Journal of Machine Learning Research 26 (2025) 1-45 Submitted 2/24; Revised 8/25; Published 11/25

Client Selection for Federated Policy Optimization with Environment Heterogeneity

Zhijie Xie ZHIJIE.XIE@CONNECT.UST.HK
Shenghui Song* EESHSONG@UST.HK
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong

Editor: Shiqian Ma

Abstract

The development of Policy Iteration (PI) has inspired many recent algorithms for Reinforcement Learning (RL), including several policy gradient methods that gained both theoretical soundness and empirical success on a variety of tasks. The theory of PI is rich in the context of centralized learning, but its study under the federated setting is still in the infant stage. This paper investigates the federated version of Approximate PI (API) and derives its error bound, taking into account the approximation error introduced by environment heterogeneity. We theoretically prove that a proper client selection scheme can reduce this error bound. Based on the theoretical result, we propose a client selection algorithm to alleviate the additional approximation error caused by environment heterogeneity. Experiment results show that the proposed algorithm outperforms other biased and unbiased client selection methods on the federated mountain car problem, the MuJoCo Hopper problem, and the SUMO-based autonomous vehicle training problem by effectively selecting clients with a lower level of heterogeneity from the population distribution.

Keywords: federated reinforcement learning, client selection, data heterogeneity, policy iteration, communication efficiency

1. Introduction

Reinforcement Learning (RL) has been applied to many real-world applications ranging from gaming and robotics to recommender systems (Silver et al., 2016; Chen et al., 2019). However, single-agent RL often suffers from poor sample efficiency, resulting in slow convergence and a high cost of sample collection (Ciosek and Whiteson, 2020; Fan et al., 2021; Papini et al., 2018). Therefore, it is desirable to deploy RL algorithms to large-scale and distributed systems where multiple agents can contribute to the learning collaboratively. However, Multi-Agent RL (MARL) (Zhang et al., 2021) and parallel RL (Nair et al., 2015; Mnih et al., 2016) require intensive communication among agents or data sharing, which may not be practical due to both the communication bottleneck and privacy concerns of many real-world applications. For example, privacy is a major concern in autonomous driving (Liang et al., 2023; Li et al., 2022; Xianjia et al., 2021), and sharing data among vehicles is not allowed. To this end, Federated Learning (FL) (McMahan et al., 2017; Kairouz et al.,

*. Corresponding author.

2021), which enables multiple clients to jointly train a global model without violating user privacy, is an appealing solution for addressing the sample inefficiency and privacy issue of RL in innovative applications such as autonomous driving, IoT network, and healthcare (Zhou et al., 2024). As a result, Federated RL (FRL) has attracted much research attention (Qi et al., 2021).

Despite the significant progress of empirical works on FRL (Qi et al., 2021), the community's understanding of FRL is still in its infancy, especially from the theoretical perspective. For example, the sample efficiency of Policy Gradient (PG) methods is typically low due to the large variance in gradient estimation. This issue could be exacerbated in the context of FL, where clients with heterogeneous environments can generate a diverse range of trajectories. To address this problem, a variance-reduced policy gradient method, namely Federated Policy Gradient with Byzantine Resilience (FedPG-BR), was proposed together with an analysis of the sample efficiency and convergence guarantee (Fan et al., 2021). While clients are assumed to be homogeneous in FedPG-BR, another line of work, termed FedKL (Xie and Song, 2023), noticed that the environment heterogeneity imposes an extra layer of difficulty in learning and proved that a Kullback-Leibler (KL) penalized local objective can generate a monotonically improving sequence of policies to accelerate convergence. The authors of QAvg & PAvg (Jin et al., 2022) provided a convergence proof for the federated Q-Learning and federated PG. QAvg offered important insights regarding how the Bellman operators can be generalized to the federated setting and proposed a useful tool, i.e., the imaginary environment (the average of all clients' environments), for analyzing FRL. More recently, FedTD (Wang et al., 2024), FedSARSA (Zhang et al., 2024), and FedSynQ (Woo et al., 2025) studied the integration of FRL and Temporal Difference (TD) learning algorithms. However, there has not been any convergence analysis regarding Policy Iteration (PI) in FRL in the literature. Given PI's application and theoretical importance, it is desirable to fill this knowledge gap and derive efficient FRL algorithms accordingly.

Among existing RL methods, PI is one of the most popular ones and serves as the foundation of many policy optimization methods, e.g., Safe Policy Iteration (SPI) (Pirotta et al., 2013; Metelli et al., 2021), Trust Region Policy Optimisation (TRPO) (Schulman et al., 2015), and Deep Conservative Policy Iteration (DCPI) (Vieillard et al., 2020). With exact PI, convergence to the optimal policy is guaranteed under mild conditions. However, exact policy evaluation and policy improvement are normally impractical. With Approximate Policy Iteration (API) (Bertsekas, 2022; Bertsekas and Tsitsiklis, 1996), it is assumed that the approximation error is inevitable, and only estimates of the value function and improved policy with bounded errors are available. In the presence of these approximation errors, convergence is not ensured, but the difference in value functions between the generated policy and the optimal policy is bounded (Bertsekas, 2022). In some cases, the algorithm ends up generating a cycle of policies, which is called the policy oscillation/chattering phenomenon (Bertsekas, 2011; Wagner, 2011). Unfortunately, FRL with heterogeneous environments introduces an extra error into the policy iteration process, making the associated analysis more challenging. As will be shown in the following sections, this error is proportional to the level of heterogeneity of the system, and client selection is an effective way to alleviate this problem.

There exist various client selection schemes for Federated Supervised Learning (FSL), and most of them can be classified into two categories: (1) unbiased client selection, and

(2) biased client selection. Convergence guarantee for both schemes has been studied and generalized to tackle the heterogeneity issue of FSL (Nishio and Yonetani, 2019; Li et al., 2020; Li et al., 2020; Jee Cho et al., 2022). However, to the best of the authors' knowledge, there is no known client selection scheme specifically designed to tackle the heterogeneity issue of FRL.

Contributions. In this paper, we derive the error bound of Federated Approximate Policy Iteration (FAPI) under heterogeneous environments, which is not yet available in the literature. The derived error bound takes the level of heterogeneity and parameter aggregation into consideration and explicitly reveals their impacts. Based on the error bound, we propose a client selection algorithm to improve the convergence speed of federated policy optimization. The efficacy of the proposed algorithm is validated on the federated mountain car problem, the MuJoCo Hopper problems, and the SUMO-based autonomous vehicle training problem.

2. Background

In Section 2.1, we introduce the optimization problem of FRL. In Section 2.2, we review some known results on API. An imaginary environment is introduced in Section 2.3 to assist the following analysis.

2.1 Federated Reinforcement Learning

The system setup of FRL in this paper is similar to that of FL (McMahan et al., 2017), i.e., a federated system consisting of one central server and N distributed clients. clients' behavior are controlled by parameterized policies π^θ where $\pi^\theta(a|s)$ is a differentiable function of the parameter vector θ. For parameterized policies with t-indexed notation, we omit the parameter vector for notation simplicity and write π^t and π_n^t for the global π^{θ^t} and the n-th local $\pi^{\theta_n^t}$ policies, respectively. In the t-th training round, the central server broadcasts the current global policy π^t to K selected clients which will perform I iterations of local training. In each iteration, the n-th client interacts with its environment to collect E trajectories and utilize them to update its local policy to π_n^{t+1}. At the end of each round, the training results will be uploaded to the central server for aggregation to obtain the new global policy π^{t+1}. This step is done by parameter aggregation, i.e., $\theta^t = \sum_{n=1}^N q_n \theta_n^t$, where q_n is the weighting factor of the n-th client.

We model the local learning problem of each client as a finite-state infinite-horizon discounted Markov Decision Process (MDP). Accordingly, the FRL system consists of N finite MDPs $\{(\mathcal{S}, \mu, \mathcal{A}, P_n, \mathcal{R}, \gamma) : n \in \{1, ..., N\}\}$, where $\mathcal{S}$ denotes a finite set of states, μ represents the initial state distribution, $\mathcal{A}$ is a finite set of actions, and $\gamma \in (0, 1)$ is the discount factor. The transition function $P_n(s'|s, a) : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ represents the probability that the n-th MDP transits from state s to s' after taking action a (Sutton and Barto, 2018). The reward function $\mathcal{R}(s, a) : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$ gives the expected reward for taking action a in state s, and we assume rewards are bounded and non-negative. As a result, the n-th MDP $\mathcal{M}_n$ can be represented by a 6-tuple $(\mathcal{S}, \mu, \mathcal{A}, P_n, \mathcal{R}, \gamma)$ sharing the same state space, action space, initial state distribution, and reward function with other clients, but with possibly different transition probabilities. Throughout this work, we consider stochastic policies $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ which output the probability of taking an action a in a given

state s. Furthermore, we define the state-value function

$$V_n^\pi(s) = \mathbb{E}_{\pi,P_n}\left[\sum_{l=0}^{\infty}\gamma^l \mathcal{R}(s_{t+l}, a_{t+l})|s_t = s\right],$$

where the expectation is performed over actions sampled from policy π and states sampled from the transition probability P_n. It gives the expected return when the client starts from state s and follows policy π thereafter in the n-th MDP. For parameterized value functions with t-indexed notation, we omit the parameter vector w for notation simplicity and write V^t and V_n^t for the global V^{w^t} and the n-th local $V^{w_n^t}$ value functions, respectively. In each round, every client aims to train a local policy to maximize its expected discounted reward

$$\eta_n(\pi) = \mathbb{E}_{s_0\sim\mu, a_t\sim\pi, s_{t+1}\sim P_n}\left[\sum_{t=0}^{\infty}\gamma^t \mathcal{R}(s_t, a_t)\right], \tag{1}$$

or equivalently, $\eta_n(\pi) = \mathbb{E}_{s_0\sim\mu}\left[V_n^\pi(s_0)\right]$. The notation $\mathbb{E}_{s_0\sim\mu_n, a_t\sim\pi, s_{t+1}\sim P_n}$ indicates that the reward is averaged over all states and actions according to the initial state distribution, the transition probability, and the policy. Accordingly, the optimization problem for FRL can be formulated as

$$\max_\pi \eta(\pi) \quad \text{where} \quad \eta(\pi) = \sum_{n=1}^{N} q_n \eta_n(\pi). \tag{2}$$

Denote the averaged value function of policy π as

$$\bar{V}^\pi(s) = \sum_{n=1}^{N} q_n V_n^\pi(s),\ \forall s \in \mathcal{S},$$

then we can rewrite (2) as

$$\max_\pi \eta(\pi) \quad \text{where} \quad \eta(\pi) = \mathbb{E}_{s_0\sim\mu}\left[\bar{V}^\pi(s_0)\right]. \tag{3}$$

The above formulation covers both heterogeneous and homogeneous cases. In particular, the different MDPs, i.e., different transition probabilities, represent the heterogeneous environments experienced by clients. All MDPs will be identical for the homogeneous case (Fan et al., 2021). It is worth noting that the optimization problem in (3) is often referred to as the Weighted Value Problem (WVP) in the latent MDP literature. Finding the optimal solution of WVP is NP-hard (Steimle et al., 2021). In contrast, an error bound showing the distance between the obtained policy and the optimal policy is feasible as demonstrated in Section 3.

2.2 Approximate Policy Iteration

Given any MDP $\mathcal{M}_n$ defined in Section 2.1, it is well known (Sutton and Barto, 2018) that the value function V_n^π is the unique fixed point of the Bellman operator $T_n^\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$, i.e., $\forall s \in \mathcal{S}, V \in \mathbb{R}^{|\mathcal{S}|}$

$$T_n^\pi V(s) = \sum_a \pi(a|s)\left(\mathcal{R}(s,a) + \gamma\sum_{s'} P_n(s'|s,a)V(s')\right), \quad V_n^\pi(s) = T_n^\pi V_n^\pi(s),$$

where $|\mathcal{S}|$ denotes the cardinality of $\mathcal{S}$. Similarly, the optimal value function V_n^* is the unique fixed point of the Bellman operator $T_n : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$, i.e., $\forall s \in \mathcal{S}, V \in \mathbb{R}^{|\mathcal{S}|}$

$$T_n V(s) = \arg\max_a \left(\mathcal{R}(s,a) + \gamma \sum_{s'} P_n(s'|s,a) V(s') \right), \quad V_n^*(s) = T_n V_n^*(s).$$

Note that the subscript n of the Bellman operators denotes the index of transition probability with which the operator is applied. Both operators are monotonic and sup-norm contractive (Bertsekas and Tsitsiklis, 1996).

Now we describe the classic API, which is an iterative algorithm that generates a sequence of policies and the associated value functions. Let $\|\cdot\|$ denote the sup-norm, i.e. $\|V\| = \sup_{s\in\mathcal{S}} |V(s)|, \forall V \in \mathbb{R}^{|\mathcal{S}|}$, $\|\cdot\|_2$ denote the l_2-norm, i.e. $\|V\|_2 = \sqrt{\sum_{s\in\mathcal{S}} V(s)^2}, \forall V \in \mathbb{R}^{|\mathcal{S}|}$, and V^* denote the value function of the optimal policy π^*. Given the current policy π^t, each iteration consists of two phases, where δ and ϵ are some scalars:

Policy Evaluation. The value function V^{π^t} of the current policy is approximated by V^t satisfying

$$\left\| V^t - V^{\pi^t} \right\| \leq \delta, \ t = 0, 1, \cdots . \tag{4}$$

Policy Improvement. A greedy improvement is made to the policy with an approximation error

$$\left\| T^{\pi^{t+1}} V^t - T V^t \right\| \leq \epsilon, \ t = 0, 1, \cdots . \tag{5}$$

The following proposition gives the error bound of API.

Proposition 1 *The sequence $\left(\pi^t\right)_{t=0}^{\infty}$ generated by the API algorithm described by (4), (5) satisfies*

$$\limsup_{t\to\infty} \left\| V^{\pi^t} - V^* \right\| \leq \frac{\epsilon + 2\gamma\delta}{(1-\gamma)^2}, \tag{6}$$

The detailed proof of Proposition 1 can be found in Proposition 2.4.3 of Bertsekas (2022).

2.3 Imaginary MDP

We define the imaginary MDP as in QAvg (Jin et al., 2022). Specifically, it is a MDP represented by the 6-tuple $(\mathcal{S}, \mu, \mathcal{A}, \bar{P}, \mathcal{R}, \gamma)$ where

$$\bar{P}(s'|s,a) = \sum_{n=1}^{N} q_n P_n(s'|s,a), \forall s', s \in \mathcal{S}, a \in \mathcal{A},$$

denotes the average transition probability. Accordingly, we denote the Bellman operators in the imaginary MDP as T_I^π and T_I, where the subscript I indicates that the transition probability is $\bar{P}(s'|s,a), \forall s, s' \in \mathcal{S}, a \in \mathcal{A}$. Moreover, denoting π_I^* the optimal policy in the imaginary MDP $\mathcal{M}_I$, we have the value function of π and the optimal value function in the imaginary MDP

$$V_I^\pi(s) = T_I^\pi V_I^\pi(s), \quad V_I^{\pi_I^*}(s) = V_I^*(s) = T_I V_I^*(s),$$

respectively. The imaginary MDP is a handy tool to analyze the behavior of FAPI since it provides a unified view of all clients in the context of MDP where the theory of API is richly supplied with operator theory and the fixed point theorem.

3. Error Bound of FAPI

In this section, we establish the error bound of FAPI under the framework of API and the imaginary MDP. The analysis shows that the FAPI process can be considered as learning in the imaginary MDP, which facilitates the analysis. We focus our discussion on policy with function approximation, where the error term introduced by the aggregation step is nontrivial.

To establish FAPI's error bound, we first examine the approximation error of policy evaluation and improvement within the FRL framework. Then, we can derive the distance between the optimal value function and the value function produced by FAPI.

Throughout this work, we assume the policy iteration performed by each client is approximate. In particular, let V_n^t denote the evaluated value function of the n-th client in round t, we have

$$\left\|V_n^t - V_n^{\pi^t}\right\| \leq \delta_n, \quad \max_{V\in\mathbb{R}^{|\mathcal{S}|}} \left\|T_n^{\pi_n^{t+1}} V - T_n V\right\| \leq \epsilon_n, \quad t = 0, 1, \cdots. \tag{7}$$

In the following, we define a metric to quantify the level of heterogeneity of the FRL system. Specifically, we provide two metrics, which will be used for bounding the error of policy evaluation and policy improvement, respectively.

Definition 2 *(Level of heterogeneity). We define two parameters to measure the level of heterogeneity as*

$$\kappa_1 = \sum_{n=1}^{N} q_n \kappa_{n,I}, \quad \kappa_2 = \sum_{i,j} q_i q_j \kappa_{i,j},$$

where $P_n^\pi(s'|s) = \sum_a \pi(a|s) P_n(s'|s,a), \forall s, s' \in \mathcal{S}$, $\kappa_{i,j} = \max_{\pi,s} \sum_{s'} \left|P_i^\pi(s'|s) - P_j^\pi(s'|s)\right|$ *and* $\kappa_{n,I} = \max_{\pi,s} \sum_{s'} \left|P_n^\pi(s'|s) - \sum_{j=1}^N q_j P_j^\pi(s'|s)\right|$.

Here, κ_1 measures the average deviation of clients' MDPs from the imaginary MDP, and κ_2 measures the average distance between each pair of clients in terms of the transition probability. With homogeneous environments, both κ_1 and κ_2 will be equal to 0. As the transition probability of each MDP gets farther away from each others, κ_1 and κ_2 tend to be larger and indicate a heterogeneous network. It is trivial to show that $\kappa_1 \leq \kappa_2$.

3.1 Federated Policy Evaluation

The following lemmas describe the relation between the averaged value function $\bar{V}^\pi(s)$ of policy π and the value function V_I^π of policy π in the imaginary MDP.

Lemma 3 *For all states s and policies π, we have $\bar{V}^\pi(s) \geq V_I^\pi(s)$.*

Lemma 4 *For all policies π, we have $\left\|\bar{V}^\pi - V_I^\pi\right\| \leq \frac{\gamma R_{\max} \kappa_1}{(1-\gamma)^2}$.*

Readers are referred to Jin et al. (2022) for detailed proofs and discussions of Lemmas 3 and 4. Briefly, V_I^π serves as a lower bound of $\bar{V}^\pi$. Among all policies, the optimal policy π_I^* in

the imaginary MDP is of particular interest since its value function $V_I^{\pi_I^*} = V_I^*$ is the largest lower bound of the averaged value function $\bar{V}^\pi$.

