_k\big\|_2^2+\big(1+(4L_\omega^2+C_\psi^2C_{v,\mu}^2+1)\alpha\big)\|\bar{\omega}_{k+1}-\hat{\omega}_k^*\|_2^2\\&\quad-\mathbb{E}\big\|\bar{\omega}_k-\hat{\omega}_k^*\big\|_2^2+\Big(\frac{C_1}{2}+\frac{C_p^4L_{\omega,2}^2}{2}\Big)(I-1)^2\alpha^3+\frac{C_2}{2}(I-1)^2\alpha\beta^2+\frac{(L_\tau+L_\omega^2)C_p^2}{N}\alpha^2+\alpha\epsilon\end{aligned} \tag{48}$$

where $\epsilon=\tilde{\mathcal{O}}\Big(\sqrt{\frac{\log\frac{3}{\delta}}{NM}}\Big)+\mathcal{O}\big(\epsilon_{\text{app}}\big)+\mathcal{O}\big(\epsilon_\lambda\big)$ is an error term. The last inequality follows from

$$\begin{aligned}\mathbb{E}\big\|\bar{\theta}_{k+1}-\bar{\theta}_k\big\|_2^2&=\alpha^2\mathbb{E}\Big\|\frac{1}{N}\sum_{n=1}^N\hat{p}_k^n\Big\|_2^2=\alpha^2\bigg(\mathbb{E}\Big\|\frac{1}{N}\sum_{n=1}^N\big(\hat{p}_k^n-\mathbb{E}\big[\hat{p}_k^n\big|\mathcal{F}_k\big]\big)\Big\|_2^2+\mathbb{E}\big\|p_k\big\|_2^2\bigg)\\&=\alpha^2\bigg(\frac{1}{N}\sum_{n=1}^N\mathbb{E}\big\|\hat{p}_k^n-\mathbb{E}\big[\hat{p}_k^n\big|\mathcal{F}_k\big]\big\|_2^2+\mathbb{E}\big\|p_k\big\|_2^2\bigg)\end{aligned}$$

$$\leq \alpha^2\big(\frac{2C_p^2}{N} + \mathbb{E}\big\|p_k\big\|_2^2\big), \tag{49}$$

where the second equality follows from the basic equation $\mathbb{E}\|Z\|^2 = \mathbb{E}\|Z - \mathbb{E}[Z]\|^2 + \|\mathbb{E}[Z]\|^2$, and the third equality is due to the fact that $\hat{p}_k^n - \mathbb{E}\big[\hat{p}_k^n\big|\mathcal{F}_k\big]$ has zero mean and is conditionally independent across $n \in \{1, 2, \ldots, N\}$ given $\mathcal{F}_k$.

In the following proof, we will hide some unimportant constants with $\mathcal{O}(\cdot)$ for brevity. We assume K is large enough such that $\alpha \leq 1$. Then applying (87b) in Theorem 4 to (48) gives

$$\begin{aligned}
&\mathbb{E}[\mathcal{L}_{k+1} - \mathcal{L}_k] \\
&\leq -\frac{\alpha}{2}\mathbb{E}\big\|\nabla J_\tau(\bar{\theta}_k)\big\|_2^2 - \Big(\frac{\alpha}{4} - \frac{L_\tau + 2L_\omega^2}{2}\alpha^2\Big)\mathbb{E}\big\|p_k\big\|_2^2 \\
&\quad + \Big(\big(1 + (4L_\omega^2 + C_\psi^2 C_{v,\mu}^2 + 1)\alpha\big)\big(1 - C_\lambda\beta + C_6'\beta^2\big) - 1\Big)\mathbb{E}\big\|\bar{\omega}_k - \hat{\omega}_k^*\big\|_2^2 \\
&\quad + \mathcal{O}\Big((I-1)^2\alpha^3\Big) + \mathcal{O}\Big((I-1)^2\alpha\beta^2\Big) + \mathcal{O}\Big((I-1)^2\beta^3\Big) + \mathcal{O}\Big(\frac{\alpha^2 + \beta^2}{N}\Big) + \alpha\epsilon.
\end{aligned} \tag{50}$$

Choose $\alpha = \sqrt{\frac{N}{K}}$, and $\beta = \frac{2(4L_\omega^2 + C_\psi^2 + C_{v,\mu}^2 + 1)}{C_\lambda}\alpha$. Then for $K \geq \max\{4(L_\tau + 2L_\omega^2)^2, 16(4L_\omega^2 + C_\varphi^2 + C_{v,\mu}^2 + 1)^2/C_\lambda^4\}N$, we have that

$$\frac{\alpha}{4} - \frac{L_\tau + 2L_\omega^2}{2}\alpha^2 \geq 0, \tag{51}$$

$$\big(1 + (4L_\omega^2 + C_\psi^2 C_{v,\mu}^2 + 1)\alpha\big)\big(1 - C_\lambda\beta + C_6'\beta^2\big) \leq 1. \tag{52}$$

Then dropping the negative term in RHS of (50) and rearranging give

$$\begin{aligned}
\frac{\alpha}{2}\mathbb{E}\big\|\nabla J_\tau(\bar{\theta}_k)\big\|_2^2 \leq \mathbb{E}[\mathcal{L}_k - \mathcal{L}_{k+1}] + \mathcal{O}\Big((I-1)^2\alpha^3\Big) + \mathcal{O}\Big((I-1)^2\alpha\beta^2\Big) + \mathcal{O}\Big((I-1)^2\beta^3\Big) \\
+ \mathcal{O}\Big(\frac{\alpha^2 + \beta^2}{N}\Big) + \alpha\epsilon.
\end{aligned} \tag{53}$$

Assume $I^4N^3 = \mathcal{O}(K)$, then by the choice of step sizes, the last inequality implies

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\big\|\nabla J_\tau(\bar{\theta}_k)\big\|_2^2 = \mathcal{O}\Big(\frac{1}{\sqrt{NK}}\Big) + \tilde{\mathcal{O}}\Big(\sqrt{\frac{\log\frac{3}{\delta}}{NM}}\Big) + \mathcal{O}\big(\epsilon_{\text{app}}\big) + \mathcal{O}\big(\epsilon_\lambda\big) \tag{54}$$

which completes the proof. □

B.2 Bounding the gradient bias

Let $\bar{v}_k$ and $\bar{\mu}_k$ be parametrized by the averaged parameters $\bar{\omega}_{v,k}$ and $\bar{\omega}_{\mu,k}$ respectively. We first give the theorem regarding the convergence of v and μ, which will help us prove Lemma 4.

Theorem 3. *Consider Algorithm 1. Suppose v^n and μ^n are parametrized linearly, i.e. $v^n(s,a) = \phi(s,a)^\top\omega_v^n$ and $\mu^n(s,a) = \phi(s,a)^\top\omega_\mu^n$. Suppose Assumption 1-3 hold, then with probability greater than $1-\delta$ we have*

$$\begin{aligned}
&\Big\|\hat{\mathbb{E}}\big[\bar{\mu}_{k+1}(s,a)\bar{v}_{k+1}(s,a)\psi_{\bar{\theta}_k}(s,a)\big] - \bar{\mathbb{E}}\big[\mu_{\bar{\theta}_k}^*(s,a)\hat{v}_{\bar{\theta}_k}^*(s,a)\psi_{\bar{\theta}_k}(s,a)\big]\Big\|_2^2 \\
&\leq 2C_\psi^2 C_{v,\mu}^2\|\bar{\omega}_{k+1} - \hat{\omega}^*(\bar{\theta}_k)\|_2^2 + \tilde{\mathcal{O}}\Big(\sqrt{\frac{\log\frac{3}{\delta}}{NM}}\Big) + \mathcal{O}\big(\epsilon_{\text{app}}(\mathcal{F}_v, \mathcal{F}_\mu)\big),
\end{aligned} \tag{55a}$$

and

$$\mathbb{E}\big[\big(\hat{v}_{\bar{\theta}_k}^*(s,a) - v_{\bar{\theta}_k}^*(s,a)\big)^2\big] = \tilde{\mathcal{O}}\Big(\sqrt{\frac{\log\frac{3}{\delta}}{NM}}\Big) + \mathcal{O}\big(\epsilon_{\text{app}}(\mathcal{F}_v)\big) + \mathcal{O}\big(\epsilon_{\text{app}}(\mathcal{F}_\mu)\big), \tag{55b}$$

where $\epsilon_{\text{app}}(\mathcal{F}_v)$, $\epsilon_{\text{app}}(\mathcal{F}_\mu)$ and $\epsilon_{\text{app}}(\mathcal{F}_v, \mathcal{F}_\mu)$ are function approximation errors.

Given Theorem 3, we are ready to bound the gradient bias.

Lemma 4 (Estimation error of policy gradient)**.** *Consider Algorithm 1. Under the same conditions of Theorem 1, it holds with probability at least* $1-\delta$ *that*

$$\mathbb{E}\big\|\nabla J_\tau(\bar\theta_k)-p_k\big\|_2^2 \le 2C_\psi^2C_{v,\mu}^2\|\bar\omega_{k+1}-\hat\omega^*(\bar\theta_k)\|_2^2 + C_1(I-1)^2\alpha^2 + C_2(I-1)^2\beta^2 + \tilde{\mathcal{O}}\bigg(\sqrt{\frac{\log\frac{3}{\delta}}{NM}}\bigg) + \mathcal{O}\big(\epsilon_{\rm app}\big) + \mathcal{O}\big(\epsilon_\lambda\big), \tag{56}$$

where $C_{v,\mu}$*,* C_1*,* C_2 *are some positive constants.*

Proof. We first define the virtual stochastic gradient with averaged parameters as

$$\bar p_k^n \coloneqq \bar\mu_{k+1}(\tilde s,\tilde a)\bar v_{k+1}(\tilde s,\tilde a)\psi_k(\tilde s,\tilde a)+\tau\psi_k(s_p,a_p), \tag{57}$$

where and ψ_k is the shorthand notations for $\psi_{\bar\theta_k}$. Note that random samples in (57) is the same as that in $\hat p_k^n$. We slightly abuse the notation and omit the subscriptions k,n on samples.

We then define the client averaged expectation of $\bar p_k^n$

$$\begin{aligned}\bar p_k &\coloneqq \mathbb{E}\Big[\frac{1}{N}\sum_{n=1}^N \bar p_k^n\Big|\mathcal{F}_k\Big]\\ &= \mathbb{E}_{s,a\sim\hat d_D}\big[\bar\mu_{k+1}(s,a)\bar v_{k+1}(s,a)\psi_k(s,a)\big]+\tau\nabla R(\bar\theta_k).\end{aligned} \tag{58}$$

We also define the optimal $\bar p_k$ as

$$p_k^* \coloneqq \mathbb{E}_{s,a\sim\bar d_D}\big[\mu_k^*(s,a)v_k^*(s,a)\psi_k(s,a)\big]+\tau\nabla R(\bar\theta_k), \tag{59}$$

where v_k^* and μ_k^* are the shorthand notations for $v_{\bar\theta_k}^*$ and $\mu_{\bar\theta_k}^*$ respectively. Recall v_θ^* is defined in (8) and $\mu_\theta^*=d_{\pi_\theta}/\bar d_D$.

We can decompose the policy gradient error as

$$\big\|\nabla J_\tau(\bar\theta_k)-p_k\big\|_2^2 \le 3\underbrace{\Big\|p_k-\bar p_k\Big\|_2^2}_{I_1}+3\underbrace{\big\|\bar p_k-p_k^*\big\|_2^2}_{I_2}+3\underbrace{\big\|\nabla J_\tau(\bar\theta_k)-p_k^*\big\|_2^2}_{I_3}. \tag{60}$$

Consider I_1 first. By definition of p_k,

$$I_1=\Big\|\mathbb{E}\Big[\frac{1}{N}\sum_{n=1}^N\hat p_k^n-\bar p_k^n\Big|\mathcal{F}_k\Big]\Big\|_2^2 \le \frac{1}{N}\sum_{n=1}^N\mathbb{E}\Big[\big\|\hat p_k^n-\bar p_k^n\big\|_2^2\Big|\mathcal{F}_k\Big]. \tag{61}$$

Consider the term $\big\|\hat p_k^n-\bar p_k^n\big\|_2^2$ in the last inequality. By definition we have

$$\begin{aligned}&\big\|\hat p_k^n-\bar p_k^n\big\|_2^2\\ &\le 4\Big\|\mu_{k+1}^n(s,a)\big(v_{k+1}^n(s,a)-\bar v_{k+1}(s,a)\big)\psi_k(s,a)\Big\|_2^2+4\Big\|\mu_{k+1}^n(s,a)\bar v_{k+1}(s,a)\big(\psi_{\theta_k^n}(s,a)-\psi_k(s,a)\big)\Big\|_2^2\\ &\quad+4\Big\|\big(\mu_{k+1}^n(s,a)-\bar\mu_{k+1}(s,a)\big)\bar v_{k+1}(s,a)\psi_k(s,a)\Big\|_2^2+4\Big\|\tau\psi_{\bar\theta_k}(s_p,a_p)-\tau\psi_{\theta_k^n}(s_p,a_p)\Big\|_2^2\\ &\le 4C_\mu^2C_\psi^2\big\|\omega_{v,k+1}^n-\bar\omega_{v,k+1}^n\big\|_2^2+4(C_\mu^2C_v^2+\tau^2)\big\|\theta_k^n-\bar\theta_k\big\|_2^2+4C_v^2C_\psi^2\big\|\omega_{\mu,k+1}^n-\bar\omega_{\mu,k+1}\big\|_2^2,\end{aligned} \tag{62}$$

Substituting (62) into (61) gives

$$\begin{aligned}I_1 &\le 4C_\mu^2C_\psi^2\big\|\omega_{v,k+1}^n-\bar\omega_{v,k+1}^n\big\|_2^2+4(C_\mu^2C_v^2+\tau^2)\big\|\theta_k^n-\bar\theta_k\big\|_2^2+4C_v^2C_\psi^2\big\|\omega_{\mu,k+1}^n-\bar\omega_{\mu,k+1}\big\|_2^2\\ &\le C_1(I-1)^2\alpha^2+C_2(I-1)^2\beta^2,\end{aligned} \tag{63}$$

where the last inequality follows from Lemma 5, $C_1=8(C_\mu^2C_v^2+\tau^2)C_p^2$, and $C_2=8(C_v^2+C_\mu^2)C_\psi^2C_q^2$.

Term I_2 can be further decomposed as

$$\begin{aligned}\left\|\bar{p}_k - p_k^*\right\|_2^2 &\leq 2\left\|\hat{\mathbb{E}}\Big[\bar{\mu}_{k+1}(s,a)\bar{v}_{k+1}(s,a)\psi_k(s,a)\Big] - \bar{\mathbb{E}}\Big[\mu_k^*(s,a)\hat{v}_k^*(s,a)\psi_k(s,a)\Big]\right\|_2^2 \\ &\quad + 2\left\|\bar{\mathbb{E}}\Big[\mu_k^*(s,a)\big(\hat{v}_k^* - v_k^*\big)(s,a)\psi_k(s,a)\Big]\right\|_2^2 \\ &\leq 2\left\|\hat{\mathbb{E}}\Big[\bar{\mu}_{k+1}(s,a)\bar{v}_{k+1}(s,a)\psi_k(s,a)\Big] - \bar{\mathbb{E}}\Big[\mu_k^*(s,a)\hat{v}_k^*(s,a)\psi_k(s,a)\Big]\right\|_2^2 \\ &\quad + 2C_d^2C_\psi^2\bar{\mathbb{E}}\Big[\big(\hat{v}_k^* - v_k^*\big)^2(s,a)\Big] \end{aligned} \tag{64}$$

where we use $\hat{\mathbb{E}}$ and $\bar{\mathbb{E}}$ as shorthand notations for $\mathbb{E}_{s,a\sim\hat{d}_D}$ and $\mathbb{E}_{s,a\sim\bar{d}_D}$ respectively.

Applying Theorem 3 yield

$$\mathbb{E}[I_3] = 2C_\psi^2C_{v,\mu}^2\|\bar{\omega}_{k+1} - \hat{\omega}^*(\bar{\theta}_k)\|_2^2 + \tilde{\mathcal{O}}\left(\sqrt{\frac{\log\frac{3}{\delta}}{NM}}\right) + \mathcal{O}\big(\epsilon_{\text{app}}\big) \tag{65}$$

which holds with probability greater than $1-\delta$, and function approximation error $\epsilon_{\text{app}} \coloneqq \epsilon_{\text{app}}(\mathcal{F}_v) + \epsilon_{\text{app}}(\mathcal{F}_\mu) + \epsilon_{\text{app}}(\mathcal{F}_v, \mathcal{F}_\mu)$.

Now we consider I_4. By the definition of μ_θ^*, we have

$$p_k^* = \mathbb{E}_{s,a\sim d_{\pi_k}}\big[v_k^*(s,a)\psi_k(s,a)\big] + \tau\nabla R(\bar{\theta}_k), \tag{66}$$

then we have

$$\begin{aligned}\left\|\nabla J_\tau(\bar{\theta}_k) - p_k^*\right\|_2^2 &= \left\|\mathbb{E}_{s,a\sim d_{\pi_k}}\big[\big(v_k^* - Q_{\pi_k}\big)(s,a)\psi_k(s,a)\big]\right\|_2^2 \\ &= \left\|\mathbb{E}_{s,a\sim d_{\pi_k}}\left[\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\lambda\frac{d_{\pi_k}}{\bar{d}_D}(s_t,a_t)\Big|s_0=s,a_0=a\right]\psi_k(s,a)\right]\right\|_2^2 \\ &\leq \epsilon_\lambda = \left(\lambda\frac{C_\psi C_d}{1-\gamma}\right)^2. \end{aligned} \tag{67}$$

Taking expected running average of both sides of (60), then substituting the upper bounds in (63), (65) and (67) into (60) gives

$$\begin{aligned}\mathbb{E}\left\|\nabla J_\tau(\bar{\theta}_k) - p_k\right\|_2^2 &= 2C_\psi^2C_{v,\mu}^2\|\bar{\omega}_{k+1} - \hat{\omega}^*(\bar{\theta}_k)\|_2^2 + C_1(I-1)^2\alpha^2 + C_2(I-1)^2\beta^2 \\ &\quad + \tilde{\mathcal{O}}\left(\sqrt{\frac{\log\frac{3}{\delta}}{NM}}\right) + \mathcal{O}\big(\epsilon_{\text{app}}\big) + \mathcal{O}\big(\epsilon_\lambda\big) \end{aligned} \tag{68}$$

which holds with probability greater than $1-\delta$. This completes the proof. □

B.3 Bounding the consensus error

The following lemma bounds the difference between local sequences and their average.

Lemma 5. *Under linear parametrization of v^n and μ^n, consider Algorithm 1 under Assumption 2. For any iteration k, it holds that*

$$\|\theta_k^n - \bar{\theta}_k\|_2^2 \leq 2(I-1)^2C_p^2\alpha^2 \tag{69a}$$

$$\|\omega_{v,k}^n - \bar{\omega}_{v,k}\|_2^2 \leq 2(I-1)^2C_q^2\beta^2 \tag{69b}$$

$$\|\omega_{\mu,k}^n - \bar{\omega}_{\mu,k}\|_2^2 \leq 2(I-1)^2C_q^2\beta^2. \tag{69c}$$

Proof. At any iteration k, the server update in Algorithm 1 guarantees that there exists a iteration number $k_0 \in [k-I+1, k]$ such that $\theta_{k_0}^n = \bar{\theta}_{k_0}$. Then we have

$$\begin{aligned}\|\theta_k^n - \bar{\theta}_k\|_2^2 &= \left\|\theta_{k_0}^n + \sum_{i=k_0}^{k-1} \alpha \hat{p}_i^n - \bar{\theta}_{k_0} - \frac{1}{N}\sum_{n=1}^N \sum_{i=k_0}^{k-1} \alpha \hat{p}_i^n\right\|_2^2 \\ &= \left\| \sum_{i=k_0}^{k-1} \alpha \left(\hat{p}_i^n - \frac{1}{N}\sum_{n=1}^N \hat{p}_i^n\right)\right\|_2^2 \\ &\leq (k-k_0)\sum_{i=k_0}^{k-1} \left\|\alpha\left(\hat{p}_i^n - \frac{1}{N}\sum_{n=1}^N \hat{p}_i^n\right)\right\|_2^2 \\ &\leq (k-k_0)\sum_{i=k_0}^{k-1} 2\alpha^2 C_p^2 \leq 2(I-1)^2 C_p^2 \alpha^2 \end{aligned} \tag{70}$$

where the last inequality follows from $k_0 \in [k-I+1, t]$. Similarly, it can be proven that

$$\begin{aligned}\|\omega_{v,k}^n - \bar{\omega}_{v,k}\|_2^2 &\leq 2(I-1)^2 C_q^2 \beta^2, \\ \|\omega_{\mu,k}^n - \bar{\omega}_{\mu,k}\|_2^2 &\leq 2(I-1)^2 C_q^2 \beta^2, \end{aligned} \tag{71}$$

where $C_q \coloneqq \max\{2C_\mu + 1, \lambda C_\mu + 2C_v + r_{\max}\}$. □

C Analysis of critic

C.1 Proof of Theorem 3

In the proof, we write $\hat{v}_{\bar{\theta}_k}^*$, $\bar{v}_{\bar{\theta}_k}^*$ and $v_{\bar{\theta}_k}^*$ in short as $\hat{v}_k^*$, $\bar{v}_k^*$ and v_k^* respectively (and likewise for μ).