There are two sources of error when model aggregation is utilized to obtain an approximation V^t for the (desired) averaged value function $\bar{V}^{\pi^t}$. On the one hand, there are the approximation errors from local policy evaluation (7). On the other hand, there is also the error between the target "aggregated function" and the function realized by the aggregated model parameters, due to the nonlinearity of models.

Given the approximation error in each client, we can only aggregate the approximations for the local value functions by $\bar{V}^t(s) = \sum_n^N q_n V_n^t(s), \forall s \in \mathcal{S}$ to approximate the averaged value function $\bar{V}^{\pi^t}$ with a bounded error

$$\left\|\bar{V}^t - \bar{V}^{\pi^t}\right\| \leq \sum_{n=1}^{N} q_n \delta_n = \bar{\delta}. \tag{8}$$

Furthermore, with the linear parameter aggregation, i.e., $w^t = \sum_{n=1}^N q_n w_n^t$, there will be an additional discrepancy $\bar{\varepsilon}_w$ between the function V^t parameterized by w^t and the aggregated local approximations $\bar{V}^t$, i.e.,

$$\left\|V^t - \bar{V}^t\right\| \leq \max_t \left\|V^t - \bar{V}^t\right\| = \bar{\varepsilon}_w. \tag{9}$$

Therefore, by the triangle inequality, we have

$$\left\|V^t - V_I^{\pi^t}\right\| \leq \left\|V^t - \bar{V}^t\right\| + \left\|\bar{V}^t - \bar{V}^{\pi^t}\right\| + \left\|\bar{V}^{\pi^t} - V_I^{\pi^t}\right\| \leq \bar{\varepsilon}_w + \bar{\delta} + \frac{\gamma R_{\max} \kappa_1}{(1-\gamma)^2} = \dot{\delta}. \tag{10}$$

Note that we postpone the discussion of the error $\bar{\varepsilon}_w$ induced by value function aggregation to the end of this section.

3.2 Federated Policy Improvement

Note that the approximation errors in (8) and (10) match that of the policy evaluation of API (4). This motivates us to further obtain an API-style approximation error for the policy improvement phase of FAPI. We consider two variants of FAPI: FAPI with federated policy evaluation (Algorithm 1) and FAPI without federated policy evaluation (Algorithm 2). Algorithm 1 is communication inefficient in practice since it introduces an extra round of communication for policy evaluation (lines 3 - 8 in Algorithm 1). The two algorithms lead to different approximation errors of the policy improvement phase, as will be shown by Lemmas 5 and 6, respectively.

Lemma 5 *With federated policy evaluation, the sequence $\left(\pi^t\right)_{t=0}^{\infty}$ generated by Algorithm 1 satisfies*

$$\left\|T_I^{\pi^{t+1}} V^t - T_I V^t\right\| \leq \frac{\bar{\varepsilon}_\theta \sqrt{|\mathcal{A}|} R_{\max}}{1-\gamma} + \frac{2\gamma R_{\max} \kappa_1}{1-\gamma} + \bar{\epsilon} = \epsilon',$$

where $\bar{\epsilon} = \sum_{n=1}^N q_n \epsilon_n$ and $\bar{\varepsilon}_\theta = \max_{t>0, s\in\mathcal{S}} \left\|\pi^{t+1}(\cdot|s) - \sum_{n=1}^N q_n \pi_n^{t+1}(\cdot|s)\right\|_2$.

Algorithm 1 FAPI with federated policy evaluation (full client participation)

Input: N, T.
for $t = 0, 1, \ldots, T$ **do**
  Synchronize the global policy π^t to every client.
  for $n = 0, 1, \ldots, N$ **do**
    Approximate the value function $V_n^{\pi^t}$ with V_n^t.
    Upload V_n^t to the central server.
  end for
  The central server aggregates V_n^t to obtain V^t: $w^{t+1} \leftarrow \sum_{n=1}^N q_n w_n^{t+1}$.
  Synchronize the global policy π^t and value function V^t to every client.
  for $n = 0, 1, \ldots, N$ **do**
    Local update of client policy: $\left\| T_n^{\pi_n^{t+1}} V^t - T_n V^t \right\| \leq \epsilon_n$.
    Upload π_n^{t+1} to the central server.
  end for
  The central server aggregates π_n^{t+1} to obtain π^{t+1}: $\theta^{t+1} \leftarrow \sum_{n=1}^N q_n \theta_n^{t+1}$.
end for

Algorithm 2 FAPI without federated policy evaluation (full client participation)

Input: N, T.
for $t = 0, 1, \ldots, T$ **do**
  Synchronize the global policy π^t to every client.
  for $n = 0, 1, \ldots, N$ **do**
    Approximate the value function $V_n^{\pi^t}$ with V_n^t.
    Local update of client policy: $\left\| T_n^{\pi_n^{t+1}} V_n^t - T_n V_n^t \right\| \leq \epsilon_n$.
    Upload π_n^{t+1} to the central server.
  end for
  The central server aggregates π_n^{t+1} to obtain π^{t+1}: $\theta^{t+1} \leftarrow \sum_{n=1}^N q_n \theta_n^{t+1}$.
end for

See Appendix A for the detailed proof. Similar to $\bar{\varepsilon}_w$, $\bar{\varepsilon}_\theta$ is equal to zero when the policy is linear with respect to its parameters, i.e., $\pi^{\theta^t}(a|s) = \sum_{n=1}^N q_n \pi^{\theta_n^t}(a|s), \forall n \in [N], s \in \mathcal{S}, a \in \mathcal{A}$ where $\theta^t = \sum_{i=1}^N q_i \theta_i^t$. Note that we postpone the discussion of the error ε_θ induced by policy aggregation to the end of this section.

Lemma 6 *Without federated policy evaluation, the sequence $\left(\pi^t\right)_{t=0}^\infty$ generated by Algorithm 2 satisfies*

$$\left\| T_I^{\pi^{t+1}} \bar{V}^t - T_I \bar{V}^t \right\| \leq \frac{\bar{\varepsilon}_\theta \sqrt{|\mathcal{A}|} R_{\max}}{1-\gamma} + \frac{2\gamma^2 R_{\max} \kappa_2}{(1-\gamma)^2} + \frac{\gamma R_{\max} \kappa_1}{1-\gamma} + 4\gamma \bar{\delta} + \bar{\epsilon} = \dot{\epsilon}. \tag{11}$$

See Appendix B for the detailed proof. As shown by Lemma 5 and Lemma 6, FAPI with federated policy evaluation provides a tighter bound. Intuitively, forcing clients to optimize their local policy from the same starting point (value function $\bar{V}^{\pi^t}$) helps them mitigate the negative impact of heterogeneity. While Algorithm 2 is what FRL applications typically

employ in practice, it is also more vulnerable to environment heterogeneity as its error bound is inferior to the one of Algorithm 1 roughly by a factor of $1-\gamma$.

By far, we have assumed full client participation, i.e., all clients participate in every round of training. However, partial client participation is more favorable in practice. Proposition 7 accounts for this scenario.

Proposition 7 *Let $\mathcal{C}$ denote the set of selected clients and $q'_m = \frac{q_m}{\sum_{m\in\mathcal{C}} q_m}$. With partial client participation and without federated policy evaluation, the sequence $\left(\pi^t\right)_{t=0}^{\infty}$ generated by Algorithm 2 satisfies*

$$\left\|T_I^{\pi^{t+1}}\bar{V}^t - T_I\bar{V}^t\right\| \leq \frac{\tilde{\varepsilon}_\theta\sqrt{|\mathcal{A}|}R_{\max}}{1-\gamma} + \sum_{m\in\mathcal{C}}\sum_{n=1}^{N} q'_m q_n \frac{\left(\gamma+\gamma^2\right)R_{\max}\kappa_{m,n}}{(1-\gamma)^2} + \sum_{m\in\mathcal{C}} q'_m \frac{\gamma R_{\max}\kappa_{m,I}}{1-\gamma} + 2\gamma\sum_{m\in\mathcal{C}} q'_m\delta_m + 2\gamma\bar{\delta} + \sum_{m\in\mathcal{C}} q'_m\epsilon_m = \hat{\epsilon}, \quad (12)$$

where $\tilde{\varepsilon}_\theta = \max_{t>0,s\in\mathcal{S}} \left\|\pi^t(\cdot|s) - \sum_{m\in\mathcal{C}} q'_m\pi_m^t(\cdot|s)\right\|_2$.

See Appendix C for the detailed proof. To better understand how to minimize the right-hand side of (12), one can consider the optimization problem, $\min_x f(x) = \frac{1}{N}\sum_{n=1}^{N}|x-a_j|$, which corresponds to the case $q_n = \frac{1}{N}, n = 1,\cdots,N$ and $|\mathcal{C}| = 1$ in (12). It can be easily shown that $f(x)$ is minimal when x is the median of $\{a_1,\cdots,a_N\}$.

Remark 8 *Proposition 7 reveals a remarkable fact that a proper client selection method can effectively reduce the error bound of the policy improvement phase. In particular, to have the right-hand side of (12) smaller than the right-hand side of (11), the selected clients shall have an average $\kappa_{m,n}$ that is smaller than $\frac{2\gamma}{1+\gamma}\kappa_2$. This encourages FAPI to select clients that are closer to the imaginary MDP, which is a reasonable approximation to the median of all transition probabilities.*

For completeness, we provide the error bound of the policy improvement phase with partial client participation and federated policy evaluation by the following proposition.

Proposition 9 *Let $\mathcal{C}$ denote the set of selected clients and $q'_m = \frac{q_m}{\sum_{m\in\mathcal{C}} q_m}$. With partial client participation and federated policy evaluation, the sequence $\left(\pi^t\right)_{t=0}^{\infty}$ generated by Algorithm 1 satisfies*

$$\left\|T_I^{\pi^{t+1}}V^t - T_IV^t\right\| \leq \frac{\tilde{\varepsilon}_\theta\sqrt{|\mathcal{A}|}R_{\max}}{1-\gamma} + \sum_{m\in\mathcal{C}}\sum_{n=1}^{N} q'_m q_n \frac{\gamma R_{\max}\kappa_{m,n}}{1-\gamma} + \sum_{m\in\mathcal{C}} q'_m \frac{\gamma R_{\max}\kappa_{m,I}}{1-\gamma} + \sum_{m\in\mathcal{C}} q'_m\epsilon_m = \acute{\epsilon}. \quad (13)$$

See Appendix E for the detailed proof.

3.3 Bounding the distance between value functions

The following theorem provides the error bound of FAPI in terms of the distance between value functions.

Theorem 10 *Let $\pi^* = \arg\max_\pi \eta(\pi)$. The sequence $\left(\pi^t\right)_{t=0}^{\infty}$ generated by FAPI satisfies*

$$\limsup_{t\to\infty} \left|\bar{V}^{\pi^t}(s) - \bar{V}_s^{\max}\right| \leq \frac{\tilde{\epsilon} + 2\gamma\tilde{\delta}}{(1-\gamma)^2} + 2\frac{\gamma R_{\max}\kappa_1}{(1-\gamma)^2}, \tag{14}$$

where $\bar{V}_s^{\max} = \max\left\{\bar{V}^{\pi^*}(s), \bar{V}_I^{\pi_I^*}(s)\right\}, \forall s \in \mathcal{S}$, $\tilde{\delta}$ may be $\bar{\delta}$ or $\dot{\delta}$, and $\tilde{\epsilon}$ may be one of $\dot{\epsilon}$, ϵ', $\hat{\epsilon}$ or $\acute{\epsilon}$. More specifically, the error bound for Algorithm 2 with partial client participation is

$$\begin{aligned}\limsup_{t\to\infty} \left|\bar{V}^{\pi^t}(s) - \bar{V}_s^{\max}\right| \leq\ & C_1\kappa_1 + C_2\sum_{m\in\mathcal{C}} q'_m\kappa_{m,I} + C_3\sum_{m\in\mathcal{C}}\sum_{n=1}^{N} q'_m q_n \kappa_{m,n} \\ & + \tilde{\mathcal{O}}\left(\tilde{\varepsilon}_\theta\sqrt{|\mathcal{A}|}R_{\max} + \bar{\delta} + \sum_{m\in\mathcal{C}} q'_m\delta_m + \sum_{m\in\mathcal{C}} q'_m\epsilon_m\right),\end{aligned} \tag{15}$$

where $\tilde{\mathcal{O}}$ omits some constants related to γ, $C_1 = \frac{2\gamma\left(\gamma^2-\gamma+1\right)}{(1-\gamma)^4}R_{\max}$, $C_2 = \frac{\gamma}{(1-\gamma)^3}R_{\max}$, and $C_3 = \frac{\gamma+\gamma^2}{(1-\gamma)^4}R_{\max}$.

See Appendix D for the detailed proof.

Remark 11 *The bound given by Theorem 10 is similar to that of the centralized API as in Proposition 1 and inversely proportional to the heterogeneity level κ_1 and κ_2, which explicitly unveils the impact of the data heterogeneity. The second term in (14) stems from the difference between $\left\|\bar{V}^{\pi^t} - \bar{V}^{\pi^*}\right\|$ and $\left\|V_I^{\pi^t} - V_I^*\right\|$. Although the imaginary MDP enables us to analyze the theoretical properties of FAPI and shows the error bound in terms of V_I^*, $\bar{V}^{\pi^*}$ is the actual target (refer to (3)) that FAPI wants to achieve. Fortunately, this bound is still useful, since (14) is dominated by its first term. To this end, the optimal policy in the imaginary MDP is a good estimation of the optimal policy for (3) unless there is a bound sharper than $\tilde{\delta}$. Again, the client selection scheme is the key to reducing error.*

3.4 Impact of Aggregation Error

We have defined three terms to quantify the impact of policy aggregation and value function aggregation, i.e., $\tilde{\varepsilon}_\theta$, $\bar{\varepsilon}_\theta$, and $\bar{\varepsilon}_w$. To further investigate how these errors affect the convergence of FAPI, there are two possible approaches: 1) Performing a general analysis with a few standard assumptions in the convex optimization literature; and 2) Carrying out the analysis with a specific neural network parameterization to gain insight into the network configuration that can affect the approximation error. In light of the recent breakthrough in overparameterized neural networks (Arora et al., 2019; Cai et al., 2019; Wang et al., 2020; Liu et al., 2019), we employ the second strategy. To that end, we analyze the impact of

aggregation error with the following two-layer ReLU-activated Neural Networks (NNs) to parameterize the state value function and policy, respectively, as

$$u_w(s) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \cdot \sigma(w_i^T(s)), \tag{16}$$

$$f_\theta(s,a) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \cdot \sigma(\theta_i^T(s,a)), \tag{17}$$

where m is the width of the network, $\theta = (\theta_1^T, \ldots, \theta_m^T)^T \in \mathcal{R}^{m(d_s+d_a)}$, $w = (w_1^T, \ldots, w_m^T)^T \in \mathcal{R}^{md_s}$ denotes the input weights, and $b_i \in \{-1, 1\}, \forall i \in [m]$ is the output weight. Without loss of generality, we assume that $a \in \mathbb{R}^{d_a}, s \in \mathbb{R}^{d_s}$, $\|(s,a)\|_2 \leq 1$, and denote $d = d_s + d_a$. Given arbitrary $\hat{R}_\vartheta < \infty$, we consider the following parameter initialization, where ϑ is a placeholder for θ and w:

$$\begin{aligned} &\mathbb{E}_{\text{init}}\left[\vartheta_{i,j}^0\right] = 0, \mathbb{E}_{\text{init}}\left[\left(\vartheta_{i,j}^0\right)^2\right] = \frac{1}{d \cdot m}, \forall i \in [m], j \in [d], \\ &b_i \sim \text{Unif}(\{-1,1\}), \mathbb{E}_{\text{init}}\left[\left\|\vartheta_i^0\right\|_2^{-2}\right] < \infty, \left\|\vartheta_i^0\right\|_2 \leq \hat{R}_\vartheta, \forall i \in [m], \end{aligned} \tag{18}$$

where $\mathbb{E}_{\text{init}}[\ldots]$ denotes the expectation over parameter initialization. In other words, $0 < \left\|\vartheta_i^0\right\|_2 \leq \hat{R}_\vartheta, \forall i \in [m]$. The output weights $b_i, \forall i \in [m]$ are fixed. The input weights $\vartheta_i, \forall i \in [m]$ are projected into a ball centered at the initial parameter, i.e., $\theta \in \mathcal{B}_{\mathcal{R}_\theta}^0 = \left\{\theta : \left\|\theta - \theta^0\right\|_2 \leq \mathcal{R}_\theta\right\}, w \in \mathcal{B}_{\mathcal{R}_w}^0 = \left\{w : \left\|w - w^0\right\|_2 \leq \mathcal{R}_w\right\}$. Then, the state value function and Softmax policy are parameterized by

$$V^t(s) = u_{w^t}(s), \pi^t(a|s) = \frac{\exp(f_{\theta^t}(s,a))}{\sum_{a' \in \mathcal{A}} \exp(f_{\theta^t}(s,a'))}, \forall a \in \mathcal{A}, s \in \mathcal{S}. \tag{19}$$

The following lemma quantifies the aggregation error under the above setting.

Lemma 12 *By utilizing the initialization scheme in (18), the policy parameterization in (19), and the neural network parameterization in (16) and (17), we have*

$$\mathbb{E}_{init}\left[\bar{\varepsilon}_w\right] = \mathcal{O}\left(R_w^{6/5} m^{-1/10} \hat{R}_w^{2/5}\right), \tag{20}$$

$$\mathbb{E}_{init}\left[\bar{\varepsilon}_\theta\right] = \mathcal{O}\left(R_\theta^{1/2}\right), \tag{21}$$

$$\mathbb{E}_{init}\left[\tilde{\varepsilon}_\theta\right] = \mathcal{O}\left(R_\theta^{1/2}\right). \tag{22}$$

See Appendix G for the detailed proof.