Proof. First we begin to prove (55a). We decompose the error as

$$\begin{aligned}&\left\|\hat{\mathbb{E}}\big[\bar{\mu}_{k+1}(s,a)\bar{v}_{k+1}(s,a)\psi_k(s,a)\big] - \bar{\mathbb{E}}\big[\mu_k^*(s,a)\hat{v}_k^*(s,a)\psi_k(s,a)\big]\right\|_2^2 \\ &\leq 4\underbrace{\left\|\hat{\mathbb{E}}\big[\bar{\mu}_{k+1}(s,a)\bar{v}_{k+1}(s,a)\psi_k(s,a)\big] - \hat{\mathbb{E}}\big[\hat{\mu}_k^*(s,a)\hat{v}_k^*(s,a)\psi_k(s,a)\big]\right\|_2^2}_{I_1} \\ &\quad + 4\underbrace{\left\|\hat{\mathbb{E}}\big[\hat{\mu}_k^*(s,a)\hat{v}_k^*(s,a)\psi_k(s,a)\big] - \bar{\mathbb{E}}\big[\hat{\mu}_k^*(s,a)\hat{v}_k^*(s,a)\psi_k(s,a)\big]\right\|_2^2}_{I_2} \\ &\quad + 4\underbrace{\left\|\bar{\mathbb{E}}\big[\hat{\mu}_k^*(s,a)\hat{v}_k^*(s,a)\psi_k(s,a)\big] - \bar{\mathbb{E}}\big[\tilde{\mu}_k^*(s,a)\hat{v}_k^*(s,a)\psi_k(s,a)\big]\right\|_2^2}_{I_3} \\ &\quad + 4\underbrace{\left\|\bar{\mathbb{E}}\big[\tilde{\mu}_k^*(s,a)\hat{v}_k^*(s,a)\psi_k(s,a)\big] - \bar{\mathbb{E}}\big[\mu_k^*(s,a)\hat{v}_k^*(s,a)\psi_k(s,a)\big]\right\|_2^2}_{I_4}. \end{aligned} \tag{72}$$

We will bound the terms one by one. The first term I_1 can be bounded as

$$\begin{aligned} I_1 &\leq \hat{\mathbb{E}}\left\|\bar{\mu}_{k+1}(s,a)\bar{v}_{k+1}(s,a)\psi_k(s,a) - \hat{\mu}_k^*(s,a)\hat{v}_k^*(s,a)\psi_k(s,a)\right\|_2^2 \\ &\leq 2\hat{\mathbb{E}}\left\|\big(\bar{\mu}_{k+1} - \hat{\mu}_k^*\big)(s,a)\bar{v}_{k+1}(s,a)\psi_k(s,a)\right\|_2^2 + 2\hat{\mathbb{E}}\left\|\hat{\mu}_k^*(s,a)\big(\bar{v}_{k+1} - \hat{v}_k^*\big)(s,a)\psi_k(s,a)\right\|_2^2 \\ &\leq 2C_v^2 C_\psi^2 \hat{\mathbb{E}}\Big[\big(\bar{\mu}_{k+1}(s,a) - \hat{\mu}_k^*(s,a)\big)^2\Big] + 2R_\mu^2 C_\psi^2 \hat{\mathbb{E}}\Big[\big(\bar{v}_{k+1}(s',a') - \hat{v}_k^*(s',a')\big)^2\Big] \end{aligned}$$

$$\leq 2C_v^2C_\psi^2\varphi\frac{1}{\varphi}\big\|\bar{\omega}_{\mu,k+1}-\hat{\omega}_\mu^*(\bar{\theta}_k)\big\|_2^2+2R_\mu^2C_\psi^2\big\|\bar{\omega}_{v,k}-\hat{\omega}_v^*(\bar{\theta}_k)\big\|_2^2$$
$$\leq 2C_\psi^2C_{v,\mu}^2\|\bar{\omega}_{k+1}-\hat{\omega}^*(\bar{\theta}_k)\|_2^2, \tag{73}$$

where $C_{v,\mu}\coloneqq\max\{C_v\sqrt{\varphi},R_\mu\}$, and the last inequality uses the fact that

$$\|\bar{\omega}-\hat{\omega}^*(\theta)\|_2^2=\frac{1}{\varphi}\|\bar{\omega}_v-\hat{\omega}_v^*(\theta)\|_2^2+\|\bar{\omega}_\mu-\hat{\omega}_\mu^*(\theta)\|_2^2. \tag{74}$$

The term I_2 is introduced by the difference between $\hat{d}_D$ and $\bar{d}_D$, which is essentially caused by finite data set. By Lemma 11, with probability greater than $1-\delta/3$ we have

$$I_2=\tilde{\mathcal{O}}\Big(\frac{\log\frac{3}{\delta}}{NM}\Big). \tag{75}$$

Now we consider I_3. First we have

$$I_3=\Big\|\bar{\mathbb{E}}\big[\hat{\mu}_k^*(s,a)\hat{v}_k^*(s,a)\psi_k(s,a)\big]-\bar{\mathbb{E}}\big[\tilde{\mu}_k^*(s,a)\hat{v}_k^*(s,a)\psi_k(s,a)\big]\Big\|_2^2$$
$$\leq R_v^2C_\psi^2\bar{\mathbb{E}}\Big[\big(\hat{\mu}_k^*(s,a)-\tilde{\mu}_k^*(s,a)\big)^2\Big]\leq R_v^2C_\psi^2\big\|\hat{\omega}_\mu^*(\bar{\theta}_k)-\tilde{\omega}_\mu^*(\bar{\theta}_k)\big\|_2^2. \tag{76}$$

We see that I_3 is introduced by the difference between $\bar{\omega}_\mu^*$ and $\hat{\omega}_\mu^*$, which is again due to finite data set. By the strong-concavity of $\hat{F}_\lambda(\theta,v,\mu)$ with respect to ω_μ and the optimality condition of $\hat{\omega}_\mu^*(\theta)$, we have for any $\theta\in\mathbb{R}^d$

$$\big\|\hat{\omega}_\mu^*(\theta)-\tilde{\omega}_\mu^*(\theta)\big\|_2^2\leq\frac{2}{\eta}\Big(\hat{F}_\lambda(\theta,\hat{v}_\theta^*,\hat{\mu}_\theta^*)-\hat{F}_\lambda(\theta,\hat{v}_\theta^*,\tilde{\mu}_\theta^*)\Big)$$
$$\leq\frac{2}{\eta}\Big(\hat{F}_\lambda(\theta,\hat{v}_\theta^*,\hat{\mu}_\theta^*)-\min_{v\in\mathcal{F}_v}\hat{F}_\lambda(\theta,v,\tilde{\mu}_\theta^*)\Big)$$
$$\leq\frac{2}{\eta}\Big(\hat{F}_\lambda(\theta,\hat{v}_\theta^*,\hat{\mu}_\theta^*)-\min_{v\in\mathcal{F}_v}\hat{F}_\lambda(\theta,v,\tilde{\mu}_\theta^*)+F_\lambda(\theta,\tilde{v}_\theta^*,\tilde{\mu}_\theta^*)-\min_{v\in\mathcal{F}_v}F_\lambda(\theta,v,\hat{\mu}_\theta^*)\Big), \tag{77}$$

where η is the smallest eigenvalue of matrix C.

With $v_{(1)}\in\arg\min_{v\in\mathcal{F}_v}F_\lambda(\theta,v,\hat{\mu}_\theta^*)$ and $v_{(2)}\in\arg\min_{v\in\mathcal{F}_v}\hat{F}_\lambda(\theta,v,\tilde{\mu}_\theta^*)$, from (77), we have

$$\big\|\hat{\omega}_\mu^*(\theta)-\tilde{\omega}_\mu^*(\theta)\big\|_2^2\leq\frac{2}{\eta}\Big(\hat{F}_\lambda(\theta,\hat{v}_\theta^*,\hat{\mu}_\theta^*)-F_\lambda(\theta,v_{(1)},\hat{\mu}_\theta^*)+F_\lambda(\theta,\tilde{v}_\theta^*,\tilde{\mu}_\theta^*)-\hat{F}_\lambda(\theta,v_{(2)},\tilde{\mu}_\theta^*)\Big)$$
$$\leq\frac{2}{\eta}\Big(\hat{F}_\lambda(\theta,v_{(1)},\hat{\mu}_\theta^*)-F_\lambda(\theta,v_{(1)},\hat{\mu}_\theta^*)+F_\lambda(\theta,v_{(2)},\tilde{\mu}_\theta^*)-\hat{F}_\lambda(\theta,v_{(2)},\tilde{\mu}_\theta^*)\Big)$$
$$\leq\frac{4}{\eta}\sup_{\theta\in\mathbb{R}^d,v\in\mathcal{F}_v,\mu\in\mathcal{F}_\mu}\Big|F_\lambda(\theta,v,\mu)-\hat{F}_\lambda(\theta,v,\mu)\Big|=\tilde{\mathcal{O}}\Big(\sqrt{\frac{\log\frac{1}{\delta}}{NM}}\Big), \tag{78}$$

where the last inequality follows from Lemma 12. Substituting the last inequality into (76) gives

$$I_3=\tilde{\mathcal{O}}\Big(\sqrt{\frac{\log\frac{3}{\delta}}{NM}}\Big) \tag{79}$$

which holds with probability greater than $1-\delta/3$.

Next we bound I_4 as

$$I_4=\Big\|\bar{\mathbb{E}}\big[\tilde{\mu}_k^*(s,a)\hat{v}_k^*(s,a)\psi_k(s,a)\big]-\bar{\mathbb{E}}\big[\mu_k^*(s,a)\hat{v}_k^*(s,a)\psi_k(s,a)\big]\Big\|_2^2$$
$$\leq\bar{\mathbb{E}}\Big\|\tilde{\mu}_k^*(s,a)\hat{v}_k^*(s,a)\psi_k(s,a)-\mu_k^*(s,a)\hat{v}_k^*(s,a)\psi_k(s,a)\Big\|_2^2$$
$$\leq R_v^2C_\psi^2\bar{\mathbb{E}}\Big[\big(\tilde{\mu}_k^*(s,a)-\mu_k^*(s,a)\big)^2\Big]\leq R_v^2C_\psi^2\epsilon_{\text{app}}(\mathcal{F}_v,\mathcal{F}_\mu), \tag{80}$$

where $\epsilon_{\text{app}}(\mathcal{F}_v, \mathcal{F}_\mu) \coloneqq \sup_{\theta\in\mathbb{R}^d} \bar{\mathbb{E}}_{\pi_\theta}\Big[\big(\tilde{\mu}^*_\theta(s,a) - \mu^*_\theta(s,a)\big)^2\Big]$ is the function approximation error. Substituting the upper bounds in (73), (75), (79) and (80) into (72) gives the convergence result of μ

$$\begin{aligned}&\Big\|\hat{\mathbb{E}}\big[\bar{\mu}_{k+1}(s,a)\bar{v}_{k+1}(s,a)\psi_k(s,a)\big] - \bar{\mathbb{E}}\big[\mu^*_k(s,a)\hat{v}^*_k(s,a)\psi_k(s,a)\big]\Big\|_2^2\\ &= 2C_\psi^2 C_{v,\mu}^2\|\bar{\omega}_{k+1} - \hat{\omega}^*(\bar{\theta}_k)\|_2^2 + \tilde{\mathcal{O}}\bigg(\sqrt{\frac{\log\frac{3}{\delta}}{NM}} + \epsilon_{\text{app}}(\mathcal{F}_v,\mathcal{F}_\mu)\bigg),\end{aligned} \tag{81}$$

This completes the proof of (55a).