Remark 13 *Lemma 12 indicates that the aggregation error is determined by the neural network parameterization (m) and optimization method (R_θ and R_w). While a small projection radius, e.g., setting R_θ and R_w to zero, is beneficial for reducing aggregation errors ($\bar{\varepsilon}_\theta$, $\tilde{\varepsilon}_\theta$, and $\bar{\varepsilon}_w$), the algorithm's performance will be limited by the representation power of the corresponding function class $\mathcal{F}_{\theta^0,R_\theta,m} = \left\{\frac{1}{\sqrt{m}} \sum_{i=1}^m b_i \cdot \sigma(\theta_i^T(s,a)) : \theta \in \mathcal{B}_{\mathcal{R}_\theta}^0\right\}$. In other words, there is a trade-off between approximation error (a rich function class) and aggregation error (a small parameter space). A powerful function approximator can fit training data well, but suffers from high aggregation error. In contrast, a restricted function approximator may not be able to solve complex problems, but it leads to smaller aggregation error.*

Remark 14 *The aggregation errors ($\bar{\varepsilon}_\theta$, $\tilde{\varepsilon}_\theta$, and $\bar{\varepsilon}_w$) stem from the nonlinearity of function approximators and will be zero when linear models (approximators) are utilized. In particular, the error for value function aggregation will be zero when value functions are linear, such as linear regression and tabular implementation, and the error for policy aggregation will also be zero when using linear policy parameterization, such as tabular policies. Lemma 12 indicates that, in the infinite-width limit, the aggregation error for the two-layer ReLU-activated NNs is very small. This observation is consistent with the well-known fact that, in the infinite-width limit, an NN is approximately linear with respect to its parameters during gradient descent (Jacot et al., 2018; Cao and Gu, 2019; Lee et al., 2019).*

3.5 Connection to Centralized Learning

When the environments are homogeneous, the policy is linear w.r.t. parameters, and the value function is linear w.r.t. parameters, the learning process of FAPI will be equivalent to learning from N copies of the same environment (that is identical to the imaginary MDP) with federated policy evaluation. Under such circumstances, Theorem 10 degenerates to that for centralized learning (6) (Bertsekas, 2022) with the same error bound, i.e., $\frac{\bar{\epsilon}+2\gamma\bar{\delta}}{(1-\gamma)^2}$, as shown in the following proposition.

Proposition 15 *With homogeneous environments and federated policy evaluation, the error bound for partial/full client participation is*

$$\limsup_{t\to\infty} \left\| \bar{V}^{\pi^t} - \bar{V}^{\pi^*} \right\| \leq \frac{\hat{\varepsilon}_\theta \sqrt{|\mathcal{A}|} R_{\max}}{(1-\gamma)^3} + \frac{2\gamma\bar{\varepsilon}_w}{(1-\gamma)^2} + \frac{\bar{\epsilon}+2\gamma\bar{\delta}}{(1-\gamma)^2},$$

where $\hat{\varepsilon}_\theta$ is equal to $\bar{\varepsilon}_\theta$ and $\tilde{\varepsilon}_\theta$ for full participation and partial participation, respectively.

See Appendix F for the detailed proof. As discussed in Section 3.4, $\hat{\varepsilon}_\theta = 0$ when the policy is linear w.r.t. parameters, and $\bar{\varepsilon}_w = 0$ when the value function is linear w.r.t. parameters.

4. Federated Policy Optimization with Heterogeneity-aware Client Selection

In this section, we propose a federated policy optimization algorithm based on the discussions in Section 3. The pseudocode for the proposed Federated Policy Optimization with Heterogeneity-aware Client Selection (FedPOHCS) is illustrated in Algorithm 3.

4.1 Client Selection Metric

As characterized by Remark 8, a proper client selection method should be able to capture the heterogeneity level κ_1/κ_2 of the selected clients. In general, the smaller the difference between the transition probability of selected clients and that of the imaginary MDP, the better bound we may obtain. As there are many methods to approximate and represent the transition probability, we consider it as an implementation consideration and leave it to the applications. The use of transition probability makes the proposed client selection scheme a model-based framework (Moerland et al., 2023).

Next, we make several approximations to the theoretically justified client selection metric, i.e., the level of heterogeneity of the n-th client $\kappa_{n,I}$ defined in Definition 2. It is hard to compute $\kappa_{n,I}$ by finding the maximum value over all states since the number of samples is finite in practice and this metric may suffer from high variance. Therefore, we use the current global policy to compute the metric and weight each transition with the stationary (or steady-state) distribution $d_{\pi,P_n}(s)$ of the entry state s (Bojun, 2020). Moreover, in the proofs of the lemmas and propositions, we relax the inequalities by replacing all value functions with their upper bound $\frac{R_{\max}}{1-\gamma}$. To further improve the approximation, we scale each transition with the value (advantage or Q-value) of each state-action pair (s', a). Then, for each tuple of (s, s', a) in the n-th client, we have

$$\hat{\kappa}_{n,I}(s, s', a) = \left| d_{\pi,P_n}(s) P_n^{\pi}(s'|s) A_{\pi,P_n}(s', a) - \sum_{j=1}^{N} q_j d_{\pi,P_j}(s) P_j^{\pi}(s'|s) A_{\pi,P_j}(s', a) \right|.$$

However, it is too expensive to compute all $|\mathcal{S}| \times |\mathcal{S}| \times |\mathcal{A}|$ elements for each client. Since the stationary distribution is a fixed-point distribution, i.e., $\sum_s d_{\pi,P_n}(s) P_n^{\pi}(s'|s) = d_{\pi,P_n}(s')$, we can sum over the first dimension to simplify the metric and reduce the amount of computation. In other words, we want to compute

$$\begin{aligned}
\hat{\kappa}_{n,I}(s', a) &= \left| \sum_s d_{\pi,P_n}(s) P_n^{\pi}(s'|s) A_{\pi,P_n}(s', a) - \sum_{j=1}^{N} q_j \sum_s d_{\pi,P_j}(s) P_j^{\pi}(s'|s) A_{\pi,P_j}(s', a) \right| \\
&= \left| d_{\pi,P_n}(s') A_{\pi,P_n}(s', a) - \sum_{j=1}^{N} q_j d_{\pi,P_j}(s') A_{\pi,P_j}(s', a) \right|.
\end{aligned}$$

For compact notation, we define $\mathbf{D}_{\pi,P_n}$ and $\mathbf{D}_{\pi,\mu,P_n,\gamma}$ as two $|\mathcal{S}| \times |\mathcal{S}|$ diagonal matrices whose i-th diagonal entries are the stationary distribution and discounted visitation frequencies of state s_i, respectively. We denote $\mathbf{\Pi}_{\pi}$ as a $|\mathcal{S}| \times |\mathcal{A}|$ matrix whose (i, j)-th entry is $\pi(a_j|s_i)$, and $\mathbf{A}_{\pi,P_n}$ as a $|\mathcal{S}| \times |\mathcal{A}|$ matrix whose (i, j)-th entry is the advantage of action a_j on state s_i, i.e., a matrix representation of the advantage function (Kakade and Langford, 2002). Then, we can rewrite the approximated level of heterogeneity $\sum_{s',a} \hat{\kappa}_{n,I}(s', a)^2$ as $\hat{\kappa}_{n,I} = \left\| \sum_{k=1}^{N} q_k \mathbf{D}_{\pi,P_k} \mathbf{A}_{\pi,P_k} - \mathbf{D}_{\pi,P_n} \mathbf{A}_{\pi,P_n} \right\|_F$, where we use the Frobenius norm to make the metric more sensitive to entries with large difference.

However, the metric $\hat{\kappa}_{n,I}$ has a drawback: the algorithm will continue to select clients that are closer to the imaginary MDP, even if those clients are sufficiently trained. As a solution, we propose to consider the learning potential of local policies. In fact, the learning step size of many policy optimization algorithms depends on the magnitude of value functions. For example, the advantage function (or Q-value) can affect the magnitude of the policy gradient, and the Q-value can affect the updates in Temporal Difference (TD) methods, including Q-learning and SARSA. This observation motivates us to use the magnitude of the advantage function, i.e., $\|\mathbf{D}_{\pi,P_n} \mathbf{A}_{\pi,P_n}\|_F$, to measure how much the n-th client can learn starting from the current global policy. Finally, we define the selection metric of FedPOHCS

as

$$\Delta_n = \|\mathbf{D}_{\pi,P_n}\mathbf{A}_{\pi,P_n}\|_F - \left\|\sum_{k=1}^{N} q_k \mathbf{D}_{\pi,P_k}\mathbf{A}_{\pi,P_k} - \mathbf{D}_{\pi,P_n}\mathbf{A}_{\pi,P_n}\right\|_F, \ \forall n = 1, \cdots, N. \tag{23}$$

As shown in the first phase (lines 3 - 9) of Algorithm 3, the server samples a candidate set $\mathcal{C}$ and computes Δ_n for clients in this set. Then, the server selects K clients with the largest Δ_n from the candidate set and starts the second phase (lines 10 - 15) of Algorithm 3.

As a client selection metric, Δ_n is better than the original heterogeneity measurement $\kappa_{n,I}$, because it helps skip clients that can not sufficiently contribute to the global objective (3). This helps allocate resources to clients whose information has not been fully learned instead of investing all resources into those who are just closer to the imaginary MDP throughout the learning process. Similar measurements of heterogeneity and learning potential can be found in FedKL (Xie and Song, 2023), where states are weighted by the discounted visitation frequency instead of the stationary distribution.

4.2 Implementation

To compute Δ_n, we adopt the tabular maximum likelihood model (Moerland et al., 2023; Strehl et al., 2009; Ornik and Topcu, 2021; Strehl and Littman, 2008) in the implementation of FedPOHCS. Given a set of trajectories, let $C_n(s,a)$ denote the number of times action a was taken under state s, $C_n(s,a,s')$ represent the number of times the MDP transited from state s to $s\prime$ after taking action a, and r_n denote the corresponding sequence of reward received. Generally speaking, the more trajectories we have, the more accurate the modeling will be. For Mountain Cars and Hoppers, the number of trajectories is 200. For HongKongOSMs, the number of trajectories is 1000. We can estimate the transition probability P_n of the n-th MDP and the reward function $\mathcal{R}$ by

$$\hat{P}_n = \frac{C_n(s,a,s')}{C_n(s,a)}, \quad \hat{\mathcal{R}}_n = \frac{1}{C_n(s,a)} \sum_{i=1}^{C_n(s,a)} r_n[i].$$

We estimate the state visitation frequency matrix $\hat{\mathbf{D}}_{\pi,\mu,P_n}$ in place of $\mathbf{D}_{\pi,P_n}$ as follows (Ziebart et al., 2008):

$$D_{n,s',0} = \mu(s'), \quad D_{n,s',t+1} = \sum_{s,a} D_{n,s,t}\pi(a|s)\hat{P}_n(s'|s,a), \quad D_{n,s'} = \sum_t D_{n,s',t},$$

where the time horizon t is a hyperparameter. Then, we diagonalize the vector $D_{n,s'}$ to obtain the state visitation frequency matrix $\hat{\mathbf{D}}_{\pi,\mu,P_n}$. Last, we estimate the advantage function by Generalized Advantage Estimation (GAE) (Schulman et al., 2016).

In contrast to client selection schemes that update the selection metrics together with the models at the end of each round, we employ an extra round for metric computation (lines 3 - 9) in Algorithm 3. A similar selection scheme was utilized by a biased client selection method called Power-of-Choice (Jee Cho et al., 2022).

The choice of the local learner is optional as our discussion is general and does not rely on any particular implementation of policy evaluation and policy improvement. In particular, we assume all clients adopt the Proximal Policy Optimization (PPO) algorithm proposed by Schulman et al. (2017), which is a PG method motivated by TRPO.

Algorithm 3 FedPOHCS

Input: The initial estimation of the transition probabilities $\hat{P}_k$ and the reward functions $\hat{\mathcal{R}}_k, d, T, K, \theta^0$.
for $t = 0, 1, \ldots, T$ **do**
  Sample the candidate client set $\mathcal{C}$ of d ($K \leq d \leq N$) clients without replacement.
  Synchronize the global policy π^t to every selected client.
  for $k \in \mathcal{C}$ **do**
    Compute the advantage function $\mathbf{A}_{\pi,P_k}$ and the state visitation matrix $\hat{\mathbf{D}}_{\pi^t,\mu,P_k}$.
    Upload $\mathbf{A}_{\pi,P_k}, \hat{\mathbf{D}}_{\pi^t,\mu,P_k}$ to the central server.
  end for
  Compute $\Delta_k, \forall k = 1, ..., d$. Select K clients based on Δ_k to replace $\mathcal{C}$.
  for $k \in \mathcal{C}$ **do**
    Update client policy π_k^{t+1}, transition probability $\hat{P}_k$, and reward function $\hat{\mathcal{R}}_k$.
    Upload π_k^{t+1} to the central server.
  end for
  The central server aggregates π_k^{t+1} to obtain π^{t+1}: $\theta^{t+1} \leftarrow \sum_{k=1}^{N} q_k \theta_k^{t+1}$.
end for

4.3 Limitations of the Proposed Implementation

Algorithm 3 utilizes a two-phase communication scheme, which is not communication-efficient unless the convergence improvement obtained by client selection suppresses this cost. However, we note that such a communication cost can be removed by using outdated information to compute the selection metrics at the cost of accuracy, as in Jee Cho et al. (2022). In particular, we can remove the first phase, and order selected clients to upload their local information (e.g., $\mathbf{A}_{\pi,P_n}$ and $\mathbf{D}_{\pi,\mu,P_n}$ matrices) when uploading their models at the end of each communication round. This communication-efficient variant saves a lot of communication overhead and has been shown to be effective (Jee Cho et al., 2022), i.e., with slightly worse performance. The performance of the one-phase scheme will be shown later in the experiment results.

Another limitation of Algorithm 3 is that it requires the clients to upload their state visitation frequencies $\hat{\mathbf{D}}_{\pi,\mu,P_n}$ and advantage functions $\mathbf{A}_{\pi,P_n}$ which may contain private information. This problem can be addressed by privacy protection methods, e.g., Homomorphic Encryption (HE) (Jiang et al., 2018).

5. Experiments

In this section, we introduce two federated environments for empirical evaluation and evaluate the proposed client selection method from different perspectives. All experimental results are reproducible and can be accessed from: `https://github.com/ShiehShieh/FedPOHCS`.

5.1 Environments

We evaluate the effectiveness of the proposed client selection algorithm on a federated version of the mountain car continuous control problem (Moore, 1990) and the MuJoCo Hopper

problem (Coulom, 2002). Furthermore, we utilize the Flow simulator Vinitsky et al. (2018) and OpenStreetMap (OSM) (OpenStreetMap contributors, 2017; Vargas-Munoz et al., 2021) data set to create a series of traffic networks for autonomous vehicle training. We construct the federated environments as follows:

Mountain Cars consists of 60 equally weighted MountainCarContinuous-v0 environments developed by OpenAI Gym (Brockman et al., 2016). In each environment, a car aims to reach a hill. The episode terminates when the car reaches this hill or runs out of time. The environment consists of a 2-dimensional continuous state space and a 1-dimensional continuous action space. To introduce heterogeneity into the system, we assume that the engine of each car is different and the n-th car shifts the input action by θ_n on all states. To introduce a medium-level heterogeneity, the constant shift θ_n is uniformly sampled from $[-1.5, 1.5]$ and assigned to each environment at initialization. In fact, to make the experiments traceable, we set the constant shift to $\theta_n = -1.5 + \frac{n}{20}, n = 1, \ldots, 60$. The intervals for low-level and high-level heterogeneity are $[-1.0, 1.0]$ and $[-2.0, 2.0]$, respectively.

Hoppers consists of 60 equally weighted Hopper-v3 environments developed by OpenAI Gym. The environment has an 11-dimensional continuous state space and a 3-dimensional continuous action space. We introduce the heterogeneity into this system by following Jin et al. (2022), i.e., the leg size is uniformly sampled from $[0.01, 0.07]$, $[0.01, 0.10]$, and $[0.01, 0.15]$ for low-level, medium-level, and high-level heterogeneity, respectively.

HongKongOSMs consists of 10 equally weighted traffic networks, each based on one of the OSM data sets as shown in Figure 1. Each traffic network contains one RL-controlled and 10 IDM-controlled (Intelligent Driver Model) vehicles. The 18-dimensional observation includes headway, speed, and positional information of visible neighborhoods of the RL-controlled vehicle, and only observations for vehicles running on the same and adjacent lanes are visible to the RL-controlled vehicle. The 2-dimensional action includes acceleration and lane-changing decisions. We adopted the Eclipse Simulation of Urban MObility (SUMO) simulator to conduct this experiment.

5.2 Experiment Settings

We use neural networks to represent policies as in Schulman et al. (2015); Vinitsky et al. (2018). Specifically, we use Multilayer Perceptrons (MLPs) with tanh non-linearity and hidden layers (64, 64). We use the SGD optimizer with a momentum of 0.9 and learning-rate decay of 0.98, 0.9, and 0.98 per round for Mountain Cars, Hoppers, and HongKongOSMs, respectively. Hyperparameters are carefully tuned so that they are near-optimal for each algorithm. See Appendix H for more experimental details.

We compare FedPOHCS with biased and unbiased client selection methods, including FedAvg (random selection), Power-of-Choice (Jee Cho et al., 2022), and GradientNorm (Marnissi et al., 2024; Chen et al., 2022). FedAvg randomly selects K clients. Power-of-Choice utilizes the aforementioned two-phase scheme and selects candidate clients with the highest loss (or lowest advantages/values in case of RL problem). GradientNorm selects candidate clients with the largest gradient norm. Besides the client selection scheme, all algorithms follow the same procedure described at the beginning of Section 2.1. In particular, clients perform local training with the algorithm proposed by Schulman et al. (2017), with an adaptive KL penalty term. At the end of every round, the server broadcasts the global

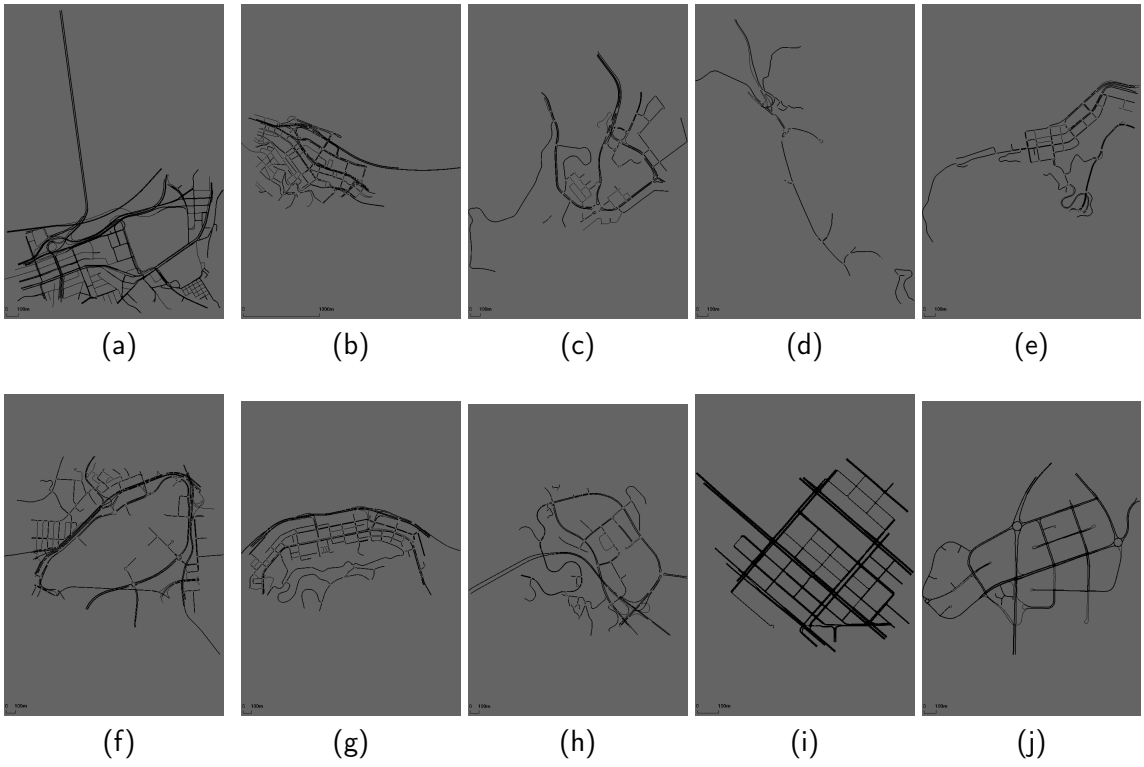


Figure 1: OSM data sets of ten areas in Hong Kong. (a) Causeway; (b) Central; (c) Chai Wan; (d) Clear Water Bay; (e) Kennedy Town; (f) Kai Tak; (g) North Point; (h) Po Lam; (i) Sham Shui Po; (j) Tseung Kwan O.

policy to all clients, orders them to interact with their MDPs for several episodes (10 for Mountain Cars, 100 for Hoppers, and 20 for HongKongOSMs) and report the mean returns to evaluate the performance. Each experiment is averaged across three independent runs with different random seeds and parameter initializations, both of which are shared among all algorithms for a fair comparison. Confidence intervals are also reported.

5.3 Performance and Stability

Although the original mountain car problem is simple and most modern RL algorithms can easily obtain a score over 90.0, the federated setting imposes great difficulties in solving it. Figure 2 shows the performance comparison between FedPOHCS and several baselines on Mountain Cars with medium-level heterogeneity. It can be observed that FedPOHCS has a faster convergence speed and a more stable learning process. To compute the selection metric, we discretize the state and action by rounding them off to the nearest 0.1, resulting in about 8000 states and 100 actions that are frequently visited. In each round, the first phase of FedPOHCS takes 1-10 seconds, and the local training takes 20-30 seconds. Although the running time of the first phase may vary depending on the implementation, FedPOHCS outperforms all baselines in terms of the number of rounds and wall-clock time in our setting.

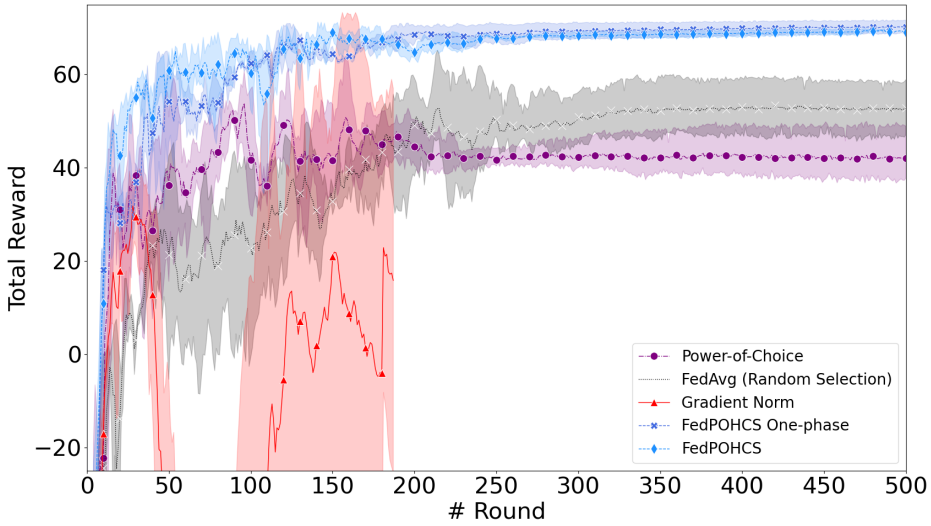


Figure 2: Comparison of FedAvg, Power-of-Choice, FedPOHCS, and GradientNorm on Mountain Cars with a medium level of heterogeneity. For FedAvg, the learning rate is 0.005, and the KL target is 0.003. For Power-of-Choice, GradientNorm, and FedPOHCS, the learning rate is 1e-3, and the KL target is 0.003.

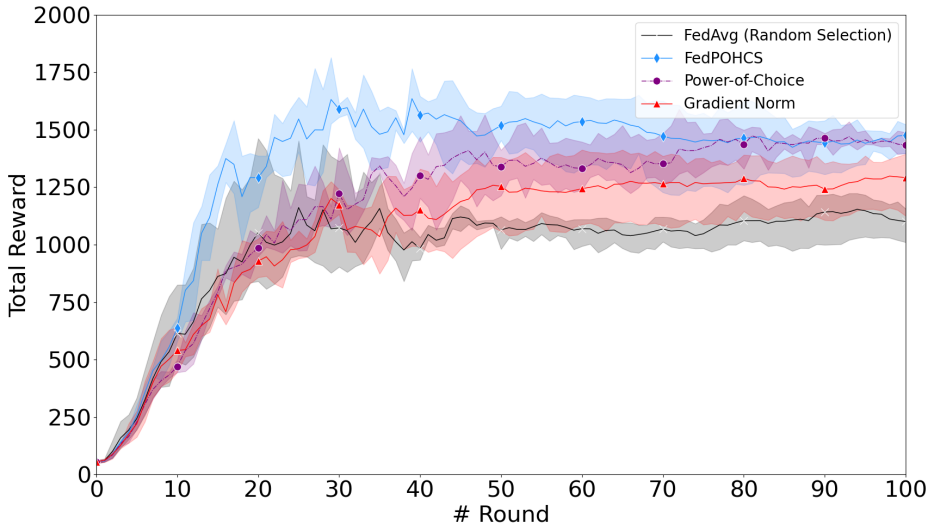


Figure 3: Comparison of FedAvg, Power-of-Choice, FedPOHCS, and GradientNorm on Hoppers with a medium level of heterogeneity. For all algorithms, the learning rate is 0.03, the learning rate decay is 0.9, and the KL target is 0.003.

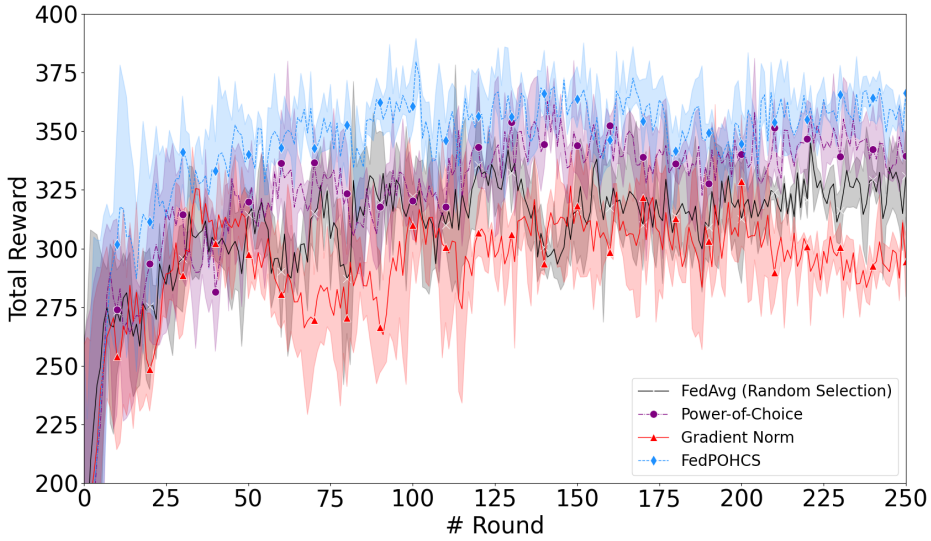


Figure 4: Comparison of FedAvg, Power-of-Choice, FedPOHCS, and GradientNorm on HongKongOSMs. For all algorithms, the learning rate is 0.0001, the learning rate decay is 0.98, and the KL target is 0.0001.

In Figure 2, we have also included the communication-efficient variant, i.e., the one-phase scheme, for comparison purposes. It can be observed that the one-phase scheme may slow down and destabilize learning in the early stage of training, but the final performance is comparable to the two-phase scheme.

We can draw a similar conclusion on Hoppers with medium-level heterogeneity. As shown in Figure 3, FedPOHCS can obtain an accumulated reward of 1450 within 20 rounds of training, while it takes about 80 rounds for Power-of-Choice to reach 1450, demonstrating the advantage of the proposed FedPOHCS algorithm in speeding up convergence.

HongKongOSMs is much more difficult to solve as shown by the high-variance curves in Figure 4. Since different OSM data sets have different numbers of lanes, target velocity, and maximum acceleration/deceleration, their state spaces and transition probabilities may be highly distinct from each other. Compared with other approaches, FedPOHCS can obtain higher rewards and converge to the highest point.

5.4 Effectiveness of Metrics

In Figure 5, we show how different algorithms select clients (the histogram) and the reward obtained by the final policy from each client (the scatter points) on Mountain Cars with medium-level heterogeneity. Note that with the constant shift $\theta_n = -1.5 + \frac{n}{20}$, clients with small IDs are very different from those with large IDs. It can be observed that, compared with random selection and Power-of-Choice, FedPOHCS refuses to allocate resources to clients with IDs in $[40, 60]$, while Power-of-Choice spends a significant amount of resources on them. The final policy generated by FedPOHCS performs well on almost all clients and gets fairly high scores on clients with IDs in $[1, 10]$ without hurting other clients. This

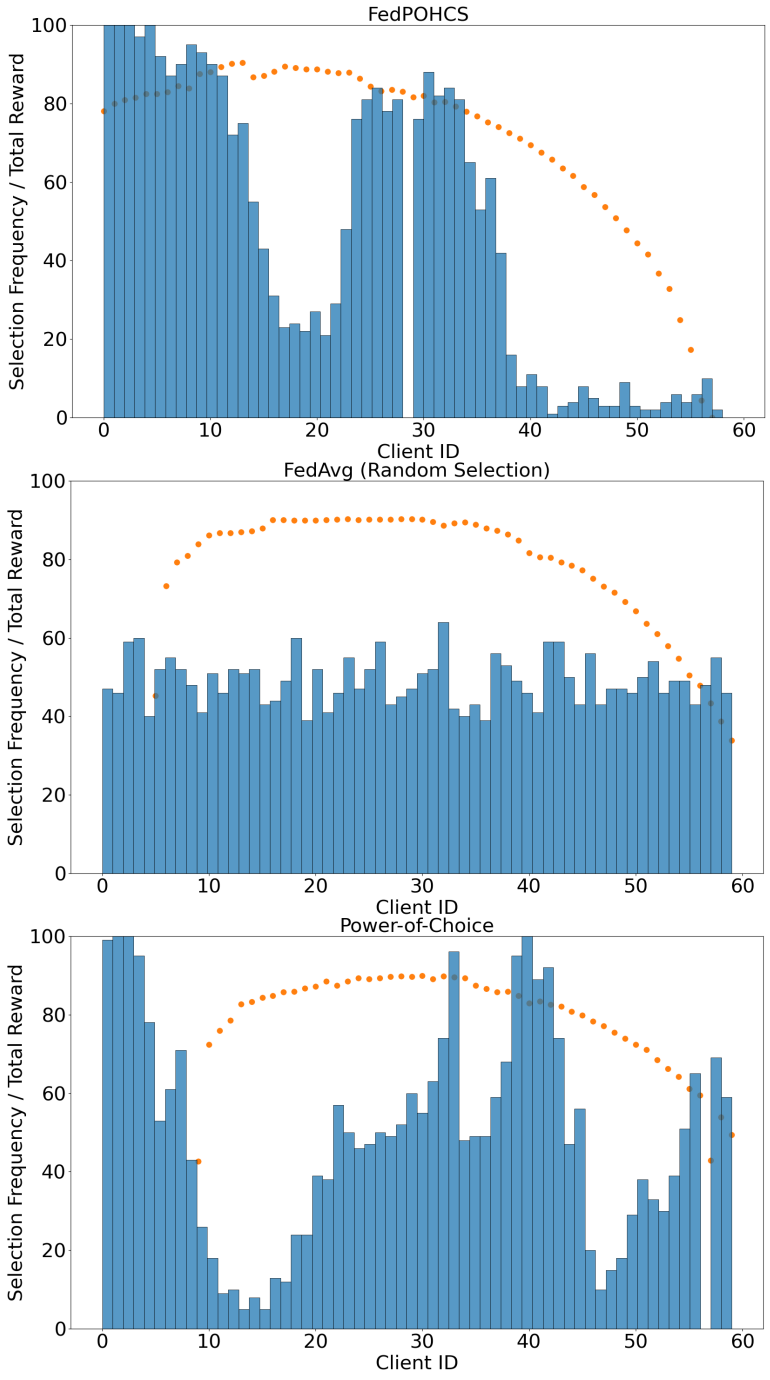
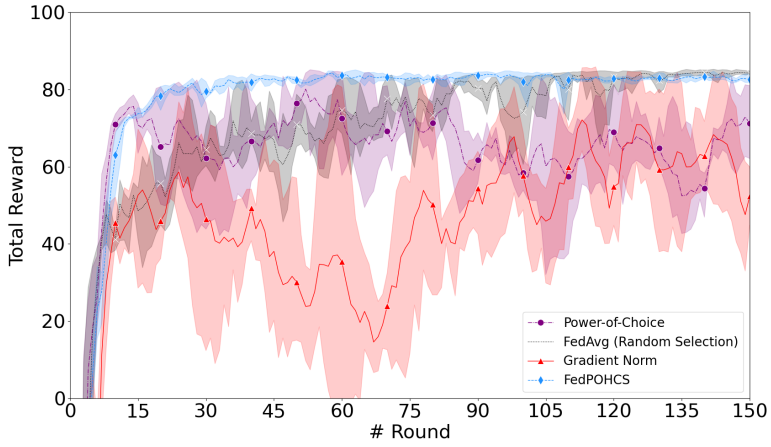
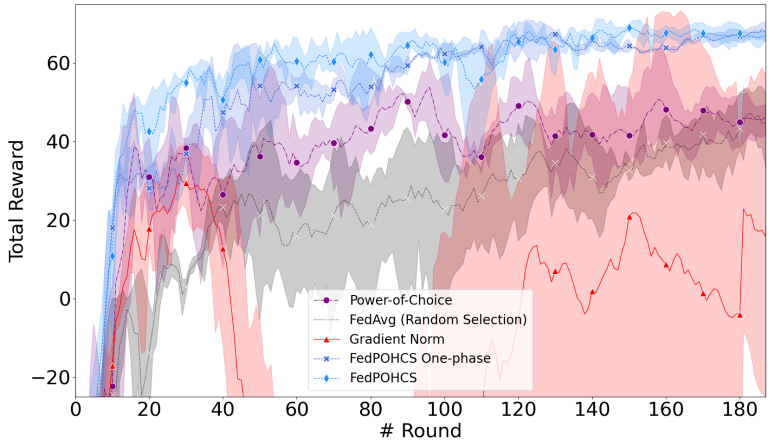


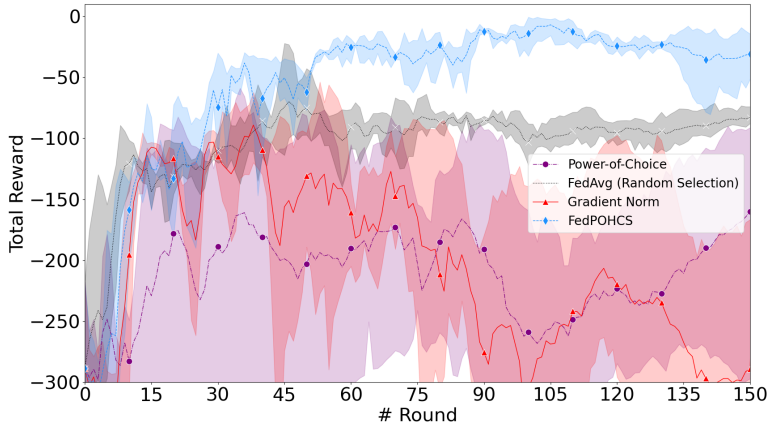
Figure 5: The histogram represents the frequency of each client being selected. The scatter points denote the return obtained by the final policy from each client. (top) FedPOHCS, (middle) FedAvg, and (bottom) Powder-of-Choice.



(a)

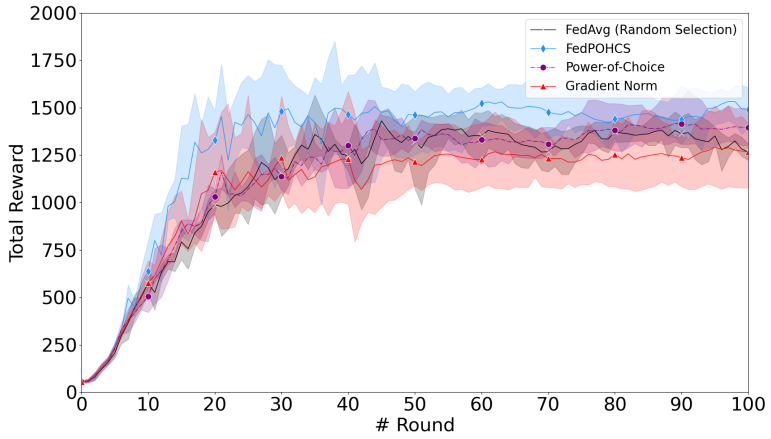


(b)

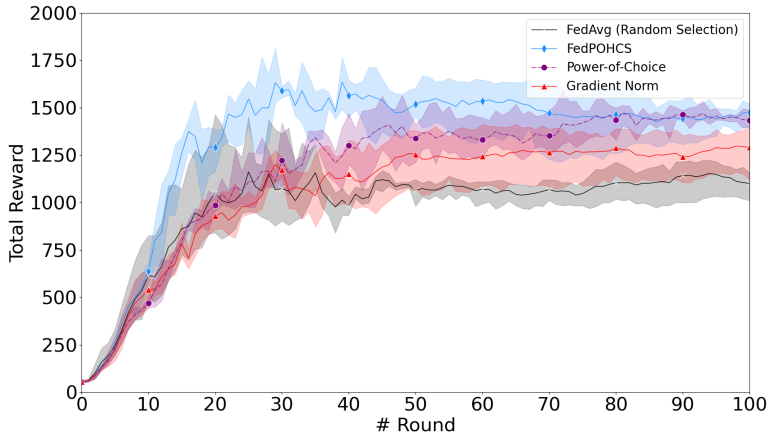


(c)

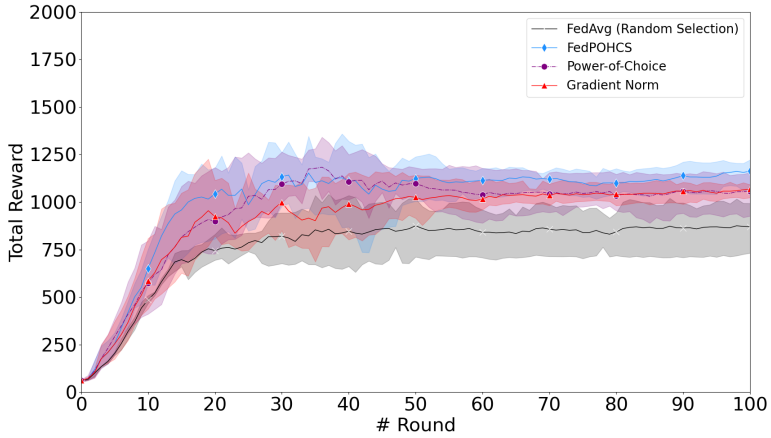
Figure 6: Comparison of FedAvg, Power-of-Choice, FedPOHCS, and GradientNorm on Mountain Cars with low/medium/high levels of heterogeneity.