Finally, we begin to prove (55b). By the towering property of the expectation, we have

$$\begin{aligned}\mathbb{E}\big[\big(\hat{v}^*_k(s,a) - v^*_k(s,a)\big)^2\big] &= \mathbb{E}\Big[\mathbb{E}_{s,a\sim\bar{d}_D}\big[\big(\hat{v}^*_k(s,a) - v^*_k(s,a)\big)^2\big]\Big]\\ &\le \frac{2}{\lambda}\mathbb{E}\big[F_\lambda(\bar{\theta}_k,\hat{v}^*_k) - F_\lambda(\bar{\theta}_k, v^*_k)\big]\end{aligned} \tag{82}$$

where the last inequality follows from the strong-convexity of $F_\lambda(\theta, v)$ with respect to v and the optimality condition of v^*_k. Term $F_\lambda(\bar{\theta}_k,\hat{v}^*_k) - F_\lambda(\bar{\theta}_k, v^*_k)$ is introduced by the difference between $\hat{v}^*_k$ and v^*_k, which is further introduced by statistical error and function approximation error. In fact, $F_\lambda(\bar{\theta}_k,\hat{v}^*_k) - F_\lambda(\bar{\theta}_k, v^*_k)$ can be bounded by following the bounding of $J(\hat{v}^*) - J(v^*)$ in section D.1 of (Nachum et al., 2019a). Without repeating the proof, the following inequality holds with probability greater than $1-\delta/3$

$$\mathbb{E}\big[\big(\hat{v}^*_k(s,a) - v^*_k(s,a)\big)^2\big] = \tilde{\mathcal{O}}\bigg(\sqrt{\frac{\log\frac{3}{\delta}}{NM}} + \epsilon_{\text{app}}(\mathcal{F}_v) + \epsilon_{\text{app}}(\mathcal{F}_\mu)\bigg). \tag{83}$$

The $\epsilon_{\text{app}}(\mathcal{F}_v)$ and $\epsilon_{\text{app}}(\mathcal{F}_\mu)$ are respectively the function approximation error of v^n and μ^n:

$$\begin{aligned}\epsilon_{\text{app}}(\mathcal{F}_v) \coloneqq \sup_{\theta\in\mathbb{R}^d}\inf_{v^*_{\mathcal{F},\theta}}\Big[&\mathbb{E}_{\bar{d}_D}\Big|\big(\mathcal{B}_{\pi_\theta}v^*_{\mathcal{F},\theta} - v^*_{\mathcal{F},\theta}\big)(s,a) - \big(\mathcal{B}_{\pi_\theta}v^*_\theta - v^*_\theta\big)(s,a)\Big|\\ &+ \mathbb{E}_{s_0\sim\rho,a_0\sim\pi_\theta}\big|v^*_{\mathcal{F},\theta}(s_0,a_0) - v^*_\theta(s_0,a_0)\big|\Big]\end{aligned} \tag{84a}$$

$$\epsilon_{\text{app}}(\mathcal{F}_\mu) \coloneqq \sup_{\theta\in\mathbb{R}^d}\mathbb{E}_{\bar{d}_D}\big|\mu^*_{\mathcal{F},\theta}(s,a) - \mu^*_{\hat{v}_\theta}(s,a)\big| \tag{84b}$$

where $v^*_{\mathcal{F},\theta}\in\arg\min_{v\in\mathcal{F}_v}F_\lambda(\theta,v)$, $\mu^*_{\mathcal{F},\theta}\in\arg\max_{\mu\in\mathcal{F}_\mu}F_\lambda(\theta,\hat{v}^*_\theta,\mu)$ and $\mu^*_{\hat{v}_\theta}\coloneqq\arg\max_{\mu\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}}F_\lambda(\theta,\hat{v}^*_\theta,\mu)$. This completes the proof. □

C.2 Convergence of critic

Given constant φ, with the concatenated variable $\omega = [\omega_v \;\; \frac{1}{\sqrt{\varphi}}\omega_\mu]^\top$, we can define the concatenated local gradient as

$$g^n_k(\omega) \coloneqq \begin{bmatrix}1 & 0\\ 0 & \frac{1}{\sqrt{\varphi}}\end{bmatrix}\begin{bmatrix}0 & \sqrt{\varphi}\hat{A}^n_k\\ -\sqrt{\varphi}(\hat{A}^n_k)^\top & \varphi\lambda\hat{C}^n_k\end{bmatrix}\omega + \begin{bmatrix}\hat{b}^n_k\\ -\hat{h}^n_k\end{bmatrix}, \tag{85}$$

in which $\hat{A}^n_{\theta,k}$, $\hat{b}^n_{\theta,k}$, $\hat{C}^n_k$ and $\hat{h}^n_k$ are respectively the unbiased stochastic estimation of $A^n_{\theta^n_k}$, $b^n_{\theta^n_k}$, C^n and h^n with the samples at iteration k.

Then the local update in Algorithm 1 can be written as:

$$\omega^n_{k+1} = \omega^n_k - Bg^n_k(\omega^n_k), \quad \text{with} \quad B\coloneqq\begin{bmatrix}1 & 0\\ 0 & \frac{1}{\sqrt{\varphi}}\end{bmatrix}\beta. \tag{86}$$

Recall $\hat{\omega}^*(\theta) = [\hat{\omega}^*_v(\theta) \;\; \frac{1}{\sqrt{\varphi}}\hat{\omega}^*_\mu(\theta)]^\top$ is the optimal ω, and $\bar{\omega}_k = [\bar{\omega}_{v,k} \;\; \frac{1}{\sqrt{\varphi}}\bar{\omega}_{\mu,k}]^\top$ is the concatenated client-averaged variable.

In the following proof, We write $\hat{\omega}^*(\bar{\theta}_k)$ as $\hat{\omega}^*_k$ for notation simplicity. Now we are ready to give the convergence proof.

Theorem 4. *Consider Algorithm 1 with linear function parametrization of v^n and μ^n. Under Assumption 2-3, it holds that*

$$\mathbb{E}\big\|\bar{\omega}_{k+1}-\hat{\omega}^*(\bar{\theta}_{k+1})\big\|_2^2 \le \big(1+(4L_\omega^2+1)\alpha\big)\mathbb{E}\big\|\bar{\omega}_{k+1}-\hat{\omega}^*(\bar{\theta}_k)\big\|_2^2+\frac{\alpha}{4}\mathbb{E}\|p_k\|_2^2+\frac{C_p^4L_{\omega,2}^2}{2}\alpha^3 + L_\omega^2\|\bar{\theta}_{k+1}-\bar{\theta}_k\|_2^2 \tag{87a}$$

and

$$\mathbb{E}\big\|\bar{\omega}_{k+1}-\hat{\omega}^*(\bar{\theta}_k)\big\|_2^2 \le (1-C_\lambda\beta+C_6'\beta^2)\mathbb{E}\big\|\bar{\omega}_k-\hat{\omega}^*(\bar{\theta}_k)\big\|_2^2+(I-1)^2C_4'\beta^3 + (I-1)^2C_5'\alpha^2\beta+\frac{4C_g^2C_\varphi}{N}\beta^2 \tag{87b}$$

where C_4', C_5', C_6', C_φ and C_λ are some positive constants.

Proof. First we have

$$\begin{aligned}\big\|\bar{\omega}_{k+1}-\hat{\omega}_{k+1}^*\big\|_2^2 &= \big\|\bar{\omega}_{k+1}-\hat{\omega}_k^*+\hat{\omega}_k^*-\hat{\omega}_{k+1}^*\big\|_2^2\\ &= \big\|\bar{\omega}_{k+1}-\hat{\omega}_k^*\big\|_2^2+\big\|\hat{\omega}_k^*-\hat{\omega}_{k+1}^*\big\|_2^2+2\big\langle\bar{\omega}_{k+1}-\hat{\omega}_k^*,\hat{\omega}_k^*-\hat{\omega}_{k+1}^*\big\rangle.\end{aligned} \tag{88}$$

By the non-expansiveness of the projection operator, we have

$$\begin{aligned}\mathbb{E}\big\|\bar{\omega}_{k+1}-\hat{\omega}_k^*\big\|_2^2 &\le \Big\|\bar{\omega}_k-\frac{1}{N}\sum_{n=1}^N Bg_k^n(\omega_k^n)-\hat{\omega}_k^*\Big\|_2^2\\ &\le \mathbb{E}\big\|\bar{\omega}_k-\hat{\omega}_k^*\big\|_2^2+2C_\varphi\beta^2\mathbb{E}\Big\|\frac{1}{N}\sum_{n=1}^N g_k^n(\omega_k^n)\Big\|_2^2-2\mathbb{E}\Big\langle\bar{\omega}_k-\hat{\omega}_k^*,\frac{B}{N}\sum_{n=1}^N g_k^n(\omega_k^n)\Big\rangle\end{aligned} \tag{89}$$

where $C_\varphi \coloneqq (1+\frac{1}{\varphi})$ is the upper bound of $\|B\|^2$. By Lemma 8, for some positive constants C_4, C_5, C_λ, C_g $\tilde{C}_4$, $\tilde{C}_5$ and $\tilde{C}_6$, we can bound the second and third term respectively as

$$\mathbb{E}\Big\|\frac{1}{N}\sum_{n=1}^N g_k^n(\omega_k^n)\Big\|_2^2 \le \frac{2C_g^2}{N}+3(I-1)^2(\tilde{C}_4^2\beta^2+\tilde{C}_5^2\alpha^2)+\tilde{C}_6\mathbb{E}\big\|\bar{\omega}_k-\hat{\omega}_k^*\big\|_2^2, \tag{90}$$

and

$$\begin{aligned}&-\mathbb{E}\Big\langle\bar{\omega}_k-\hat{\omega}_k^*,\frac{B}{N}\sum_{n=1}^N g_k^n(\omega_k^n)\Big\rangle\\ &\le \beta(I-1)\big(C_4\beta+C_5\alpha\big)\mathbb{E}\big\|\bar{\omega}_k-\hat{\omega}_k^*\big\|_2-C_\lambda\beta\mathbb{E}\big\|\bar{\omega}_k-\hat{\omega}_k^*\big\|_2^2\\ &\le \beta\Big(\frac{(I-1)^2C_4^2}{C_\lambda}\beta^2+\frac{(I-1)^2C_5^2}{C_\lambda}\alpha^2+\frac{C_\lambda}{2}\mathbb{E}\|\bar{\omega}_k-\hat{\omega}_k^*\|_2^2\Big)-C_\lambda\beta\mathbb{E}\|\bar{\omega}_k-\hat{\omega}_k^*\|_2^2\\ &\le -\frac{C_\lambda}{2}\beta\mathbb{E}\|\bar{\omega}_k-\hat{\omega}_k^*\|_2^2+\frac{(I-1)^2C_4^2}{C_\lambda}\beta^3+\frac{(I-1)^2C_5^2}{C_\lambda}\alpha^2\beta\end{aligned} \tag{91}$$

where the second inequality follows from Young's inequality. Substituting (90) and (91) into (89), and using $\beta \le C_\beta$ to simplify the inequality give

$$\mathbb{E}\big\|\bar{\omega}_{k+1}-\hat{\omega}_k^*\big\|_2^2 \le (1-C_\lambda\beta+C_6'\beta^2)\mathbb{E}\big\|\bar{\omega}_k-\hat{\omega}_k^*\big\|_2^2+(I-1)^2C_4'\beta^3+(I-1)^2C_5'\alpha^2\beta. \tag{92}$$

where $C_4' \coloneqq \frac{2C_4^2}{C_\lambda}+6C_\psi\tilde{C}_4^2C_\beta$ and $C_5' \coloneqq \frac{2C_5^2}{C_\lambda}+6C_\psi\tilde{C}_5^2C_\beta$.

The third term in (88) can be bounded as

$$\begin{aligned}&\mathbb{E}\big\langle\bar{\omega}_{k+1}-\hat{\omega}_k^*,\hat{\omega}_k^*-\hat{\omega}_{k+1}^*\big\rangle\\ &=\mathbb{E}\big\langle\bar{\omega}_{k+1}-\hat{\omega}_k^*,\hat{\omega}_k^*-\hat{\omega}_{k+1}^*-\langle\nabla\hat{\omega}^*(\bar{\theta}_k),\bar{\theta}_{k+1}-\bar{\theta}_k\rangle\big\rangle\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{E}\big\langle \bar{\omega}_{k+1}-\hat{\omega}_k^*, \langle \nabla\hat{\omega}^*(\bar{\theta}_k), \bar{\theta}_{k+1}-\bar{\theta}_k\rangle\big\rangle\\
&\le \mathbb{E}\Big[\big\|\bar{\omega}_{k+1}-\hat{\omega}_k^*\big\|_2\big\|\hat{\omega}_k^*-\hat{\omega}_{k+1}^*-\langle\nabla\hat{\omega}^*(\bar{\theta}_k),\bar{\theta}_{k+1}-\bar{\theta}_k\rangle\big\|_2\Big]\\
&\quad+\mathbb{E}\big\langle \bar{\omega}_{k+1}-\hat{\omega}_k^*, \langle \nabla\hat{\omega}^*(\bar{\theta}_k), \mathbb{E}\big[\bar{\theta}_{k+1}-\bar{\theta}_k|\mathcal{F}_k\big]\rangle\big\rangle\\
&\le \frac{L_{\omega,2}}{2}\mathbb{E}\Big[\big\|\bar{\omega}_{k+1}-\hat{\omega}_k^*\big\|_2\|\bar{\theta}_{k+1}-\bar{\theta}_k\|_2^2\Big]+L_\omega\alpha\mathbb{E}\Big[\big\|\bar{\omega}_{k+1}-\hat{\omega}_k^*\big\|_2\|p_k\|\Big]
\end{aligned}\tag{93}
$$

where the first inequality follows from the towering property of expectation, and last inequality follows from the L_ω-Lipschitz and $L_{\omega,2}$-smoothness of $\hat{\omega}^*(\theta)$ shown in Lemma 6. Continuing from the last inequality, we have

$$
\begin{aligned}
&\mathbb{E}\big\langle \bar{\omega}_{k+1}-\hat{\omega}_k^*, \hat{\omega}_k^*-\hat{\omega}_{k+1}^*\big\rangle\\
&\le \frac{L_{\omega,2}C_p^2\alpha^2}{2}\mathbb{E}\Big[\big\|\bar{\omega}_{k+1}-\hat{\omega}_k^*\big\|_2\Big]+L_\omega\alpha\mathbb{E}\Big[\big\|\bar{\omega}_{k+1}-\hat{\omega}_k^*\big\|_2\|p_k\|\Big]\\
&\le \frac{C_p^4L_{\omega,2}^2}{4}\alpha^3+\frac{\alpha}{4}\mathbb{E}\|\bar{\omega}_{k+1}-\hat{\omega}_k^*\|_2^2+\frac{\alpha}{8}\mathbb{E}\|p_k\|_2^2+2L_\omega^2\alpha\mathbb{E}\|\bar{\omega}_{k+1}-\hat{\omega}_k^*\|_2^2\\
&=\Big(2L_\omega^2+\frac{1}{2}\Big)\alpha\mathbb{E}\big\|\bar{\omega}_{k+1}-\hat{\omega}_k^*\big\|_2^2+\frac{\alpha}{8}\mathbb{E}\|p_k\|_2^2+\frac{C_p^4L_{\omega,2}^2}{4}\alpha^3.
\end{aligned}\tag{94}
$$

Lastly, the second term in (88) can be bounded as

$$
\|\hat{\omega}_k^*-\hat{\omega}_{k+1}^*\|_2^2\le L_\omega^2\|\bar{\theta}_{k+1}-\bar{\theta}_k\|_2^2.\tag{95}
$$

Substituting the last inequality along with (92) into (88) gives

$$
\begin{aligned}
\mathbb{E}\big\|\bar{\omega}_{k+1}-\hat{\omega}_{k+1}^*\big\|_2^2\le{}&\big(1+(4L_\omega^2+1)\alpha\big)\mathbb{E}\big\|\bar{\omega}_{k+1}-\hat{\omega}_k^*\big\|_2^2+\frac{\alpha}{4}\mathbb{E}\|p_k\|_2^2+\frac{C_p^4L_{\omega,2}^2}{2}\alpha^3\\
&+L_\omega^2\|\bar{\theta}_{k+1}-\bar{\theta}_k\|_2^2.
\end{aligned}\tag{96}
$$

This completes the proof. □

C.3 Supporting lemmas

In this section, we give several supporting lemmas used to prove our main theorems.

C.3.1 Lipschitz continuity and smoothness of optimal solution

The following lemma proves the L-smoothness of $\hat{\omega}^*(\theta)$ w.r.t. θ, which is then used to prove Theorem 4. The idea is inspired by (Chen et al., 2021).

Lemma 6. *Suppose Assumption 2 and 3 hold, then there exist positive constants L_ω and $L_{\omega,2}$ such that*

$$
\big\|\hat{\omega}^*(\theta_1)-\hat{\omega}^*(\theta_2)\big\|_2\le L_\omega\big\|\theta_1-\theta_2\big\|_2,\tag{97a}
$$

and

$$
\big\|\nabla\hat{\omega}^*(\theta_1)-\nabla\hat{\omega}^*(\theta_2)\big\|_2\le L_{\omega,2}\big\|\theta_1-\theta_2\big\|_2.\tag{97b}
$$

Proof. Under Assumption 2, the optimal solution of $\hat{F}_\lambda(\theta,\omega_v,\omega_\mu)$ takes the following form

$$
\hat{\omega}_\mu^*(\theta)=-A_\theta^{-1}b_\theta,\quad \hat{\omega}_v^*(\theta)=-(A_\theta^\top)^{-1}(\lambda CA_\theta^{-1}b_\theta+h).\tag{98}
$$

We write A_1, A_2 and b_1, b_2 as shorthand notations for A_{θ_1}, A_{θ_2} and b_{θ_1}, b_{θ_2} respectively. By definition of b_θ, we have

$$
\begin{aligned}
b_1-b_2&=\mathbb{E}_{s_0\sim\hat{\rho},a_0\sim\pi_1}\big[\phi(s_0,a_0)\big]-\mathbb{E}_{s_0\sim\hat{\rho},a_0\sim\pi_2}\big[\phi(s_0,a_0)\big]\\
&\le 2\sup_{s_0\in\mathcal{S},a_0\in\mathcal{A}}\|\phi(s_0,a_0)\|d_{TV}(\hat{\rho}\otimes\pi_1,\hat{\rho}\otimes\pi_2)\\
&=2d_{TV}(\hat{\rho}\otimes\pi_1,\hat{\rho}\otimes\pi_2),
\end{aligned}\tag{99}
$$

in which

$$d_{TV}(\hat{\rho}\otimes\pi_1,\hat{\rho}\otimes\pi_2)=\frac{1}{2}\sum_{s_0\in\hat{D}_0}\int_{\mathcal{A}}\left|\hat{\rho}(s_0)\pi_1(da_0|s_0)-\hat{\rho}(s_0)\pi_2(da_0|s_0)\right|$$
$$\leq\frac{1}{2}L_\pi|\mathcal{A}|\left\|\theta_1-\theta_2\right\|_2 \tag{100}$$

where the last inequality follow from the L_π-lipschitz continuity of π_θ. Thus we have

$$b_1-b_2\leq L_\pi|\mathcal{A}|\left\|\theta_1-\theta_2\right\|_2. \tag{101}$$

Similarly, we can prove that

$$A_1-A_2\leq 2L_\pi|\mathcal{A}|\left\|\theta_1-\theta_2\right\|_2. \tag{102}$$

By definition of $\hat{\omega}_\mu^*(\theta)$, we have

$$\begin{aligned}\left\|\hat{\omega}_\mu^*(\theta_1)-\hat{\omega}_\mu^*(\theta_2)\right\|_2&=\left\|-A_1^{-1}b_1+A_2^{-1}b_2\right\|_2=\left\|-A_1^{-1}(A_2-A_1)A_2^{-1}b_1+A_2^{-1}(b_2-b_1)\right\|_2\\&\leq\frac{1}{\sigma_{\inf}^2}\|A_1-A_2\|_2+\frac{1}{\sigma_{\inf}}\|b_1-b_2\|\\&\leq L_1\left\|\theta_1-\theta_2\right\|_2\end{aligned} \tag{103}$$

where the last inequality follows from (101) and (102), and $L_1\coloneqq\left(\frac{2}{\sigma_{\inf}^2}+\frac{1}{\sigma_{\inf}}\right)L_\pi|\mathcal{A}|$.

Observing that $\hat{\omega}_v^*(\theta)$ is in a similar form as $\hat{\omega}_\mu^*(\theta)$, it can be proven that similar result holds for $\hat{\omega}_v^*(\theta)$

$$\left\|\hat{\omega}_v^*(\theta_1)-\hat{\omega}_v^*(\theta_2)\right\|_2\leq L_2\left\|\theta_1-\theta_2\right\|_2, \tag{104}$$

where $L_2\coloneqq\left(\frac{2\left(r_{\max}+\frac{\lambda}{\sigma_{\inf}}\right)}{\sigma_{\inf}^2}L_\pi|\mathcal{A}|+\frac{L_1}{\sigma_{\inf}}\right)$.

Given (103) and (104), we have

$$\left\|\hat{\omega}^*(\theta_1)-\hat{\omega}^*(\theta_2)\right\|_2=\sqrt{\left\|\hat{\omega}_v^*(\theta_1)-\hat{\omega}_v^*(\theta_2)\right\|_2^2+\frac{1}{\varphi}\left\|\hat{\omega}_\mu^*(\theta_1)-\hat{\omega}_\mu^*(\theta_2)\right\|_2^2}\leq L_\omega\left\|\theta_1-\theta_2\right\|_2$$

where $L_\omega\coloneqq\sqrt{L_1^2+\frac{1}{\varphi}L_2^2}$. Now we begin to prove (97b). Let ∇_i be a shorthand notation for ∇_{θ_i} where θ_i is the ith element of θ. Then we have

$$\nabla_i\hat{\omega}_\mu^*(\theta)=-A_\theta^{-1}\nabla_i b_\theta+A_\theta^{-1}\nabla_i A_\theta A_\theta^{-1}b_\theta. \tag{105}$$

In order for $\nabla_i\hat{\omega}_\mu^*(\theta_1)$ to be lipschitz continuous, it suffices to show A_θ^{-1}, b_θ, $\nabla_i A_\theta$ and $\nabla_i b_\theta$ are bounded and lipschitz continuous. By previous derivations (101)–(103), A_θ^{-1} and b_θ are indeed bounded lipschitz continuous. Thus it suffices to check $\nabla_i A_\theta$ and $\nabla_i b_\theta$.

$$\nabla_i b_\theta=(1-\gamma)\int_{\mathcal{S}}\hat{\rho}(ds_0)\int_{\mathcal{A}}\nabla_i\pi_\theta(da_0|ds_0)\phi(ds_0,da_0). \tag{106}$$

It is easy to check $\|\nabla_i b_\theta\|\leq C_\psi$. It also holds that

$$\begin{aligned}\|\nabla_i b_{\theta_1}-\nabla_i b_{\theta_2}\|&\leq(1-\gamma)\int_{\mathcal{S}}\hat{\rho}(ds_0)\int_{\mathcal{A}}\left\|\nabla_i\pi_{\theta_1}(da_0|ds_0)-\nabla_i\pi_{\theta_1}(da_0|ds_0)\right\|\\&\leq|\mathcal{A}|(L_\pi C_\psi+L_\psi)\|\theta_1-\theta_2\|\end{aligned} \tag{107}$$

where the last inequality follows from

$$\begin{aligned}\left\|\nabla_i\pi_{\theta_1}(a|s)-\nabla_i\pi_{\theta_1}(a|s)\right\|&=\left\|\pi_{\theta_1}(a|s)\psi_{\theta_1}(s,a)-\pi_{\theta_2}(a|s)\psi_{\theta_2}(s,a)\right\|\\&\leq(L_\pi C_\psi+L_\psi)\|\theta_1-\theta_2\|.\end{aligned} \tag{108}$$

Thus we know $\nabla_i b_\theta$ is bounded lipschitz continuous. Using similar technique, we can check that $\nabla_i A_\theta$ is also bounded lipschitz:

$$\|\nabla_i A_{\theta_1} - \nabla_i A_{\theta_2}\| \leq 2|\mathcal{A}|(L_\pi C_\psi + L_\psi)\|\theta_1 - \theta_2\|, \ \ \|\nabla_i A_\theta\| \leq 2C_\psi. \tag{109}$$

This makes $\nabla_i \hat{\omega}_\mu^*(\theta)$ lipschitz continuous, which implies

$$\|\nabla \hat{\omega}_\mu^*(\theta) - \nabla \hat{\omega}_\mu^*(\theta)\| \leq L_3\|\theta_1 - \theta_2\| \tag{110}$$

for some positive constant L_3. Similarly, we have for some positive constant L_4,

$$\|\nabla \hat{\omega}_v^*(\theta) - \nabla \hat{\omega}_v^*(\theta)\| \leq L_4\|\theta_1 - \theta_2\|. \tag{111}$$

The last two inequality implies (97b) with $L_{\omega,2} = \sqrt{L_3^2 + L_4^2}$. □

C.3.2 Bounding the critic bias and variance

Before we start to bound the critic bias and variance, we first give a lemma on the boundedness of critic variables under projection.