(a)



(b)



(c)

Figure 7: Comparison of FedAvg, Power-of-Choice, FedPOHCS, and GradientNorm on Hoppers with low/medium/high levels of heterogeneity.

indicates that clients with IDs in $[1, 10]$ and clients with IDs in $[40, 60]$ are competing with each other. This is because they have constant shifts θ_n of different directions. Moreover, policies that perform well on clients with IDs in $[1, 10]$ may get fairly good scores on clients with IDs in $[40, 60]$, while the opposite is not true. We conjecture that the relation between the constant shifts θ_n and the transition probabilities is not linear and the imaginary MDP is closer to clients with small IDs, and hence learning on clients with IDs in $[40, 60]$ can be harmful to the overall performance.

5.5 Different levels of heterogeneity

In Figure 6, we show the performance comparison with different levels of heterogeneity, i.e., low (a), medium (b), and high (c), on Mountain Cars. It can be observed that, as the level of heterogeneity increases, the performance of all algorithms decreases, but the gap between FedPOHCS and the others increases. In other words, FedPOHCS demonstrates a bigger advantage when the level of heterogeneity is high, though its performance is also affected by severe heterogeneity. We have also conducted similar experiments on Hoppers as shown in Figure 7. While we can draw the same conclusion on FedAvg, i.e., its performance decreases as the level of heterogeneity increases, all three biased client selection methods are less affected by the level of heterogeneity.

6. Conclusion

In this work, we derived an error bound for federated policy optimization that explicitly unveils the impact of environment heterogeneity. The associated analysis covered various scenarios in FRL and offered insights into the effects of different federated settings. In particular, it was shown that clients whose environment dynamics are close to the population distribution are preferable for training. Based on these results, a client selection algorithm was proposed for FRL with heterogeneous clients. Experiment results demonstrated that the proposed client selection scheme outperforms other baselines on two federated RL problems. The results of this work represent a small step in understanding FRL and may motivate further research efforts in client selection for FRL.

Acknowledgments

This work was fully supported by a grant from the NSFC/RGC Joint Research Scheme sponsored by the Research Grants Council of the Hong Kong Special Administrative Region, China and National Natural Science Foundation of China (Project No. N_HKUST656/22).

APPENDICES

This appendix contains the proof of Lemma 5 in Appendix A, the proof of Lemma 6 in Appendix B, the proof of Proposition 7 in Appendix C, the proof of Theorem 10 in Appendix D, the proof of Proposition 9 in Appendix E, the proof of Proposition 15 in Appendix F, and additional experiment setting in Appendix H.

Appendix A. Proof of Lemma 5

Before proving Lemma 5, we introduce the following lemmas which will be frequently used throughout the appendix. Give the same value function V, Lemma 16 evaluates the difference induced by applying bellman operators on different MDPs, including the imaginary MDP.

Lemma 16 *For any value function $V \in \mathbb{R}^{|\mathcal{S}|}$, policy π and client n, m, we have*

$$\|T_m V - T_n V\| \leq \frac{\gamma R_{\max} \kappa_{m,n}}{1-\gamma}, \tag{24}$$

$$\|T_I V - T_n V\| \leq \frac{\gamma R_{\max} \kappa_{n,I}}{1-\gamma}, \tag{25}$$

$$\|T_m^\pi V - T_n^\pi V\| \leq \frac{\gamma R_{\max} \kappa_{m,n}}{1-\gamma}, \tag{26}$$

$$\|T_I^\pi V - T_n^\pi V\| \leq \frac{\gamma R_{\max} \kappa_{n,I}}{1-\gamma}. \tag{27}$$

Proof For any $s \in \mathcal{S}$ and $m, n = 1, \ldots, N$, let $a\_$, a_n, and a_m denote the greedy actions taken by T_I, T_n and T_m on $V(s)$, respectively. For example, $a\_ = \arg\max_a \mathcal{R}(s,a) + \gamma \sum_{s'} \bar{P}(s'|s,a) V(s')$ and $a_n = \arg\max_a \mathcal{R}(s,a) + \gamma \sum_{s'} P_n(s'|s,a) V(s')$. Then, for any state s, we have

$$|T_m V(s) - T_n V(s)| = \left| \mathcal{R}(s,a_m) - \mathcal{R}(s,a_n) + \gamma \sum_{s'} \left(P_m(s'|s,a_m) - P_n(s'|s,a_n) \right) V(s') \right|.$$

Without loss of generality, we assume that $T_m V(s) \geq T_n V(s)$. As a result, we can obtain the following inequality by replacing a_n with a_m

$$\begin{aligned} |T_m V(s) - T_n V(s)| &\leq \left| \gamma \sum_{s'} \left(P_m(s'|s,a_m) - P_n(s'|s,a_m) \right) V(s') \right| \\ &\leq \gamma \sum_{s'} \left| \left(P_m(s'|s,a_m) - P_n(s'|s,a_m) \right) \right| \left| V(s') \right| \\ &\leq \frac{\gamma R_{\max} \kappa_{m,n}}{1-\gamma}, \end{aligned}$$

where the last inequality follows from the fact that all value functions are bounded by $\frac{R_{\max}}{1-\gamma}$. This completes the proof of (24) and the proof of (25) is similar. Next, we prove (26). By

following a similar procedure, we can obtain

$$\begin{aligned}
|T_m^\pi V(s) - T_n^\pi V(s)| &= \left| \gamma \sum_{a,s'} \pi(a|s) \left(P_m(s'|s,a) - P_n(s'|s,a) \right) V(s') \right| \\
&\leq \gamma \sum_{s'} \left| \left(P_m^\pi(s'|s) - P_n^\pi(s'|s) \right) V(s') \right| \\
&\leq \gamma \sum_{s'} \left| P_m^\pi(s'|s) - P_n^\pi(s'|s) \right| \left| V(s') \right| \\
&\leq \frac{\gamma R_{\max} \kappa_{m,n}}{1-\gamma}.
\end{aligned}$$

This completes the proof of (26), and the proof of (27) is similar. ■

Given the same value function V and MDP, Lemma 17 evaluates the difference induced by applying bellman operators for different policies.

Lemma 17 *Let $\bar{\pi}^t(a|s) = \sum_{n=1}^N q_n \pi_n^t(a|s)$ denote the expected output of all local policies and $\tilde{\pi}^t(a|s) = \sum_{m \in \mathcal{C}} q'_m \pi_m^t(a|s), \forall s \in \mathcal{S}, a \in \mathcal{A}$ represent the expected output of a set $\mathcal{C}$ of local policies. Define $\bar{\varepsilon}_\theta = \max_t \left\| \pi^t(\cdot|s) - \bar{\pi}_n^t(\cdot|s) \right\|_2$ and $\tilde{\varepsilon}_\theta = \max_t \left\| \pi^t(\cdot|s) - \tilde{\pi}_m^t(\cdot|s) \right\|_2$ for full participation and partial participation, respectively. For any value function $V \in \mathbb{R}^{|\mathcal{S}|}$ and policy π^t at round t, we have*

$$\left\| T_I^{\pi^t} V - T_I^{\bar{\pi}^t} V \right\| \leq \frac{\bar{\varepsilon}_\theta \sqrt{|\mathcal{A}|} R_{\max}}{1-\gamma}, \tag{28}$$

$$\left\| T_n^{\pi^t} V - T_n^{\bar{\pi}^t} V \right\| \leq \frac{\bar{\varepsilon}_\theta \sqrt{|\mathcal{A}|} R_{\max}}{1-\gamma}, \tag{29}$$

$$\left\| T_I^{\pi^t} V - T_I^{\tilde{\pi}^t} V \right\| \leq \frac{\tilde{\varepsilon}_\theta \sqrt{|\mathcal{A}|} R_{\max}}{1-\gamma}, \tag{30}$$

$$\left\| T_n^{\pi^t} V - T_n^{\tilde{\pi}^t} V \right\| \leq \frac{\tilde{\varepsilon}_\theta \sqrt{|\mathcal{A}|} R_{\max}}{1-\gamma}. \tag{31}$$

Proof For any state s, we have

$$\begin{aligned}
\left| T_I^{\pi^t} V^t(s) - T_I^{\bar{\pi}^t} V^t(s) \right| &= \left| \sum_a \sum_{n=1}^N q_n \left(\pi^t(a|s) - \pi_n^t(a|s) \right) \left(\mathcal{R}(s,a) + \gamma \sum_{s'} \bar{P}(s'|s,a) V^t(s') \right) \right| \\
&\leq \left\| \pi^t(\cdot|s) - \sum_{n=1}^N q_n \pi_n^t(\cdot|s) \right\|_2 \left\| \mathcal{R}(s,\cdot) + \gamma \sum_{s'} \bar{P}(s'|s,\cdot) V^t(s') \right\|_2 \\
&\leq \frac{\sqrt{|\mathcal{A}|} R_{\max}}{1-\gamma} \left\| \pi^t(\cdot|s) - \sum_{n=1}^N q_n \pi_n^t(\cdot|s) \right\|_2 \\
&\leq \frac{\bar{\varepsilon}_\theta \sqrt{|\mathcal{A}|} R_{\max}}{1-\gamma},
\end{aligned}$$

which completes the proof of (28) and the proofs of (29), (30), and (31) are similar. ■

Next, we prove Lemma 5.

Proof Let $\bar{\pi}^t(a|s) = \sum_{n=1}^N q_n \pi_n^t(a|s)$ denote the expected output of all local policies. For any state s, we have

$$\left|T_I^{\pi^{t+1}} V^t(s) - T_I V^t(s)\right| \leq \left|T_I^{\pi^{t+1}} V^t(s) - T_I^{\bar{\pi}^{t+1}} V^t(s)\right| + \left|T_I^{\bar{\pi}^{t+1}} V^t(s) - T_I V^t(s)\right|. \quad (32)$$

By Lemma 17, the first term on the right-hand side (RHS) of (32) is upper bounded by

$$\left|T_I^{\pi^{t+1}} V^t(s) - T_I^{\bar{\pi}^{t+1}} V^t(s)\right| \leq \frac{\bar{\varepsilon}_\theta \sqrt{|\mathcal{A}|} R_{\max}}{1-\gamma}. \quad (33)$$

Now we bound the second term on the RHS of (32). For any state s, we have

$$\begin{aligned}
&\left|T_I^{\bar{\pi}^{t+1}} V^t(s) - T_I V^t(s)\right| \\
&= \left|\sum_a \bar{\pi}^{t+1}(a|s) \left(\mathcal{R}(s,a) + \gamma \sum_{s'} \bar{P}(s'|s,a) V^t(s')\right) - T_I V^t\right| \\
&\overset{(a)}{=} \left|\sum_{n=1}^N q_n \left[\sum_a \pi_n^{t+1}(a|s) \left(\mathcal{R}(s,a) + \gamma \sum_{s'} \bar{P}(s'|s,a) V^t(s')\right) - T_I V^t\right]\right| \\
&\overset{(b)}{=} \left|\sum_{n=1}^N q_n \left[\gamma \sum_{a,s'} \pi_n^{t+1}(a|s) \left(\bar{P}(s'|s,a) - P_n(s'|s,a)\right) V^t(s') + T_n^{\pi_n^{t+1}} V^t(s) - T_I V^t\right]\right| \\
&\overset{(c)}{\leq} \left|\sum_{n=1}^N q_n \gamma \sum_{a,s'} \pi_n^{t+1}(a|s) \left(\bar{P}(s'|s,a) - P_n(s'|s,a)\right) V^t(s')\right| \\
&\quad + \left|\sum_{n=1}^N q_n \left(T_n^{\pi_n^{t+1}} V^t(s) - T_n V^t(s)\right)\right| + \left|\sum_{n=1}^N q_n \left(T_n V^t(s) - T_I V^t(s)\right)\right|. \quad (34)
\end{aligned}$$

Step (a) follows from $\bar{\pi}^{t+1}(a|s) = \sum_{n=1}^N q_n \pi_n^{t+1}(a|s), \forall s \in \mathcal{S}, a \in \mathcal{A}$. In step (b), we add and then subtract the term $\sum_{a,s'} \pi_n^{t+1}(a|s) P_n(s'|s,a) V^t(s')$. The added term is combined with $\sum_{a,s'} \pi_n^{t+1}(a|s) \mathcal{R}(s,a)$ to form $T_n^{\pi_n^{t+1}} V^t(s)$. Step (c) follows from the triangle inequality.

By Lemma 16, the first term on the RHS of (34) is upper bounded by

$$\begin{aligned}
\left|\sum_{n=1}^N q_n \gamma \sum_{a,s'} \pi_n^{t+1}(a|s) (\bar{P}(s'|s,a) - P_n(s'|s,a)) V^t(s')\right| &= \left|\sum_{n=1}^N q_n \left(T_I^{\pi_n^{t+1}} V^t(s) - T_n^{\pi_n^{t+1}} V^t(s)\right)\right| \\
&\leq \frac{\gamma R_{\max} \kappa_1}{1-\gamma}. \quad (35)
\end{aligned}$$

By (7), the second term on the RHS of (34) is upper bounded by $\bar{\epsilon}$. By Lemma 16, the third term on the RHS of (34) is upper bounded by

$$\left|\sum_{n=1}^N q_n \left(T_n V^t(s) - T_I V^t(s)\right)\right| \leq \frac{\gamma R_{\max} \kappa_1}{1-\gamma}. \quad (36)$$

By substituting the above-mentioned three upper bounds into (34), we can obtain

$$\left|T_I^{\pi^{t+1}} V^t(s) - T_I V^t(s)\right| \leq \frac{2\gamma R_{\max}\kappa_1}{1-\gamma} + \bar{\epsilon}. \tag{37}$$

By substituting (33) and (37) into (32), we can obtain

$$\left|T_I^{\pi^{t+1}} V^t(s) - T_I V^t(s)\right| \leq \frac{\bar{\varepsilon}_\theta \sqrt{|\mathcal{A}|} R_{\max}}{1-\gamma} + \frac{2\gamma R_{\max}\kappa_1}{1-\gamma} + \bar{\epsilon}.$$

■

Appendix B. Proof of Lemma 6

To prove Lemma 6, we first introduce Lemma 18, which bounds the difference between value functions of policy π in different clients.

Lemma 18 *For any state s, policy π and clients m, n, we have*

$$|V_m^\pi(s) - V_n^\pi(s)| \leq \frac{\gamma R_{\max}\kappa_{m,n}}{(1-\gamma)^2}.$$

Proof For any states s, the distance between $V_m^\pi(s)$ and $V_n^\pi(s)$ can be bounded as

$$\begin{aligned}
&|V_m^\pi(s) - V_n^\pi(s)| \\
&\overset{(a)}{=} \left|\sum_a \pi(a|s)\left(\mathcal{R}(s,a) + \gamma \sum_{s'\in\mathcal{S}} P_m(s'|s,a)V_m^\pi(s') - \mathcal{R}(s,a) - \gamma \sum_{s'\in\mathcal{S}} P_n(s'|s,a)V_n^\pi(s')\right)\right| \\
&\overset{(b)}{=} \left|\gamma \sum_{s'\in\mathcal{S}} \left(P_m^\pi(s'|s)V_m^\pi(s') - P_m^\pi(s'|s)V_n^\pi(s') + P_m^\pi(s'|s)V_n^\pi(s') - P_n^\pi(s'|s)V_n^\pi(s')\right)\right| \\
&\overset{(c)}{\leq} \gamma \sum_{s'\in\mathcal{S}} P_m^\pi(s'|s)\left|V_m^\pi(s') - V_n^\pi(s')\right| + \gamma \sum_{s'\in\mathcal{S}} \left|P_m^\pi(s'|s) - P_n^\pi(s'|s)\right|\left|V_n^\pi(s')\right| \\
&\overset{(d)}{\leq} \gamma \max_{s'} \left|V_m^\pi(s') - V_n^\pi(s')\right| + \frac{\gamma R_{\max} \sum_{s'\in\mathcal{S}} \left|P_m^\pi(s'|s) - P_n^\pi(s'|s)\right|}{1-\gamma}.
\end{aligned} \tag{38}$$

Step (a) follows from Bellman's equation. In step (b), we added and then subtracted the term $P_m^\pi(s'|s)V_n^\pi(s')$. Step (c) follows from the triangle inequality. Step (d) follows from the fact that all value functions are bounded by $\frac{R_{\max}}{1-\gamma}$. By taking the maximum of both sides of (38) over state s and after some mathematical manipulations, we can finally obtain

$$|V_m^\pi(s) - V_n^\pi(s)| \leq \frac{\gamma R_{\max}\kappa_{m,n}}{(1-\gamma)^2}.$$

■

Note that we prove Lemma 18 for the state-value function, and a similar result for the action-value function was given by a previous work (Strehl and Littman, 2008, Lemma 1).

Next, we prove Lemma 6.

Proof By the triangle inequality, we have

$$\left\|T_I^{\pi^{t+1}}\bar{V}^t - T_I\bar{V}^t\right\| \leq \left\|T_I^{\pi^{t+1}}\bar{V}^t - T_I^{\bar{\pi}^{t+1}}\bar{V}^t\right\| + \left\|T_I^{\bar{\pi}^{t+1}}\bar{V}^t - T_I\bar{V}^t\right\|,$$

from which we can obtain

$$\left\|T_I^{\pi^{t+1}}V^t - T_IV^t\right\| \leq \frac{\bar{\varepsilon}_\theta\sqrt{|\mathcal{A}|}R_{\max}}{1-\gamma} + \left\|T_I^{\bar{\pi}^{t+1}}\bar{V}^t - T_I\bar{V}^t\right\|, \tag{39}$$

by Lemma 17. To finish the proof, it suffices to bound the third term on the RHS of (39). By the definition of the Bellman operators, we have

$$\left\|T_I^{\bar{\pi}^{t+1}}\bar{V}^t - T_I\bar{V}^t\right\| = \left\|\sum_{n=1}^N q_n\left(T_I^{\pi_n^{t+1}}\bar{V}^t - T_n\bar{V}^t\right)\right\| \leq \sum_{n=1}^N q_n\left\|T_I^{\pi_n^{t+1}}\bar{V}^t - T_n\bar{V}^t\right\|. \tag{40}$$

Next, we further bound the RHS of (40). By the triangle inequality, we have

$$\left\|T_I^{\pi_n^{t+1}}\bar{V}^t - T_n\bar{V}^t\right\| \leq \left|T_I^{\pi_n^{t+1}}\bar{V}^t - T_I^{\pi_n^{t+1}}V_n^t\right\| + \left\|T_I^{\pi_n^{t+1}}V_n^t - T_nV_n^t\right\| + \left\|T_nV_n^t - T_n\bar{V}^t\right\|,$$

from which we can obtain

$$\left\|T_I^{\pi_n^{t+1}}\bar{V}^t - T_n\bar{V}^t\right\| \leq 2\gamma\left\|\bar{V}^t - V_n^t\right\| + \left\|T_I^{\pi_n^{t+1}}V_n^t - T_nV_n^t\right\| \tag{41}$$

by the contraction property of the Bellman operators.

Next, we bound the first term on the RHS of (41). For any state s, we have

$$\begin{aligned}\left|\bar{V}^t(s) - V_n^t(s)\right| &= \left|\sum_{j=1}^N q_j\left(V_n^t(s) - V_j^t(s)\right)\right| \\ &\leq \sum_{j=1}^N q_j\left|V_n^{\pi^t}(s) - V_j^{\pi^t}(s)\right| + \left|V_n^{\pi^t}(s) - V_n^t(s)\right| + \sum_{j=1}^N q_j\left|V_j^{\pi^t}(s) - V_j^t(s)\right|\end{aligned}$$

due to the triangle inequality. By Lemma 18 and (7), we can further obtain

$$\left|\bar{V}^t(s) - V_n^t(s)\right| \leq \sum_{j=1}^N q_j\frac{\gamma R_{\max}\kappa_{n,j}}{(1-\gamma)^2} + \bar{\delta} + \delta_n.$$

Thus, we have the following bound

$$\left\|\bar{V}^t - V_n^t\right\| \leq \sum_{j=1}^N q_j\frac{\gamma R_{\max}\kappa_{n,j}}{(1-\gamma)^2} + \bar{\delta} + \delta_n. \tag{42}$$

Now we bound the second term on the RHS of (41). For any $s \in \mathcal{S}$, we have

$$\begin{aligned}\left|T_I^{\pi_n^{t+1}}V_n^t(s) - T_nV_n^t(s)\right| &\leq \left|T_I^{\pi_n^{t+1}}V_n^t(s) - T_n^{\pi_n^{t+1}}V_n^t(s)\right| + \left|T_n^{\pi_n^{t+1}}V_n^t(s) - T_nV_n^t(s)\right| \\ &\leq \frac{\gamma R_{\max}\kappa_{n,I}}{1-\gamma} + \epsilon_n,\end{aligned}$$

where the last inequality follows from Lemma 16 and (7). Thus, we can obtain

$$\left\|T_I^{\pi_n^{t+1}} V_n^t - T_n V_n^t\right\| \leq \frac{\gamma R_{\max} \kappa_{n,I}}{1-\gamma} + \epsilon_n. \tag{43}$$

By substituting (42) and (43) into (41), and then substituting (41) into (40), we have

$$\left\|T_I^{\pi^{t+1}} \bar{V}^t - T_I \bar{V}^t\right\| \leq \frac{2\gamma^2 R_{\max} \kappa_2}{(1-\gamma)^2} + \frac{\gamma R_{\max} \kappa_1}{1-\gamma} + 4\gamma\bar{\delta} + \bar{\epsilon}. \tag{44}$$

By substituting (44) into (39), we can obtain

$$\left\|T_I^{\pi^{t+1}} \bar{V}^t - T_I \bar{V}^t\right\| \leq \frac{\bar{\varepsilon}_\theta \sqrt{|\mathcal{A}|} R_{\max}}{1-\gamma} + \frac{2\gamma^2 R_{\max} \kappa_2}{(1-\gamma)^2} + \frac{\gamma R_{\max} \kappa_1}{1-\gamma} + 4\gamma\bar{\delta} + \bar{\epsilon}.$$

■

Appendix C. Proof of Proposition 7

Proof Let $\tilde{\pi}^{t+1}(a|s) = \sum_{m\in\mathcal{C}} q'_m \pi_m^{t+1}(a|s), \forall s \in \mathcal{S}, a \in \mathcal{A}$ denote the expected output of a set $\mathcal{C}$ of local policies. By the triangle inequality, we have

$$\left\|T_I^{\pi^{t+1}} \bar{V}^t - T_I \bar{V}^t\right\| \leq \left\|T_I^{\pi^{t+1}} \bar{V}^t - T_I^{\tilde{\pi}^{t+1}} \bar{V}^t\right\| + \left\|T_I^{\tilde{\pi}^{t+1}} \bar{V}^t - T_I \bar{V}^t\right\|,$$

from which we can obtain

$$\left\|T_I^{\pi^{t+1}} V^t - T_I V^t\right\| \leq \frac{\tilde{\varepsilon}_\theta \sqrt{|\mathcal{A}|} R_{\max}}{1-\gamma} + \left\|T_I^{\tilde{\pi}^{t+1}} \bar{V}^t - T_I \bar{V}^t\right\|, \tag{45}$$

by Lemma 17. To finish the proof, it suffices to bound the third term on the RHS of (45). By the definition of the Bellman operators, we have

$$\begin{aligned}
\left\|T_I^{\tilde{\pi}^{t+1}} \bar{V}^t - T_I \bar{V}^t\right\| &= \left\|\sum_{m\in\mathcal{C}} q'_m \sum_{n=1}^{N} q_n \left(T_I^{\pi_m^{t+1}} \bar{V}^t - T_n \bar{V}^t\right)\right\| \\
&\leq \sum_{m\in\mathcal{C}} q'_m \sum_{n=1}^{N} q_n \left\|T_I^{\pi_m^{t+1}} \bar{V}^t - T_n \bar{V}^t\right\|.
\end{aligned} \tag{46}$$

The RHS of (46) can be further bounded by

$$\begin{aligned}
&\left\|T_I^{\pi_m^{t+1}} \bar{V}^t - T_n \bar{V}^t\right\| \\
&\leq \left\|T_I^{\pi_m^{t+1}} \bar{V}^t - T_I^{\pi_m^{t+1}} V_m^t\right\| + \left\|T_I^{\pi_m^{t+1}} V_m^t - T_m V_m^t\right\| + \left\|T_m V_m^t - T_n V_m^t\right\| + \left\|T_n V_m^t - T_n \bar{V}^t\right\| \\
&\overset{(a)}{\leq} 2\gamma \left(\left\|\bar{V}^t - V_m^{\pi^t}\right\| + \left\|V_m^{\pi^t} - V_m^t\right\|\right) + \left\|T_I^{\pi_m^{t+1}} V_m^t - T_m^{\pi_m^{t+1}} V_m^t\right\| \\
&\quad + \left\|T_m^{\pi_m^{t+1}} V_m^t - T_m V_m^t\right\| + \left\|T_m V_m^t - T_n V_m^t\right\| \\
&\overset{(b)}{\leq} \sum_{j=1}^{N} q_j \frac{2\gamma^2 R_{\max} \kappa_{m,j}}{(1-\gamma)^2} + 2\gamma\bar{\delta} + 2\gamma\delta_m + \epsilon_m + \frac{\gamma R_{\max} \kappa_{m,n}}{1-\gamma} + \frac{\gamma R_{\max} \kappa_{m,I}}{1-\gamma},
\end{aligned} \tag{47}$$

where step (a) follows from the contraction property of the Bellman operators and the triangle inequality. Step (b) follows from Lemma 16 and (7). Thus, by substituting (47) into the RHS of (46), we can obtain

$$\begin{aligned}\left\|T_I^{\pi_I^{t+1}}\bar{V}^{\pi^t}-T_I\bar{V}^{\pi^t}\right\| &\le \sum_{m\in\mathcal{C}}\sum_{n=1}^{N} q'_m q_n \left\|T_I^{\pi_m^{t+1}}\bar{V}^{\pi^t}-T_n\bar{V}^{\pi^t}\right\| \\ &\le \sum_{m\in\mathcal{C}}\sum_{n=1}^{N} q'_m q_n \frac{\left(\gamma+\gamma^2\right)R_{\max}\kappa_{m,n}}{(1-\gamma)^2}+\sum_{m\in\mathcal{C}} q'_m\frac{\gamma R_{\max}\kappa_{m,I}}{1-\gamma} \\ &\quad + 2\gamma\sum_{m\in\mathcal{C}} q'_m\delta_m + 2\gamma\bar{\delta}+\sum_{m\in\mathcal{C}} q'_m\epsilon_m. \end{aligned} \tag{48}$$

By substituting (48) into (45), we can conclude

$$\begin{aligned}\left\|T_I^{\pi_I^{t+1}}\bar{V}^{t}-T_I\bar{V}^{t}\right\| \le{}& \frac{\tilde{\varepsilon}_\theta\sqrt{|\mathcal{A}|}R_{\max}}{1-\gamma}+\sum_{m\in\mathcal{C}}\sum_{n=1}^{N} q'_m q_n \frac{\left(\gamma+\gamma^2\right)R_{\max}\kappa_{m,n}}{(1-\gamma)^2} \\ &+\sum_{m\in\mathcal{C}} q'_m\frac{\gamma R_{\max}\kappa_{m,I}}{1-\gamma}+2\gamma\sum_{m\in\mathcal{C}} q'_m\delta_m+2\gamma\bar{\delta}+\sum_{m\in\mathcal{C}} q'_m\epsilon_m.\end{aligned}$$

■

Appendix D. Proof of Theorem 10

Proof By the triangle inequality, for any state s, we have

$$\left|\bar{V}^{\pi^t}(s)-\bar{V}^{\pi^*}(s)\right| \le \left|\bar{V}^{\pi^t}(s)-V_I^{\pi^t}(s)\right|+\left|V_I^{\pi^t}(s)-V_I^{\pi_I^*}(s)\right|+\left|V_I^{\pi_I^*}(s)-\bar{V}^{\pi^*}(s)\right|, \tag{49}$$

where the first and second terms on the RHS of (49) can be upper bounded by Lemma 4 and Proposition 1, respectively.

To bound the third term on the RHS of (49), we notice that, for certain states, the distance between the value function $V_I^{\pi_I^*}=V_I^*$ for the optimal policy π_I^* in the imaginary MDP $\mathcal{M}_I$ and the averaged value function $\bar{V}^{\pi^*}$ of the optimal policy π^* for (3) is upper bounded. Specifically, for state s such that $\bar{V}^{\pi^*}(s)\ge\bar{V}^{\pi_I^*}(s)$, we have

$$\left|V_I^{\pi_I^*}(s)-\bar{V}^{\pi^*}(s)\right| \le \left|V_I^{\pi^*}(s)-\bar{V}^{\pi^*}(s)\right| \le \frac{\gamma R_{\max}\kappa_1}{(1-\gamma)^2}. \tag{50}$$

By substituting (50), Lemma 4 and Proposition 1 into (49), for a given state s with $\bar{V}^{\pi^*}(s)\ge\bar{V}^{\pi_I^*}(s)$, we have

$$\limsup_{t\to\infty}\left|\bar{V}^{\pi^t}(s)-\bar{V}^{\pi^*}(s)\right| \le \frac{\gamma R_{\max}\kappa_1}{(1-\gamma)^2}+\frac{\tilde{\epsilon}+2\gamma\tilde{\delta}}{(1-\gamma)^2}+\frac{\gamma R_{\max}\kappa_1}{(1-\gamma)^2}, \tag{51}$$

where $\tilde{\delta}$ may be $\bar{\delta}$ (8) or $\dot{\delta}$ (10), $\tilde{\epsilon}$ may be one of $\dot{\epsilon}$ (Lemma 6), ϵ' (Lemma 5), $\hat{\epsilon}$ (Proposition 7) or $\acute{\epsilon}$ (Proposition 9).

When the aforementioned condition does not hold, we can alternatively bound the distance between $\bar{V}^{\pi^t}(s)$ and $\bar{V}^{\pi_I^*}(s)$ as

$$\left|\bar{V}^{\pi^t}(s)-\bar{V}^{\pi_I^*}(s)\right| \leq \left|\bar{V}^{\pi^t}(s)-V_I^{\pi^t}(s)\right| + \left|V_I^{\pi^t}(s)-V_I^{\pi_I^*}(s)\right| + \left|V_I^{\pi_I^*}(s)-\bar{V}^{\pi_I^*}(s)\right|, \quad (52)$$

and actually π_I^* performs better on these states.

By substituting Lemma 4 and Proposition 1 into (52), for a given state s with $\bar{V}^{\pi^*}(s) \leq \bar{V}^{\pi_I^*}(s)$, we have

$$\limsup_{t\to\infty}\left|\bar{V}^{\pi^t}(s)-\bar{V}^{\pi_I^*}(s)\right| \leq \frac{\gamma R_{\max}\kappa_1}{(1-\gamma)^2} + \frac{\tilde{\epsilon}+2\gamma\tilde{\delta}}{(1-\gamma)^2} + \frac{\gamma R_{\max}\kappa_1}{(1-\gamma)^2}. \quad (53)$$

Let $\bar{V}_s^{\max} = \max\left\{\bar{V}^{\pi^*}(s), \bar{V}^{\pi_I^*}(s)\right\}, \forall s \in \mathcal{S}$. We can then combine (51) and (53) into

$$\limsup_{t\to\infty}\left|\bar{V}^{\pi^t}(s)-\bar{V}_s^{\max}\right| \leq \frac{\tilde{\epsilon}+2\gamma\tilde{\delta}}{(1-\gamma)^2} + 2\frac{\gamma R_{\max}\kappa_1}{(1-\gamma)^2},$$

which proves (14) in Theorem 10. The error bound for Algorithm 2 with partial client participation can be obtained by replacing $\tilde{\delta}$ with $\bar{\delta}$ and replacing $\tilde{\epsilon}$ with $\hat{\epsilon}$ as

$$\begin{aligned}
\limsup_{t\to\infty}\left|\bar{V}^{\pi^t}(s)-\bar{V}_s^{\max}\right| &\leq \frac{\hat{\epsilon}+2\gamma\bar{\delta}}{(1-\gamma)^2} + 2\frac{\gamma R_{\max}\kappa_1}{(1-\gamma)^2} \\
&= \frac{2\gamma(\gamma^2-\gamma+1)}{(1-\gamma)^4}R_{\max}\kappa_1 + \frac{\gamma}{(1-\gamma)^3}R_{\max}\sum_{m\in\mathcal{C}} q'_m\kappa_{m,I} \\
&\quad + \frac{\gamma+\gamma^2}{(1-\gamma)^4}R_{\max}\sum_{m\in\mathcal{C}}\sum_{n=1}^{N} q'_m q_n \kappa_{m,n} \\
&\quad + \frac{\tilde{\varepsilon}_\theta\sqrt{|\mathcal{A}|}R_{\max}}{(1-\gamma)^3} + \tilde{\mathcal{O}}\left(\bar{\delta} + \sum_{m\in\mathcal{C}} q'_m\delta_m + \sum_{m\in\mathcal{C}} q'_m\epsilon_m\right),
\end{aligned}$$

where $\tilde{\mathcal{O}}$ omits some constants related to γ. This proves (15) in Theorem 10. ■

Next, we will discuss what Theorem 10 indicates.

D.1 Interpretation of the Error Bound

It is difficult to analyze FAPI because we can not directly apply the results of API to FAPI. Fortunately, the use of the imaginary MDP M_I aligns FAPI with API and enables the application of Proposition 1 in Theorem 10.

However, the imaginary MDP $\mathcal{M}_I$ also brings two problems: (1) While π^* is the optimal policy for the objective function of FRL (3), Proposition 1 can only show how far the generated policy π^t is from the optimal policy π_I^* in the imaginary MDP (refer to Remark 11); and (2) The performance (V_I^π) of any policy π in the imaginary MDP $\mathcal{M}_I$ does not reflect its performance ($\bar{V}^\pi$) on FRL (3). These problems make the proof intractable as we have to bound the distance between $\bar{V}^{\pi^*}$ and $V_I^{\pi_I^*}$, which is not always feasible. As shown in (50), their distance on a given state s is bounded only when $\bar{V}^{\pi^*}(s) \geq \bar{V}^{\pi_I^*}(s)$. The difficulty

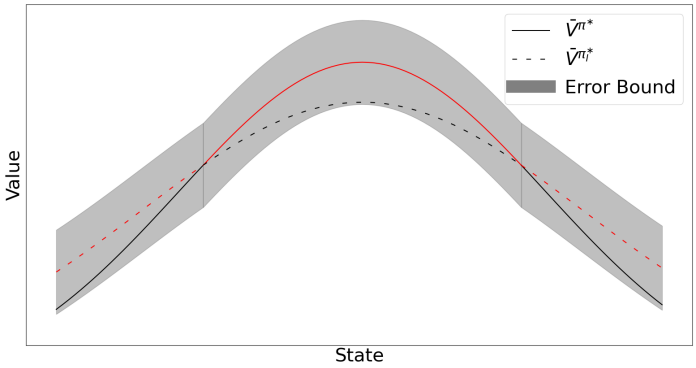


Figure 8: The area in grey indicates the error bound in (14). The maximum of the two curves on each state is highlighted in red, and the bound is drawn with respect to this highlighted curve.

of bounding their distance on all states stems from the fact that the optimal policy π^* for (3) does not necessarily outperform other policies on every state (i.e., it may not be uniformly the best).

For the states where π_I^* outperforms π^*, we can alternatively bound the error with respect to $\bar{V}^{\pi_I^*}$. Although $\bar{V}^{\pi_I^*}$ is not directly related to the objective function of FRL (3), it can be regarded as an approximation to $\bar{V}^{\pi^*}$, especially when the level of heterogeneity is low where $\bar{V}^{\pi_I^*}$ is close to $\bar{V}^{\pi^*}$. As a result, the error bound in (14) is a combination of two bounds with respect to $\bar{V}^{\pi_I^*}$ and $\bar{V}^{\pi^*}$, respectively. An illustrative example is given in Figure 8.