Lemma 7. *Consider running Algorithm 1 with linear function approximation of v^n and μ^n. Under the same assumptions as those of Theorem 3, there exist positive constants C_v and C_μ such that $\|\omega_{v,k}^n\|_2 \leq C_v$ and $\|\omega_{\mu,k}^n\|_2 \leq C_\mu$.*

Proof. To bound $\|\omega_{v,k}^n\|$ and $\|\omega_{\mu,k}^n\|$, it suffices to bound the norm of the concatenated variable $\omega_k^n = [\omega_{v,k}^n \ \ \frac{1}{\sqrt{\varphi}}\omega_{\mu,k}^n]^\top$. By the critic update in (86), we have

$$\begin{aligned}\|\omega_{k+1}^n\|_2 = \|\omega_k^n - Bg_k^n(\omega_k^n)\|_2 &\leq \|\omega_k^n\|_2 + \|Bg_k^n(\omega_k^n)\|_2 \\ &\leq \|\omega_k^n\|_2 + \beta\sqrt{1+\frac{1}{\varphi}}\sqrt{8\varphi+\varphi^2\lambda^2}\|\omega_k^n\|_2 + \sqrt{(1-\gamma)^2+1}.\end{aligned} \tag{112}$$

where the last inequality follows from

$$\begin{aligned}\|Bg_k^n(\omega_k^n)\|_2 &\leq \beta\left\|\begin{bmatrix}1 & 0\\ 0 & \frac{1}{\varphi}\end{bmatrix}\right\|_2\left\|\begin{bmatrix}0 & \sqrt{\varphi}\hat{A}_k^n\\ -\sqrt{\varphi}(\hat{A}_k^n)^\top & \varphi\lambda\hat{C}_k^n\end{bmatrix}\right\|_2\|\omega_{k,n}\|_2 + \left\|\begin{bmatrix}\hat{b}_k^n\\ -\hat{h}_k^n\end{bmatrix}\right\|_2 \\ &\leq \beta\|\omega_{k,n}\|_2\sqrt{1+\frac{1}{\varphi}}\sqrt{2\varphi\big\|\hat{A}_k^n\big\|_F^2 + \varphi^2\lambda^2\big\|\hat{C}_k^n\big\|_F^2} + \sqrt{\big\|\hat{b}_k^n\big\|_2^2 + \big\|\hat{h}_k^n\big\|_2^2} \\ &\leq \beta\sqrt{1+\frac{1}{\varphi}}\sqrt{8\varphi+\varphi^2\lambda^2}\|\omega_k^n\|_2 + \sqrt{(1-\gamma)^2+1}.\end{aligned} \tag{113}$$

Because of the projection in Algorithm 1, for any $k_0 - 1$ which are multiples of I, we have $\|\omega_{k_0}^n\|_2 \leq \sqrt{R_v^2 + \frac{1}{\varphi}R_\mu^2}$. Then by (112), it follows from induction that Lemma 7 holds for any k. □

Now we are ready to give the Lemma that bounds the bias and variance of critic sequence, which is used in the proof of Theorem 4.

Lemma 8. *We choose $\varphi = \frac{8(\gamma+1)^2}{\lambda^2\eta^2}$. Under the same conditions in Theorem 4, we have*

$$\begin{aligned}\mathbb{E}\Big\langle \bar{\omega}_k - \hat{\omega}_k^*, B\mathbb{E}\Big[\frac{1}{N}\sum_{n=1}^N g_k^n(\omega_k^n)\big|\mathcal{F}_k\Big]\Big\rangle \geq &-\beta(I-1)\big(C_4\beta + C_5\alpha\big)\mathbb{E}\big\|\bar{\omega}_k - \hat{\omega}_k^*\big\|_2 \\ &+ C_\lambda\beta\mathbb{E}\big\|\bar{\omega}_k - \hat{\omega}_k^*\big\|_2^2,\end{aligned} \tag{114}$$

and

$$\mathbb{E}\Big\|\frac{1}{N}\sum_{n=1}^N g_k^n(\omega_k^n)\Big\|_2^2 \leq \frac{2C_g^2}{N} + 3(I-1)^2(\tilde{C}_4^2\beta^2 + \tilde{C}_5^2\alpha^2) + \tilde{C}_6\mathbb{E}\big\|\bar{\omega}_k - \hat{\omega}_k^*\big\|_2^2 \tag{115}$$

where C_4, C_5, C_λ, $\tilde{C}_4$, $\tilde{C}_5$ and $\tilde{C}_6$ are some positive constants.

Proof. We first prove (114). We denote its LHS as I_1. I_1 can be decomposed as

$$I_1 \geq -\Big(1+\frac{1}{\sqrt{\varphi}}\Big)\beta\mathbb{E}\Big[\big\|\bar{\omega}_k-\hat{\omega}_k^*\big\|_2\underbrace{\Big\|\mathbb{E}\Big[\frac{1}{N}\sum_{n=1}^N g_k^n(\omega_k^n)\big|\mathcal{F}_k\Big]-\mathbb{E}\Big[\frac{1}{N}\sum_{n=1}^N g_k^n(\bar{\omega}_k)\big|\mathcal{F}_k\Big]\Big\|_2}_{I_1^{(1)}}\Big]$$
$$-\Big(1+\frac{1}{\sqrt{\varphi}}\Big)\beta\mathbb{E}\Big[\big\|\bar{\omega}_k-\hat{\omega}_k^*\big\|_2\underbrace{\Big\|\mathbb{E}\Big[\frac{1}{N}\sum_{n=1}^N g_k^n(\bar{\omega}_k)\big|\mathcal{F}_k\Big]-g_k(\bar{\omega}_k)\Big\|_2}_{I_1^{(2)}}\Big]+\underbrace{\mathbb{E}\Big\langle\bar{\omega}_k-\hat{\omega}_k^*,Bg_k(\bar{\omega}_k)\Big\rangle}_{I_1^{(3)}} \tag{116}$$

where $g_k(\omega)$ is defined as

$$\begin{aligned}g_k(\omega)&\coloneqq\frac{1}{N}\sum_{n=1}^N\begin{bmatrix}1&0\\0&\frac{1}{\sqrt{\varphi}}\end{bmatrix}\begin{bmatrix}0&\sqrt{\varphi}A_{\bar{\theta}_k}^n\\-\sqrt{\varphi}(A_{\bar{\theta}_k}^n)^\top&\lambda\varphi C^n\end{bmatrix}\omega+\begin{bmatrix}b_{\bar{\theta}_k}^n\\-h^n\end{bmatrix}\\&=\begin{bmatrix}1&0\\0&\frac{1}{\sqrt{\varphi}}\end{bmatrix}\underbrace{\begin{bmatrix}0&\sqrt{\varphi}A_{\bar{\theta}_k}\\-\sqrt{\varphi}(A_{\bar{\theta}_k})^\top&\varphi\lambda C\end{bmatrix}}_{G_{\bar{\theta}_k}}\omega+\begin{bmatrix}b_{\bar{\theta}_k}\\-h\end{bmatrix}.\end{aligned}$$

With $\Sigma\coloneqq\begin{bmatrix}1&0\\0&\frac{1}{\sqrt{\varphi}}\end{bmatrix}$, $I_1^{(1)}$ can be bounded as

$$\begin{aligned}I_1^{(1)}&\leq\frac{1}{N}\sum_{n=1}^N\big\|g_k^n(\omega_k^n)-g_k^n(\bar{\omega}_k)\big\|_2=\frac{1}{N}\sum_{n=1}^N\big\|\Sigma G_{\theta_k^n}^n(\omega_k^n-\bar{\omega}_k)\big\|_2\\&\leq C_G\frac{1}{N}\sum_{n=1}^N\big\|\omega_k^n-\bar{\omega}_k\big\|_2\\&\leq\tilde{C}_4(I-1)\beta,\end{aligned}\tag{117}$$

where the last inequality follows from Lemma 5, $C_G\coloneqq\Big(1+\frac{1}{\sqrt{\varphi}}\Big)\sqrt{8\varphi+\lambda^2\varphi^2}$ is the upper bound of $\|\Sigma G_\theta^n\|_2$, and $\tilde{C}_4\coloneqq 2\Big(1+\frac{1}{\varphi}\Big)C_GC_q$.

$I_1^{(2)}$ can be bounded as

$$\begin{aligned}I_1^{(2)}&=\Big\|\frac{1}{N}\sum_{n=1}^N\begin{bmatrix}0&\sqrt{\varphi}(A_{\theta_k^n}^n-A_{\bar{\theta}_k}^n)\\-(A_{\theta_k^n}^n-A_{\bar{\theta}_k}^n)^\top&0\end{bmatrix}\bar{\omega}_k+\begin{bmatrix}b_{\theta_k^n}^n-b_{\bar{\theta}_k}^n\\0\end{bmatrix}\Big\|_2\\&\leq\frac{1}{N}\sum_{n=1}^N\Big(\sqrt{1+\frac{1}{\varphi}}C_\omega\big\|A_{\theta_k^n}^n-A_{\bar{\theta}_k}^n\big\|_2+\big\|b_{\theta_k^n}^n-b_{\bar{\theta}_k}^n\big\|_2\Big),\end{aligned}\tag{118}$$

where the last inequality is due to the fact that $\|\bar{\omega}_k\|_2\leq\frac{1}{N}\sum_{n=1}^N\|\omega_k^n\|_2$ and $\|\omega_k^n\|_2\leq C_\omega\coloneqq\sqrt{C_v^2+\frac{1}{\varphi}C_\mu^2}$. It can be easily verified by following the derivation of (101) that

$$\big\|A_{\theta_k^n}^n-A_{\bar{\theta}_k}^n\big\|_2\leq 2L_\pi|\mathcal{A}|\|\theta_k^n-\bar{\theta}_k\|_2,\ \ \big\|b_{\theta_k^n}^n-b_{\bar{\theta}_k}^n\big\|_2\leq L_\pi|\mathcal{A}|\|\theta_k^n-\bar{\theta}_k\|_2.\tag{119}$$

Substituting the last inequality into (118) gives

$$\begin{aligned}I_1^{(2)}&\leq\frac{2L_\pi|\mathcal{A}|}{N}\sum_{n=1}^N\Big(\sqrt{1+\frac{1}{\varphi}}C_\omega\big\|\theta_k^n-\bar{\theta}_k\big\|_2+\big\|\theta_k^n-\bar{\theta}_k\big\|_2\Big)\\&\leq\tilde{C}_5(I-1)\alpha,\end{aligned}\tag{120}$$

where the last inequality follows from Lemma 5 and $\tilde{C}_5\coloneqq 4L_\pi|\mathcal{A}|\Big(1+\sqrt{1+\frac{1}{\varphi}}C_\omega\Big)C_p$.

To bound $I_1^{(3)}$, we need to first bound the eigenvalues of G_θ. Let $\varphi = \frac{8(\gamma+1)^2}{\lambda^2\eta^2}$, then it is easy to verify that $\varphi \geq \frac{8\lambda_{\max}(A_\theta(\lambda C)^{-1}A_\theta^\top)}{\lambda_{\min}(\lambda C)}$ for any $\theta \in \mathbb{R}^d$, where $\lambda_{\min}(\cdot), \lambda_{\max}(\cdot)$ denotes the minimum and maximum eigenvalue. Then by the analysis in section A.1 and A.3 of (Du et al., 2017), we have $\lambda_{\min}(G_\theta) \geq \frac{8}{9}\lambda_{\min}(A_\theta(\lambda C)^{-1}A_\theta^\top)$. Under Assumption 2, we can also show that $\lambda_{\min}(A_\theta(\lambda C)^{-1}A_\theta^\top) \geq \frac{\sigma_{\inf}^2}{\lambda}$. This leads to $\lambda_{\min}(G_\theta) \geq \lambda_{\inf} \coloneqq \frac{8}{9}\frac{\sigma_{\inf}^2}{\lambda}$.

By the optimality condition of $\hat{\omega}_k^*$, we have

$$\begin{aligned} I_1^{(3)} &= \Big\langle \bar{\omega}_k - \hat{\omega}_k^*, B\big(g_k(\bar{\omega}_k) - g_k(\hat{\omega}_k^*)\big)\Big\rangle = \beta\Big\langle \bar{\omega}_k - \hat{\omega}_k^*, \Sigma^2 G_{\bar{\theta}_k}\big(\bar{\omega}_k - \hat{\omega}_k^*\big)\Big\rangle \\ &\geq \beta \min\{1, \frac{1}{\varphi}\}\lambda_{\inf}\big\|\bar{\omega}_k - \hat{\omega}_k^*\big\|. \end{aligned} \tag{121}$$

Substituting the lower bounds in (117), (120) and (121) into (116) gives

$$I_1 \geq -\beta(I-1)\big(C_4\beta + C_5\alpha\big)\mathbb{E}\big\|\bar{\omega}_k - \hat{\omega}_k^*\big\|_2 + C_\lambda\beta\mathbb{E}\big\|\bar{\omega}_k - \hat{\omega}_k^*\big\|_2^2 \tag{122}$$

where $C_\lambda \coloneqq \min\{1, \frac{1}{\varphi}\}\lambda_{\inf}$, $C_4 \coloneqq \big(1 + \frac{1}{\sqrt{\varphi}}\big)\tilde{C}_4$ and $C_5 \coloneqq \big(1 + \frac{1}{\sqrt{\varphi}}\big)\tilde{C}_5$. This completes the proof of (114).

Now we prove (115).

$$\begin{aligned} \mathbb{E}\Big\|\frac{1}{N}\sum_{n=1}^N g_k^n(\omega_k^n)\Big\|_2^2 &= \frac{1}{N^2}\mathbb{E}\Big\|\sum_{n=1}^N \big(g_k^n(\omega_k^n) - \mathbb{E}[g_k^n(\omega_k^n)|\mathcal{F}_k]\big)\Big\|_2^2 + \mathbb{E}\Big\|\frac{1}{N}\sum_{n=1}^N \mathbb{E}\big[g_k^n(\omega_k^n)|\mathcal{F}_k\big]\Big\|_2^2 \\ &= \frac{1}{N^2}\sum_{n=1}^N \mathbb{E}\Big\|g_k^n(\omega_k^n) - \mathbb{E}[g_k^n(\omega_k^n)|\mathcal{F}_k]\Big\|_2^2 + \mathbb{E}\Big\|\frac{1}{N}\sum_{n=1}^N \mathbb{E}\big[g_k^n(\omega_k^n)|\mathcal{F}_k\big]\Big\|_2^2 \\ &\leq \frac{2C_g^2}{N} + \mathbb{E}\Big\|\frac{1}{N}\sum_{n=1}^N \mathbb{E}\big[g_k^n(\omega_k^n)|\mathcal{F}_k\big]\Big\|_2^2, \end{aligned} \tag{123}$$

where the second equality is due to the zero mean of $g_k^n(\omega_k^n) - \mathbb{E}[g_k^n(\omega_k^n)|\mathcal{F}_k]$ and its conditional independence across clients $n \in \{1, 2, ..., N\}$, and the last inequality is due to $\|g_k^n(\omega_k^n)\|_2 \leq C_g \coloneqq C_G C_\omega + \sqrt{r_{\max}^2 + 1}$.

We now consider the second term in (123).

$$\Big\|\frac{1}{N}\sum_{n=1}^N \mathbb{E}\big[g_k^n(\omega_k^n)|\mathcal{F}_k\big]\Big\|_2^2 \leq 3(I_1^{(1)})^2 + 3(I_1^{(2)})^2 + 3\Big\|g_k(\bar{\omega}_k)\Big\|_2^2 \tag{124}$$

where the first and second term are bounded in (117) and (120). It suffices to just consider the last term, which can be bounded as

$$\Big\|g_k(\bar{\omega}_k)\Big\|_2^2 = \Big\|g_k(\bar{\omega}_k) - g_k(\hat{\omega}_k^*)\Big\|_2^2 \leq \Big(1 + \frac{1}{\varphi}\Big)C_G^2\big\|\bar{\omega}_k - \hat{\omega}_k^*\big\|_2^2. \tag{125}$$

Thus we have

$$\Big\|\frac{1}{N}\sum_{n=1}^N \mathbb{E}\big[g_k^n(\omega_k^n)|\mathcal{F}_k\big]\Big\|_2^2 \leq 3(I-1)^2(\tilde{C}_4^2\beta^2 + \tilde{C}_5^2\alpha^2) + 3\Big(1 + \frac{1}{\varphi}\Big)C_G^2\big\|\bar{\omega}_k - \hat{\omega}_k^*\big\|_2^2. \tag{126}$$

Substituting (126) into (123) completes the proof of (115). □

C.3.3 Bounding the statistical errors

We first give a useful inequality which will be used to bound the statistical errors.

Lemma 9 (Pollard's tail inequality (Pollard, 1984)). *Let $\mathcal{F} : \mathcal{Z} \mapsto [-R, R]$ be a permissible class of functions. Let $\{z_i\}_{i=1}^M \in \mathcal{Z}$ be i.i.d samples from a distribution. For any $\epsilon > 0$, we have*

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}}\Big|\frac{1}{M}\sum_{i=1}^M f(z_i) - \mathbb{E}\big[f(z_i)\big]\Big|\Big) \leq 8\mathcal{N}_1\Big(\frac{\epsilon}{8}, \mathcal{F}, \{z_i\}_{i=1}^M\Big)\exp\Big(\frac{-M\epsilon^2}{512R^2}\Big) \tag{127}$$

where $\mathcal{N}_1(\cdot)$ is the covering number.

The covering number can be bounded by the following inequality.

Lemma 10 (Haussler's inequality (Haussler, 1995))**.** *For any set $\mathcal{Z}$, any points $\{z_i\}_{i=1}^M \in \mathcal{Z}$, any function class $\mathcal{F}$ on $\mathcal{Z}$ taking values in $[-R, R]$ with pseudo-dimension $D_{\mathcal{F}} < \infty$, we have*

$$\mathcal{N}_1\Big(\epsilon, \mathcal{F}, \{z_i\}_{i=1}^M\Big) \leq e(D_{\mathcal{F}}+1)\Big(\frac{2eR}{\epsilon}\Big)^{\mathcal{D}_{\mathcal{F}}}. \tag{128}$$

Now we can give the bound on statistical errors.

Lemma 11. *Under the same conditions as that of Theorem 3, for any $\theta \in \mathbb{R}^d$, with at least probability $1-\delta$ we have*

$$\Big\|\hat{\mathbb{E}}\big[\hat{\mu}_\theta^*(s,a)\hat{v}_\theta^*(s,a)\psi_\theta(s,a)\big] - \bar{\mathbb{E}}\big[\hat{\mu}_\theta^*(s,a)\hat{v}_\theta^*(s,a)\psi_\theta(s,a)\big]\Big\|_2^2 = \mathcal{O}\Big(\frac{\log(NM)+\log(\frac{d}{\delta})}{NM}\Big).$$

Proof. Define $y(s,a) \coloneqq \mu(s,a)v(s,a)\psi_\theta(s,a)] \in \mathcal{F}_y$, where $v \in \mathcal{F}_v$ and $\mu \in \mathcal{F}_\mu$. For any $\theta \in \mathbb{R}^d$, we have

$$\begin{aligned}&\Big\|\hat{\mathbb{E}}\big[\hat{\mu}_\theta^*(s,a)\hat{v}_\theta^*(s,a)\psi_\theta(s,a)\big] - \bar{\mathbb{E}}\big[\hat{\mu}_\theta^*(s,a)\hat{v}_\theta^*(s,a)\psi_\theta(s,a)\big]\Big\|_2^2 \\ &\leq \sup_{y\in\mathcal{F}_y}\big\|\mathbb{E}_{s,a\sim\hat{d}_D}\big[y(s,a,s')\big] - \mathbb{E}_{s,a\sim\bar{d}_D}\big[y(s,a,s')\big]\big\|_2^2.\end{aligned} \tag{129}$$

where the last inequality is due to $\hat{v}_\theta^* \in \mathcal{F}_v$ and $\hat{\mu}_\theta^* \in \mathcal{F}_\mu$. With $\mathbf{1}_i \in \mathbb{R}^d$ defined as the vector whose ith element is 1 and the the rest elements are 0, we have

$$\begin{aligned}&\sup_{y\in\mathcal{F}_y}\big\|\mathbb{E}_{s,a\sim\hat{d}_D}\big[y(s,a)\big] - \mathbb{E}_{s,a\sim\bar{d}_D}\big[y(s,a)\big]\big\|_2^2 \\ &\leq \sum_{i=1}^d \sup_{y\in\mathcal{F}_y}\big(\mathbb{E}_{s,a\sim\hat{d}_D}\big[\mathbf{1}_i^\top y(s,a)\big] - \mathbb{E}_{s,a\sim\bar{d}_D}\big[\mathbf{1}_i^\top y(s,a)\big]\big)^2\end{aligned} \tag{130}$$

For any $i \in \{1,2,...,d\}$, we have $|\mathbf{1}_i^\top y(s,a,s')| \leq R_1 \coloneqq R_\mu R_v C_\psi$. Denote $\mathcal{F}_{y_i}$ as the function class of $\mathbf{1}_i^\top y(s,a,s')$. By Pollard's tail inequality in Lemma 9, we have

$$\begin{aligned}&\mathbb{P}\Big(\sup_{y\in\mathcal{F}_y}\Big|\mathbb{E}_{\hat{d}_D}\big[\mathbf{1}_i^\top y(s,a)\big] - \mathbb{E}_{\bar{d}_D}\big[\mathbf{1}_i^\top y(s,a)\big]\Big| \geq \epsilon\Big) \\ &\leq 8\mathbb{E}\Big[\mathcal{N}_1\Big(\frac{\epsilon}{8}, \mathcal{F}_{y_i}, \{\mathcal{D}^n\}_{n=1}^N\Big)\Big]\exp\big(\frac{-NM\epsilon^2}{512R_1^2}\big) \\ &\leq 8C'(i)\Big(\frac{1}{\epsilon}\Big)^{D(i)}.\end{aligned} \tag{131}$$

where the last inequality follows from Haussler's inequality in Lemma 10. Constant $R'(i) = e(D(i)+1)^2(2eR_1)^{D(i)}$ and $D(i)$ is pseudo-dimension of $\mathcal{F}_{y_i}$. Let $\epsilon_i = \sqrt{\frac{R(i)(log(NM)+\log\frac{d}{\delta})}{NM}}$ where $R(i) = \max\big((8R'(i))^{\frac{2}{D(i)}}, 512NMD(i), 512NM, 1\big)$, with at least probability $1-\delta/d$ we have

$$\sup_{y\in\mathcal{F}_y}\Big|\mathbb{E}_{\hat{d}_D}\big[\mathbf{1}_i^\top y(s,a)\big] - \mathbb{E}_{\bar{d}_D}\big[\mathbf{1}_i^\top y(s,a)\big]\Big| \leq \epsilon_i. \tag{132}$$