Appendix E. Proof of Proposition 9

Proof Let $\tilde{\pi}^{t+1}(a|s) = \sum_{m\in\mathcal{C}} q'_m \pi_m^{t+1}(a|s), \forall s \in \mathcal{S}, a \in \mathcal{A}$ denote the expected output of a set $\mathcal{C}$ of local policies. For any state s, we have

$$\left|T_I^{\pi^{t+1}} V^t(s) - T_I V^t(s)\right| \leq \left|T_I^{\pi^{t+1}} V^t(s) - T_I^{\bar{\pi}^{t+1}} V^t(s)\right| + \left|T_I^{\bar{\pi}^{t+1}} V^t(s) - T_I V^t(s)\right|. \tag{54}$$

By Lemma 17, the first term on the right-hand side (RHS) of (54) is upper bounded by

$$\left|T_I^{\pi^{t+1}} V^t(s) - T_I^{\bar{\pi}^{t+1}} V^t(s)\right| \leq \frac{\bar{\varepsilon}_\theta \sqrt{|\mathcal{A}|} R_{\max}}{1-\gamma}. \tag{55}$$

To finish the proof, it suffices to bound the second term on the RHS of (54). By the definition of the Bellman operators, we have

$$\begin{aligned}\left\|T_I^{\tilde{\pi}^{t+1}} V^t - T_I V^t\right\| &= \left\|\sum_{m\in\mathcal{C}} q'_m \sum_{n=1}^{N} q_n \left(T_I^{\pi_m^{t+1}} V^t - T_n V^t\right)\right\| \\ &\leq \sum_{m\in\mathcal{C}} q'_m \sum_{n=1}^{N} q_n \left\|T_I^{\pi_m^{t+1}} V^t - T_n V^t\right\|. \end{aligned} \tag{56}$$

The RHS of (56) can be further bounded by

$$\begin{aligned}\left\|T_I^{\pi_m^{t+1}}V^t - T_nV^t\right\| &\leq \left\|T_I^{\pi_m^{t+1}}V^t - T_m^{\pi_m^{t+1}}V^t\right\| + \left\|T_m^{\pi_m^{t+1}}V^t - T_mV^t\right\| + \left\|T_mV^t - T_nV^t\right\| \\ &\leq \frac{\gamma R_{\max}\kappa_{m,I}}{1-\gamma} + \epsilon_m + \frac{\gamma R_{\max}\kappa_{m,n}}{1-\gamma}, \end{aligned} \tag{57}$$

where the last inequality follows from Lemma 16 and (7). By substituting (57) into (56), we can obtain

$$\begin{aligned}\left\|T_I^{\pi^{t+1}}V^{\pi^t} - T_IV^{\pi^t}\right\| &\leq \sum_{m\in\mathcal{C}}\sum_{n=1}^{N} q'_m q_n \left\|T_I^{\pi_m^{t+1}}V^{\pi^t} - T_nV^{\pi^t}\right\| \\ &\leq \sum_{m\in\mathcal{C}}\sum_{n=1}^{N} q'_m q_n \frac{\gamma R_{\max}\kappa_{m,n}}{1-\gamma} + \sum_{m\in\mathcal{C}} q'_m \frac{\gamma R_{\max}\kappa_{m,I}}{1-\gamma} + \sum_{m\in\mathcal{C}} q'_m\epsilon_m. \end{aligned} \tag{58}$$

By substituting (58) into (54), we can conclude

$$\left\|T_I^{\pi^{t+1}}V^t - T_IV^t\right\| \leq \frac{\tilde{\varepsilon}_\theta\sqrt{|\mathcal{A}|}R_{\max}}{1-\gamma} + \sum_{m\in\mathcal{C}} q'_m\left(\sum_{n=1}^{N} q_n\frac{\gamma R_{\max}\kappa_{m,n}}{1-\gamma} + \frac{\gamma R_{\max}\kappa_{m,I}}{1-\gamma} + \epsilon_m\right).$$

■

Appendix F. Proof of Proposition 15

Proof By Theorem 10, we have

$$\limsup_{t\to\infty}\left|\bar{V}^{\pi^t}(s) - \bar{V}_s^{\max}\right| \leq \frac{\tilde{\epsilon} + 2\gamma\dot{\delta}}{(1-\gamma)^2} + 2\frac{\gamma R_{\max}\kappa_1}{(1-\gamma)^2},$$

where $\tilde{\epsilon}$ may be one of $\acute{\epsilon}$ (Proposition 9) or ϵ' (Lemma 5). Since the environments are homogeneous, we have

$$\begin{aligned}\bar{V}_s^{\max} &= \bar{V}^{\pi^*}(s) = \bar{V}^{\pi_I^*}(s), \\ \kappa_1 &= \kappa_2 = 0, \\ \dot{\delta} &= \bar{\varepsilon}_w + \bar{\delta}, \\ \epsilon' &= \acute{\epsilon} = \bar{\epsilon} + \frac{\hat{\varepsilon}_\theta\sqrt{|\mathcal{A}|}R_{\max}}{1-\gamma}, \end{aligned}$$

where $\hat{\varepsilon}_\theta$ is equal to $\bar{\varepsilon}_\theta$ and $\tilde{\varepsilon}_\theta$ for full participation and partial participation, respectively. Thus, we can conclude

$$\limsup_{t\to\infty}\left\|\bar{V}^{\pi^t} - \bar{V}^{\pi^*}\right\| \leq \frac{\hat{\varepsilon}_\theta\sqrt{|\mathcal{A}|}R_{\max}}{(1-\gamma)^3} + \frac{2\gamma\bar{\varepsilon}_w}{(1-\gamma)^2} + \frac{\bar{\epsilon} + 2\gamma\bar{\delta}}{(1-\gamma)^2}.$$

■

Appendix G. Proof of Lemma 12

Our proof relies on the definition of local linearization for the two-layer neural network at its random initialization, which was first introduced by (Cai et al., 2019; Wang et al., 2020; Liu et al., 2019):

$$u^0_{w^t}(s) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \cdot \mathbf{1}\left\{\left(w_i^0\right)^T(s) > 0\right\}\left(w_i^t\right)^T(s),$$

$$f^0_{\theta^t}(s,a) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} b_i \cdot \mathbf{1}\left\{\left(\theta_i^0\right)^T(s,a) > 0\right\}\left(\theta_i^t\right)^T(s,a).$$

The following lemma characterizes the error induced by the above local linearization.

Lemma 19 *For $w \in \mathcal{B}^0_{R_w}, \theta \in \mathcal{B}^0_{R_\theta}, s \in \mathcal{S}$, and $a \in \mathcal{A}$, we have*

$$\mathbb{E}_{init}\left[\left|f_\theta(s,a) - f^0_\theta(s,a)\right|\right] = \mathcal{O}\left(R_\theta^{6/5} m^{-1/10} \hat{R}_\theta^{2/5}\right), \tag{59}$$

$$\mathbb{E}_{init}\left[\left|u_w(s) - u^0_w(s)\right|\right] = \mathcal{O}\left(R_w^{6/5} m^{-1/10} \hat{R}_w^{2/5}\right). \tag{60}$$

Proof Given any pair of model parameters $\theta \in \mathcal{B}^0_{R_\theta}$ and $\theta' \in \mathcal{B}^0_{R_\theta}$,

$$\mathbf{1}\left\{\theta_i^T(s,a) > 0\right\} \neq \mathbf{1}\left\{\left(\theta_i'\right)^T(s,a) > 0\right\}$$

implies

$$\left|\left(\theta_i'\right)^T(s,a)\right| \leq \left|\theta_i^T(s,a) - \left(\theta_i'\right)^T(s,a)\right| \leq \left\|\theta_i - \theta_i'\right\|_2.$$

Consequently, we have

$$\begin{aligned}
\left|f_\theta(s,a) - f^0_\theta(s,a)\right| &= \frac{1}{\sqrt{m}}\left|\sum_{i=1}^{m} b_i \cdot \left(\mathbf{1}\left\{\theta_i^T(s,a) > 0\right\} - \mathbf{1}\left\{\left(\theta_i^0\right)^T(s,a) > 0\right\}\right) \cdot \theta_i^T(s,a)\right| \\
&\leq \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \left|\mathbf{1}\left\{\theta_i^T(s,a) > 0\right\} - \mathbf{1}\left\{\left(\theta_i^0\right)^T(s,a) > 0\right\}\right| \cdot \left|\theta_i^T(s,a)\right| \\
&\leq \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \mathbf{1}\left\{\left(\theta_i^0\right)^T(s,a) \leq \left\|\theta_i - \theta_i^0\right\|_2\right\} \cdot \left\|\theta_i - \theta_i^0\right\|_2.
\end{aligned} \tag{61}$$

Next, we analyze $\mathbb{E}_{\text{init}}\left[\left|f_\theta(s,a) - f^0_\theta(s,a)\right|\right]$ by examining two cases.

- Case 1: $G_1 = \frac{1}{\sqrt{m}} \sum_{i \in C_1} \left(\mathbf{1}\left\{\left(\theta_i^0\right)^T(s,a) \leq \left\|\theta_i - \theta_i^0\right\|_2\right\}\right) \cdot \left\|\theta_i - \theta_i^0\right\|_2$, where $C_1 = \left\{i \in [m] : \left\|\theta_i - \theta_i^0\right\|_2 \leq \Delta\right\}$ and $\Delta > 0$.

  Without loss of generality, we assume that parameters $\theta(0)$ are uniformly initialized from a circle with a radius of R_0. Then, the number of neurons lying in C_1 is approximately $\frac{m\Delta}{R_0}$. Thus, we have

$$G_1 \leq \frac{1}{\sqrt{m}} \sum_{i \in C_1} \left\|\theta_i - \theta_i^0\right\|_2 = \mathcal{O}\left(m^{1/2} \Delta^2 R_0^{-1}\right). \tag{62}$$

Note that the size of C_1 decreases as Δ decreases. Given a fixed m, a sufficiently small Δ exists that makes (62) negligible.

- Case 2: $G_2 = \frac{1}{\sqrt{m}} \sum_{i \in C_2} \left(\mathbf{1} \left\{ \left(\theta_i^0\right)^T (s,a) \leq \left\|\theta_i - \theta_i^0\right\|_2 \right\} \right) \cdot \left\|\theta_i - \theta_i^0\right\|_2$, where $C_2 = \left\{ i \in [m] : \left\|\theta_i - \theta_i^0\right\|_2 > \Delta \right\}$ and $\Delta > 0$.

  We have $\left\|\theta_i - \theta_i^0\right\|_2 / \left\|\theta_i^0\right\|_2 > \Delta / \hat{R}_\theta$. Consequently, there exits a constant $c \geq \hat{R}_\theta / \Delta$ such that for any layer $i \in C_2, a \in \mathcal{A}, s \in \mathcal{S}$, it holds that

$$\mathbf{1} \left\{ \left(\theta_i^0\right)^T (s,a) \leq \left\|\theta_i - \theta_i^0\right\|_2 \right\} \leq 1 \leq c \left\|\theta_i - \theta_i^0\right\|_2 / \left\|\theta_i^0\right\|_2 . \tag{63}$$

  Next, by the Cauchy-Schwarz inequality and $\left\|\theta - \theta^0\right\|_2 \leq R_\theta$, we have

$$G_2 \leq \frac{R_\theta}{\sqrt{m}} \sqrt{\sum_{i \in C_2} \mathbf{1} \left\{ \left(\theta_i^0\right)^T (s,a) \leq \left\|\theta_i - \theta_i^0\right\|_2 \right\}}. \tag{64}$$

  By (63) and taking expectation on both sides of (64), we can obtain

$$\begin{aligned}
\mathbb{E}_{\text{init}} [G_2] &\leq \frac{R_\theta}{\sqrt{m}} \mathbb{E}_{\text{init}} \left[\sqrt{\sum_{i \in C_2} \mathbf{1} \left\{ \left(\theta_i^0\right)^T (s,a) \leq \left\|\theta_i - \theta_i^0\right\|_2 \right\}} \right] \\
&\leq \frac{R_\theta}{\sqrt{m}} \sqrt{\mathbb{E}_{\text{init}} \left[\sum_{i \in C_2} \mathbf{1} \left\{ \left(\theta_i^0\right)^T (s,a) \leq \left\|\theta_i - \theta_i^0\right\|_2 \right\} \right]} \\
&\leq \frac{R_\theta}{\sqrt{m}} \sqrt{c \mathbb{E}_{\text{init}} \left[\sum_{i \in C_2} \left\|\theta_i - \theta_i^0\right\|_2 / \|\theta_i^0\|_2 \right]}.
\end{aligned}$$

  By the Cauchy-Schwarz inequality and Jensen's inequality, we have

$$\begin{aligned}
\mathbb{E}_{\text{init}} \left[\sum_{i \in C_2} \left\|\theta_i - \theta_i^0\right\|_2 / \|\theta_i^0\|_2 \right] &\leq \mathbb{E}_{\text{init}} \left[\left(\sum_{i \in C_2} \left\|\theta_i - \theta_i^0\right\|_2^2 \right)^{1/2} \cdot \left(\sum_{i \in C_2} \|\theta_i^0\|_2^{-2} \right)^{1/2} \right] \\
&\leq R_\theta \cdot \mathbb{E}_{\text{init}} \left[\sum_{i \in C_2} \|\theta_i^0\|_2^{-2} \right]^{1/2},
\end{aligned} \tag{65}$$

  where the second inequality follows from $\sum_{i=1}^m \left\|\theta_i - \theta_i^0\right\|_2^2 = \left\|\theta - \theta^0\right\|_2^2 \leq R_\theta^2$. Since $\mathbb{E}_{\text{init}} \left[\|\theta_i\|_2^{-2} \right] \leq \infty, \forall i \in [m]$ by the initialization scheme (18), we have that the RHS of (65) is $\mathcal{O}(R_\theta m^{1/2})$. Thus, we can obtain

$$\mathbb{E}_{\text{init}} [G_2] = \mathcal{O} \left(R_\theta^{3/2} m^{-1/4} \hat{R}_\theta^{1/2} \Delta^{-1/2} \right). \tag{66}$$

By (62) and (66), we can obtain

$$\mathbb{E}_{\text{init}}\left[\left|f_\theta(s,a)-f_\theta^0(s,a)\right|\right] \le \mathbb{E}_{\text{init}}\left[G_1+G_2\right] = \mathcal{O}\left(R_\theta^{3/2}m^{-1/4}\hat{R}_\theta^{1/2}\Delta^{-1/2}+m^{1/2}\Delta^2R_0^{-1}\right).$$

We further assume $m^{1/2}\Delta^2R_0^{-1}\le\varrho$, i.e., $\Delta\le\varrho^{1/2}m^{-1/4}R_0^{1/2}$, which implies that

$$\begin{aligned}\mathbb{E}_{\text{init}}\left[\left|f_\theta(s,a)-f_\theta^0(s,a)\right|\right] &= \mathcal{O}\left(R_\theta^{3/2}m^{-1/4}\hat{R}_\theta^{1/2}\Delta^{-1/2}+\varrho\right)\\ &= \mathcal{O}\left(R_\theta^{3/2}m^{-1/8}\hat{R}_\theta^{1/2}\varrho^{-1/4}R_0^{-1/4}+\varrho\right).\end{aligned}$$

Moreover, we assume that $R_\theta^{3/2}m^{-1/8}\hat{R}_\theta^{1/2}\varrho^{-1/4}R_0^{-1/4}\ge\varrho$, i.e., $\varrho\le R_\theta^{6/5}m^{-1/10}\hat{R}_\theta^{2/5}R_0^{-1/5}$, which gives

$$\mathbb{E}_{\text{init}}\left[\left|f_\theta(s,a)-f_\theta^0(s,a)\right|\right] = \mathcal{O}\left(R_\theta^{6/5}m^{-1/10}\hat{R}_\theta^{2/5}\right).$$

This completes the proof of (59), and the proof of (60) is similar. ■

The following lemma provides the upper bound of the difference between network outputs.

Lemma 20 *For state $s\in\mathcal{S}$, any pair of actions a and a', and model parameters $\vartheta,\vartheta'\in\mathcal{B}_{R_\vartheta}^0$, which is θ for the policy and w for the value function, we have*

$$\mathbb{E}_{init}\left[\left|u_\vartheta(s,a)-u_{\vartheta'}(s,a')\right|\right] = \mathcal{O}\left(R_\vartheta\right), \tag{67}$$

$$\mathbb{E}_{init}\left[\left|u_\vartheta(s,a)-u_{\vartheta'}(s,a)\right|\right] = \mathcal{O}\left(R_\vartheta\right). \tag{68}$$

Proof By Jensen's inequality, we have

$$\begin{aligned}&\mathbb{E}_{\text{init}}\left|\left[u_\vartheta(s,a)-u_{\vartheta'}(s,a')\right]\right|^2\\ \le&\frac{1}{m}\mathbb{E}_{\text{init}}\left[\left|\sum_i^m b_i\cdot\mathbf{1}\left\{\vartheta_i^T(s,a)>0\right\}\vartheta_i^T(s,a)-\sum_i^m b_i\cdot\mathbf{1}\left\{\left(\vartheta_i'\right)^T(s,a')>0\right\}\left(\vartheta_i'\right)^T(s,a')\right|^2\right].\end{aligned}$$

By the fact that $ab-cd=a(b-d)+d(a-c)$ and $(a+b)^2\le 2a^2+2b^2$, we have

$$\begin{aligned}&\mathbb{E}_{\text{init}}\left|\left[u_\vartheta(s,a)-u_{\vartheta'}(s,a')\right]\right|^2\\ \le&\frac{1}{m}\mathbb{E}_{\text{init}}\left[\left|\sum_i^m b_i\cdot\mathbf{1}\left\{\vartheta_i^T(s,a)>0\right\}\left(\vartheta_i^T(s,a)-\left(\vartheta_i'\right)^T(s,a')\right)\right.\right.\\ &\left.\left.+\sum_i^m b_i\cdot\left(\mathbf{1}\left\{\left(\vartheta_i'\right)^T(s,a')>0\right\}-\mathbf{1}\left\{\left(\vartheta_i'\right)^T(s,a')>0\right\}\right)\left(\vartheta_i'\right)^T(s,a')\right|^2\right]\\ \le&\frac{1}{m}\mathbb{E}_{\text{init}}\left[2\left|\sum_i^m b_i\cdot\mathbf{1}\left\{\vartheta_i^T(s,a)>0\right\}\left(\vartheta_i^T(s,a)-\left(\vartheta_i'\right)^T(s,a')\right)\right|^2\right.\\ &\left.+2\left|\sum_i^m b_i\cdot\left(\mathbf{1}\left\{\left(\vartheta_i'\right)^T(s,a')>0\right\}-\mathbf{1}\left\{\left(\vartheta_i'\right)^T(s,a')>0\right\}\right)\left(\vartheta_i'\right)^T(s,a')\right|^2\right].\end{aligned}$$

Furthermore, we have

$$\|\vartheta\|_2^2 \leq \left(\left\|\vartheta - \vartheta^0\right\|_2 + \left\|\vartheta^0\right\|_2\right)^2 \leq 2R_\vartheta^2 + 2\left\|\vartheta^0\right\|_2^2. \tag{69}$$

By (69), the initialization scheme (18), and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathbb{E}_{\text{init}}\left|\left[u_\vartheta(s,a) - u_{\vartheta'}(s,a')\right|\right]^2 &\leq 2\mathbb{E}_{\text{init}}\left[\left(\sum_i^m \left(\vartheta_i^T(s,a) - \left(\vartheta_i'\right)^T(s,a')\right)^2\right) + \left\|\vartheta'\right\|_2^2\right] \\
&\leq 2\mathbb{E}_{\text{init}}\left[\left(\sum_i^m 2\left(\vartheta_i^T(s,a)\right)^2 + 2\left(\left(\vartheta_i'\right)^T(s,a')\right)^2\right) + \left\|\vartheta'\right\|_2^2\right] \\
&\leq 6\mathbb{E}_{\text{init}}\left[\left\|\vartheta'\right\|_2^2\right] + 4\mathbb{E}_{\text{init}}\left[\|\vartheta\|_2^2\right] \\
&\leq 20R_\vartheta^2 + 20.
\end{aligned} \tag{70}$$

We can then complete the proof of (67) by taking the square root of both sides of (70). The proof for (68) is similar. ■

Next, we prove Lemma 12.