Let $\epsilon = \sqrt{\frac{\bar{R}(log(NM)+\log\frac{d}{\delta})}{NM}}$ with $\bar{R} = \max_i R(i)$. Then, substituting (132) into (130) gives

$$\sup_{y\in\mathcal{F}_y}\big\|\mathbb{E}_{s,a\sim\hat{d}_D}\big[y(s,a)\big] - \mathbb{E}_{s,a\sim\bar{d}_D}\big[y(s,a)\big]\big\|_2^2 \leq d\epsilon^2 \tag{133}$$

with probability at least $1-\delta$. This along with (129) completes the proof. □

Lemma 12. *Under the same conditions of that in Theorem 3, with probability at least $1-\delta$ we have*

$$\sup_{\theta\in\mathbb{R}^d, v\in\mathcal{F}_v, \mu\in\mathcal{F}_\mu}\Big|F_\lambda(\theta,v,\mu) - \hat{F}_\lambda(\theta,v,\mu)\Big| = \mathcal{O}\Bigg(\sqrt{\frac{\log(NM)+\log(\frac{1}{\delta})}{NM}}\Bigg) \tag{134}$$

Proof. Define function $l(X) = \mathbb{E}_{a_0\sim\pi_\theta(\cdot|s_0)}[v(s_0,a_0)] + \big(r(s,a)+\gamma\mathbb{E}_{a'\sim\pi_\theta(\cdot|s')}[v(s',a')] - v(s,a)\big)\mu(s,a) - \frac{\lambda}{2}\mu(s,a)^2 \in \mathcal{F}_l$, where $X := (s_0,s,a,s')$ and $v\in\mathcal{F}_v$, $\mu\in\mathcal{F}_\mu$. We have

$$\begin{aligned}&\sup_{\theta\in\mathbb{R}^d, v\in\mathcal{F}_v,\mu\in\mathcal{F}_\mu}\Big|F_\lambda(\theta,v,\mu) - \hat{F}_\lambda(\theta,v,\mu)\Big|\\ &= \sup_{l\in\mathcal{F}_l}\Big|\mathbb{E}_{s_0\sim\rho,s,a,s'\sim\bar{d}_D}\big[l(X)\big] - \mathbb{E}_{s_0\sim\hat{\rho},s,a,s'\sim\hat{d}_D}\big[l(X)\big]\Big|.\end{aligned} \tag{135}$$

By Pollard's tail inequality in Lemma 9

$$\begin{aligned}\mathbb{P}\bigg(\sup_{l\in\mathcal{F}_l}\Big|\mathbb{E}_{\rho,\bar{d}_D}\big[l(X)\big] - \mathbb{E}_{\hat{\rho},\hat{d}_D}\big[l(X)\big]\Big| \geq \epsilon\bigg) &\leq 8\mathbb{E}\bigg[\mathcal{N}_1\bigg(\frac{\epsilon}{8},\mathcal{F}_l,\{\mathcal{D}^n\}_{n=1}^N\cup\{\mathcal{D}_0^n\}_{n=1}^N\bigg)\bigg]\exp\big(\frac{-NM\epsilon^2}{512R_2^2}\big)\\ &\leq 8R_3\bigg(\frac{1}{\epsilon}\bigg)^{D_1}\end{aligned} \tag{136}$$

where the last inequality follows from Haussler's inequality in Lemma 10. Constant $R_3 = e(D_1+1)^2(2eR_2)^{D_1}$ and D_1 is pseudo-dimension of $\mathcal{F}_l$. Let $\epsilon = \sqrt{\frac{R_4(log(NM)+\log\frac{1}{\delta})}{NM}}$ where $R_4 = \max\big((8R_3)^{\frac{2}{D_1}}, 512NMD_1, 512NM, 1\big)$, with probability at least $1-\delta$, we have

$$\sup_{l\in\mathcal{F}_l}\Big|\mathbb{E}_{\rho,\bar{d}_D}\big[l(X)\big] - \mathbb{E}_{\hat{\rho},\hat{d}_D}\big[l(X)\big]\Big| \leq \epsilon$$

which along with (135) completes the proof. □

D Proof of Theorem 2

Before we proceed with the proof, we first introduce a gradient dominance type property of $F(\theta)$:

Proposition 2 ((Agarwal et al., 2020, Theorem 5.3)). *Under softmax policy parametrization and uniform priors, if* $\|\nabla J_\tau(\theta)\| \leq \frac{\tau}{2|\mathcal{S}||\mathcal{A}|}$, *then* $J^* - J(\theta) \leq \frac{2\tau}{1-\gamma}\big\|\frac{d_{\pi^*}}{\rho}\big\|_\infty$.

Now we are ready to prove Theorem 2.

Proof. It is known that the softmax policy satisfies Assumption 3, thus we immediately know that Theorem 1 holds. Furthermore, by assumption 4, we have the optimal solution v_θ^* and μ_θ^* falls in the linear function class with respect to the feature, thus we know the function approximation error ϵ_{app} disappears. By Theorem 1, the following inequality holds with probability at least $1-\delta$

$$\sum_{k=1}^K \mathbb{E}\big\|\nabla J_\tau(\bar{\theta}_k)\big\|_2^2 = \mathcal{O}\bigg(\sqrt{\frac{K}{N}}\bigg) + \tilde{\mathcal{O}}\bigg(K\sqrt{\frac{\log\frac{3}{\delta}}{NM}}\bigg) + \mathcal{O}\bigg(K\tau^3\frac{C_d^2}{(1-\gamma)^2}\bigg) \tag{137}$$

where we have used the fact $C_\psi = 1$ for softmax policy, and $\lambda = \tau^{\frac{3}{2}}$. We define an event E_k as $\|\nabla J_\tau(\theta_k)\| \leq \frac{\tau}{2|\mathcal{S}||\mathcal{A}|}$ and its complement E_k^c as $\|\nabla J_\tau(\theta_k)\| > \frac{\tau}{2|\mathcal{S}||\mathcal{A}|}$. We use $\mathbf{1}_{E_k}$ to indicate whether the event happens or not, i.e. $\mathbf{1}_{E_k} = 1$ if E_k happens and $\mathbf{1}_{E_k} = 0$ if E_k^c happens. Then we have for the optimality gap:

$$\begin{aligned}\sum_{k=1}^K \mathbb{E}\big[J^* - J(\theta_k)\big] &= \sum_{k=1}^K \mathbb{E}\big[\big(J^* - J(\theta_k)\big)\mathbf{1}_{E_k}\big] + \sum_{k=1}^K \mathbb{E}\big[\big(J^* - J(\theta_k)\big)\mathbf{1}_{E_k^c}\big]\\ &\leq \frac{2\tau}{1-\gamma}\Big\|\frac{d_{\pi^*}}{\rho}\Big\|_\infty \sum_{k=1}^K \mathbb{E}\big[\mathbf{1}_{E_k}\big] + \sum_{k=1}^K \mathbb{E}\big[\big(J^* - J(\theta_k)\big)\mathbf{1}_{E_k^c}\big]\\ &\leq \frac{2\tau}{1-\gamma}\Big\|\frac{d_{\pi^*}}{\rho}\Big\|_\infty \sum_{k=1}^K \mathbb{E}\big[\mathbf{1}_{E_k}\big] + \sum_{k=1}^K \mathbb{E}\big[\mathbf{1}_{E_k^c}\big]\\ &\leq \frac{2\tau}{1-\gamma}\Big\|\frac{d_{\pi^*}}{\rho}\Big\|_\infty K + \sum_{k=1}^K \mathbb{E}\big[\mathbf{1}_{E_k^c}\big],\end{aligned} \tag{138}$$

where the first inequality follows from proposition 2, and the second inequality is due to the fact that $\sup_\pi F(\pi) \leq \max_{s,a} r(s,a) \leq 1$.

Now it suffices to bound $\sum_{k=1}^K \mathbb{E}\big[\mathbf{1}_{E_k^c}\big]$.

$$\begin{aligned}\sum_{k=1}^K \mathbb{E}\|\nabla J_\tau(\theta_k)\|^2 &\geq \sum_{k=1}^K \mathbb{E}\big[\|\nabla J_\tau(\theta_k)\|^2 \mathbf{1}_{E_k^c}\big] \\ &\geq \sum_{k=1}^K \frac{\tau^2}{4|\mathcal{S}|^2|\mathcal{A}|^2}\mathbb{E}\big[\mathbf{1}_{E_k^c}\big] \end{aligned} \tag{139}$$

which along with (137) implies

$$\sum_{k=1}^K \mathbb{E}\big[\mathbf{1}_{E_k^c}\big] = \mathcal{O}\bigg(\sqrt{\frac{K}{N}}\bigg) + \tilde{\mathcal{O}}\bigg(K\sqrt{\frac{\log\frac{3}{\delta}}{NM}}\bigg) + \mathcal{O}\bigg(K\tau\frac{C_d^2}{(1-\gamma)^2}\bigg). \tag{140}$$

Substituting (140) into (138), and dividing both sides by K give that the following inequality holds with probability greater than $1-\delta$ that

$$\frac{1}{K}\sum_{k=1}^K \mathbb{E}\big[J^* - J(\theta_k)\big] = \mathcal{O}\Big(\frac{1}{\sqrt{NK}}\Big) + \tilde{\mathcal{O}}\bigg(\sqrt{\frac{\log\frac{3}{\delta}}{NM}}\bigg) + \mathcal{O}\big(\epsilon_\tau\big), \tag{141}$$

which completes the proof. □

E Additional experiment details

We first present our choice of hyperparameters in the tests. All the tests are conducted in system with a 16-core CPU, an NVIDIA Geforce RTX 2080s, a Titan V and a NVIDIA Geforce RTX 3080.

Prior to training, we collect data with behavior policies that are described in Section 5. We use a data set size of 80000 (Navaigation), 100000 (cartpole) and 40000 (Frozenlake). In all tests, the critic functions v and μ are parametrized by a 3-layer neural network with a hidden dimension of 64×64 and the ReLU activation. In the cartpole and frozenlake tests, we use a 3-layer neural network with a 128×128 hidden dimension; ReLU activation and a softmax output function. In the navigation test, we use a natural softmax policy parametrization. All networks are initialized randomly using the glorot uniform. To ensure the stability of our method, we clip the gradient element-wise within 1.0.

We select $\lambda = 10^{-6}$ for all tests. In the tests of Figure 2, we select an initial step size of $\alpha = 0.0001, 0.00007, 0.00003$, $\beta = 0.0001, 0.00007, 0.00005$ for navigation, cartpole and frozenlake respectively. We set batch size as 1024 for each client and the communication interval $I = 10$. In the tests of Figure 3, we select an initial step size of $\alpha = 0.0001, 0.00007, 0.00002$, $\beta = 0.0001, 0.00007, 0.00005$ for navigation, cartpole and frozenlake respectively, and then scale it with $\sqrt{N}$ for different number of clients. The batch size is 1024 (navigation, cartpole); 512 (frozenlake) for each client and communication interval $I = 10$. In the tests of Figure 4, we choose the same hyperparameters as that of speedup tests except for a varying communication interval.