Proof Let $(t', s') = \arg\max_{t>0, s\in\mathcal{S}} \left|V^t(s) - \bar{V}^t(s)\right|$, which are dependent on the initialization w^0. By the triangle inequality and Lemma 19, we have

$$\begin{aligned}
\mathbb{E}_{\text{init}}\left[\bar{\varepsilon}_w\right] &= \mathbb{E}_{\text{init}}\left[\max_t \left\|V^t - \bar{V}^t\right\|\right] \\
&\leq \mathbb{E}_{\text{init}}\left[\left|V^{t'}(s') - V^{t',0}(s')\right| + \left|V^{t',0}(s') - \bar{V}^{t'}(s')\right|\right] \\
&= \mathbb{E}_{\text{init}}\left[\left|V^{t'}(s') - V^{t',0}(s')\right| + \left|\sum_{n=1}^N q_n V_n^{t',0}(s') - \sum_{n=1}^N q_n V_n^{t'}(s')\right|\right] \\
&= \mathcal{O}\left(R_w^{6/5} m^{-1/10} \hat{R}_w^{2/5}\right),
\end{aligned}$$

which completes the proof of (20).

Since the order of terms in the square $\left(\pi^t(a|s) - \sum_{n=1}^N q_n \pi_n^t(a|s)\right)^2$ does not affect the value, we define two sets $C_1 = \left\{t > 0, a \in \mathcal{A}, s \in \mathcal{S} : \pi^t(a|s) > \sum_{n=1}^N q_n \pi_n^t(a|s)\right\}$, and $C_2 = C_1^{\complement}$. Accordingly, by the fact that $1 - \frac{1}{x} \leq \ln x, \forall x > 0$, we have $\pi^t(a|s) - \sum_{n=1}^N q_n \pi_n^t(a|s) \leq \pi^t(a|s) \log \frac{\pi^t(a|s)}{\sum_{n=1}^N q_n \pi_n^t(a|s)}$ for all $(t, a, s) \in C_1$. By the Arithmetic Mean-Geometric Mean (AM-GM) inequality, we have

$$\begin{aligned}
\left(\pi^t(a|s) - \sum_{n=1}^N q_n \pi_n^t(a|s)\right)^2 &\leq \left(\pi^t(a|s) - \sum_{n=1}^N q_n \pi_n^t(a|s)\right) \pi^t(a|s) \left(\log \frac{\pi^t(a|s)}{\sum_{n=1}^N q_n \pi_n^t(a|s)}\right) \\
&\leq \left(\pi^t(a|s) - \sum_{n=1}^N q_n \pi_n^t(a|s)\right) \pi^t(a|s) \left(\sum_{n=1}^N q_n \log \frac{\pi^t(a|s)}{\pi_n^t(a|s)}\right) \\
&\leq \pi^t(a|s) \left(\sum_{n=1}^N q_n \log \frac{\pi^t(a|s)}{\pi_n^t(a|s)}\right).
\end{aligned}$$

For all $(t,a,s) \in C_2$, we have $\sum_{n=1}^N q_n \pi_n^t(a|s) - \pi^t(a|s) = \sum_{n=1}^N q_n \left(\pi_n^t(a|s) - \pi^t(a|s)\right) \leq \sum_{n=1}^N q_n \pi_n^t(a|s) \log \frac{\pi_n^t(a|s)}{\pi^t(a|s)}$ by the fact that $1 - \frac{1}{x} \leq \ln x, \forall x > 0$. Therefore, we have

$$\left(\pi^t(a|s) - \sum_{n=1}^N q_n \pi_n^t(a|s)\right)^2 \leq \sum_{n=1}^N q_n \pi_n^t(a|s) \log \frac{\pi_n^t(a|s)}{\pi^t(a|s)}.$$

Let $(t',s') = \arg\max_{t>0, s\in\mathcal{S}} \left\| \pi^t(\cdot|s) - \sum_{n=1}^N q_n \pi_n^t(\cdot|s) \right\|_2$, which is conditional on θ^0, we have

$$\bar{\varepsilon}_\theta = \left\| \pi^{t'}(\cdot|s') - \sum_{n=1}^N q_n \pi_n^{t'}(\cdot|s') \right\|_2 = \sqrt{\sum_{a\in\mathcal{A}} \left(\pi^{t'}(a|s') - \sum_{n=1}^N q_n \pi_n^{t'}(a|s')\right)^2}.$$

Let $\mathcal{A}_1 = \{a' \in \mathcal{A} : t', a', s' \in C_1\}$ and $\mathcal{A}_2 = \{a' \in \mathcal{A} : t', a', s' \in C_2\}$, we have

$$\begin{aligned}
\bar{\varepsilon}_\theta \leq & \left(\sum_{a\in\mathcal{A}_1} \sum_{n=1}^N q_n \pi^{t'}(a|s') \left(f_{\theta^{t'}}(s',a) - f_{\theta_n^{t'}}(s',a) + \log \frac{\sum_{a'\in\mathcal{A}} \exp\left(f_{\theta_n^t}(s',a')\right)}{\sum_{a'\in\mathcal{A}} \exp\left(f_{\theta^t}(s',a')\right)} \right) \right. \\
& \left. + \sum_{a\in\mathcal{A}_2} \sum_{n=1}^N q_n \pi_n^{t'}(a|s') \left(f_{\theta_n^{t'}}(s',a) - f_{\theta^{t'}}(s',a) + \log \frac{\sum_{a'\in\mathcal{A}} \exp\left(f_{\theta^t}(s',a')\right)}{\sum_{a'\in\mathcal{A}} \exp\left(f_{\theta_n^t}(s',a')\right)} \right) \right)^{1/2} \\
\leq & \left(\sum_{a\in\mathcal{A}} \sum_{n=1}^N q_n \max\left\{ \pi_n^{t'}(a|s'), \pi^{t'}(a|s') \right\} \left(\left| f_{\theta^{t'}}(s',a) - f_{\theta_n^{t'}}(s',a) \right| \right.\right. \\
& \left.\left. + \left| \log \frac{\sum_{a'\in\mathcal{A}} \exp\left(f_{\theta^t}(s',a')\right)}{\sum_{a'\in\mathcal{A}} \exp\left(f_{\theta_n^t}(s',a')\right)} \right| \right) \right)^{1/2}.
\end{aligned}$$

By taking the maximum over $a \in \mathcal{A}$, we have

$$\begin{aligned}
\bar{\varepsilon}_\theta \leq & \left(\max_{a\in\mathcal{A}, n\in[N]} \left(\left| f_{\theta^{t'}}(s',a) - f_{\theta_n^{t'}}(s',a) \right| + \left| \log \frac{\sum_{a'\in\mathcal{A}} \exp\left(f_{\theta^t}(s',a')\right)}{\sum_{a'\in\mathcal{A}} \exp\left(f_{\theta_n^t}(s',a')\right)} \right| \right) \cdot \right. \\
& \left. \sum_{a\in\mathcal{A}} \sum_{n=1}^N q_n \max\left\{ \pi_n^{t'}(a|s'), \pi^{t'}(a|s') \right\} \right)^{1/2} \\
\leq & \left(2 \max_{a\in\mathcal{A}, n\in[N]} \left(\left| f_{\theta^{t'}}(s',a) - f_{\theta_n^{t'}}(s',a) \right| + \left| \log \frac{\sum_{a'\in\mathcal{A}} \exp\left(f_{\theta^t}(s',a')\right)}{\sum_{a'\in\mathcal{A}} \exp\left(f_{\theta_n^t}(s',a')\right)} \right| \right) \right)^{1/2}.
\end{aligned}$$

By Jensen's inequality, we have

$$\mathbb{E}_{\text{init}}\left[\bar{\varepsilon}_\theta\right] \leq \mathbb{E}_{\text{init}} \left[2 \max_{a,n} \left| f_{\theta^{t'}}(s',a) - f_{\theta_n^{t'}}(s',a) \right| + 2 \left| \log \frac{\sum_{a'\in\mathcal{A}} \exp\left(f_{\theta^t}(s',a')\right)}{\sum_{a'\in\mathcal{A}} \exp\left(f_{\theta_n^t}(s',a')\right)} \right| \right]^{1/2}. \quad (71)$$

It remains to bound the two absolute terms on the RHS of (71). The first absolute term can be bounded by Lemma 20 as

$$\mathbb{E}_{\text{init}} \left[\max_{a\in\mathcal{A}, n\in[N]} \left| f_{\theta^{t'}}(s',a) - f_{\theta_n^{t'}}(s',a) \right| \right] = \mathcal{O}\left(R_\theta\right). \quad (72)$$

Next, we bound the second absolute term on the RHS of (71). By the log-sum inequality, the log-sum-exp trick, and Lemma 20, we can obtain

$$\log \sum_{a' \in \mathcal{A}} \exp\left(f_{\theta^t}(s,a')\right) - \log \sum_{a' \in \mathcal{A}} \exp\left(f_{\theta_n^t}(s,a')\right) \leq \max_{a \in \mathcal{A}} f_{\theta^t}(s,a) - \frac{1}{|\mathcal{A}|} \sum_{a'} \left(f_{\theta_n^t}(s,a')\right),$$

and

$$\log \sum_{a' \in \mathcal{A}} \exp\left(f_{\theta_n^t}(s,a')\right) - \log \sum_{a' \in \mathcal{A}} \exp\left(f_{\theta^t}(s,a')\right) \leq \max_{a \in \mathcal{A}} f_{\theta_n^t}(s,a) - \frac{1}{|\mathcal{A}|} \sum_{a'} \left(f_{\theta^t}(s,a')\right),$$

which indicates that

$$\mathbb{E}_{\text{init}}\left[\max_{a \in \mathcal{A}, n \in [N]} \left| \log \frac{\sum_{a' \in \mathcal{A}} \exp\left(f_{\theta^t}(s,a')\right)}{\sum_{a' \in \mathcal{A}} \exp\left(f_{\theta_n^t}(s,a')\right)} \right| \right] = \mathcal{O}\left(R_\theta\right). \tag{73}$$

By substituting (72) and (73) into (71), we can obtain

$$\mathbb{E}_{\text{init}}\left[\bar{\varepsilon}_\theta\right] = \mathcal{O}\left(R_\theta^{1/2}\right),$$

which completes the proof of (21). The proof of (22) is similar. ■

Appendix H. Additional Experiment Setting

Machines: We simulate the federated learning experiments (1 server and N devices) on a commodity machine with 16 Intel(R) Xeon(R) Gold 6348 CPU @ 2.60GHz. It took about 6 mins to finish one round of training, i.e., 50 hours to obtain 500 data points for Figure 2.

The hyperparameters for the algorithms on MountainCars, Hoppers, HongKongOSMs, and the general FRL setting are given in Table 1, Table 2, Table 3, and Table 4, respectively.

Hyperparameter	Algorithm			
	FedPOHCS	FedAvg	Power-of-Choice	GradientNorm
Learning Rate	0.001	0.005	0.001	0.001
Learning Rate Decay	0.98	0.98	0.98	0.98
Batch Size	128	128	128	128
Timestep per Iteration	2048	2048	2048	2048
Number of Epochs (E)	1	1	1	1
Discount Factor (γ)	0.99	0.99	0.99	0.99
Discount Factor for GAE	0.95	0.95	0.95	0.95
KL Target	0.003	0.003	0.003	0.003

Table 1: Hyperparameters for each algorithm on MountainCars.

Hyperparameter	Algorithm			
	FedPOHCS	FedAvg	Power-of-Choice	GradientNorm
Learning Rate	0.03	0.03	0.03	0.03
Learning Rate Decay	0.9	0.9	0.9	0.9
Batch Size	128	128	128	128
Timestep per Iteration	2048	2048	2048	2048
Number of Epochs (E)	1	1	1	1
Discount Factor (γ)	0.99	0.99	0.99	0.99
Discount Factor for GAE	0.95	0.95	0.95	0.95
KL Target	0.003	0.003	0.003	0.003

Table 2: Hyperparameters for each algorithm on Hoppers.

Hyperparameter	Algorithm			
	FedPOHCS	FedAvg	Power-of-Choice	GradientNorm
Learning Rate	0.0001	0.0001	0.0001	0.0001
Learning Rate Decay	0.98	0.98	0.98	0.98
Batch Size	128	128	128	128
Timestep per Iteration	2048	2048	2048	2048
Number of Epochs (E)	10	10	10	10
Discount Factor (γ)	0.99	0.99	0.99	0.99
Discount Factor for GAE	0.95	0.95	0.95	0.95
KL Target	0.0001	0.0001	0.0001	0.0001

Table 3: Hyperparameters for each algorithm on HongKongOSMs.

Environment	Setting			
	#Client (N)	#Candidate (d)	#Participant (K)	#Iteration (I)
MountainCars	60	18	6	5
Hoppers	60	18	6	20
HongKongOSMs	10	9	2	10

Table 4: General FRL Setting. Refer to Section 2.1 and Algorithm 3 for their definitions.

References

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *International Conference on Machine Learning*, volume 97, pages 322–332, 2019.

Dimitri Bertsekas. *Abstract dynamic programming*. Athena Scientific, 2022.

Dimitri Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

Dimitri P. Bertsekas. Approximate policy iteration: a survey and some new methods. *Journal of Control Theory and Applications*, 9(3):310–335, 2011.

Huang Bojun. Steady state analysis of episodic reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9335–9345, 2020.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Qi Cai, Zhuoran Yang, Jason D Lee, and Zhaoran Wang. Neural temporal-difference learning converges to global optima. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H. Chi. Top-k off-policy correction for a reinforce recommender system. In *ACM International Conference on Web Search and Data Mining*, page 456–464. Association for Computing Machinery, 2019.

Wenlin Chen, Samuel Horváth, and Peter Richtárik. Optimal client sampling for federated learning. *Transactions on Machine Learning Research*, 2022.

Kamil Ciosek and Shimon Whiteson. Expected policy gradients for reinforcement learning. *Journal of Machine Learning Research*, 21(52):1–51, 2020.

Rémi Coulom. *Reinforcement learning using neural networks, with applications to motor control*. PhD thesis, Institut National Polytechnique de Grenoble-INPG, 2002.

Xiaofeng Fan, Yining Ma, Zhongxiang Dai, Wei Jing, Cheston Tan, and Bryan Kian Hsiang Low. Fault-tolerant federated reinforcement learning with theoretical guarantee. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1007–1021, 2021.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, 2018.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Towards understanding biased client selection in federated learning. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *International Conference on Artificial Intelligence and Statistics*, volume 151, pages 10351–10375, 2022.

Xiaoqian Jiang, Miran Kim, Kristin Lauter, and Yongsoo Song. Secure outsourced matrix computation and application to neural networks. In *ACM SIGSAC conference on computer and communications security*, pages 1209–1222, 2018.

Hao Jin, Yang Peng, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Federated reinforcement learning with environment heterogeneity. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *International Conference on Artificial Intelligence and Statistics*, volume 151, pages 18–37, 2022.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2): 1–210, 2021.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, pages 267–274, 2002.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020.

Yijing Li, Xiaofeng Tao, Xuefei Zhang, Junjie Liu, and Jin Xu. Privacy-preserved federated learning for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):8423–8434, 2022.

Xinle Liang, Yang Liu, Tianjian Chen, Ming Liu, and Qiang Yang. *Federated Transfer Reinforcement Learning for Autonomous Driving*, pages 357–371. Springer International Publishing, Cham, 2023.

Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ouiame Marnissi, Hajar El Hammouti, and El Houcine Bergou. Client selection in federated learning based on gradients importance. In *AIP Conference Proceedings*, volume 3034, page 100005, 2024.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *International Conference on Artificial Intelligence and Statistics*, volume 54, pages 1273–1282, 2017.

Alberto Maria Metelli, Matteo Pirotta, Daniele Calandriello, and Marcello Restelli. Safe policy iteration: A monotonically improving approximate policy iteration approach. *Journal of Machine Learning Research*, 22(97):1–83, 2021.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *International Conference on Machine Learning*, volume 48, pages 1928–1937, 2016.

Thomas M. Moerland, Joost Broekens, Aske Plaat, and Catholijn M. Jonker. Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 16(1): 1–118, 2023.

Andrew William Moore. Efficient memory-based learning for robot control. Technical report, University of Cambridge, 1990.

Arun Nair, Praveen Srinivasan, Sam Blackwell, Cagdas Alcicek, Rory Fearon, Alessandro De Maria, Vedavyas Panneershelvam, Mustafa Suleyman, Charles Beattie, Stig Petersen, Shane Legg, Volodymyr Mnih, Koray Kavukcuoglu, and David Silver. Massively parallel methods for deep reinforcement learning. *arXiv preprint arXiv:1507.04296*, 2015.

Takayuki Nishio and Ryo Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *IEEE International Conference on Communications*, pages 1–7, 2019.

OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`, 2017.

Melkior Ornik and Ufuk Topcu. Learning and planning for time-varying mdps using maximum likelihood estimation. *Journal of Machine Learning Research*, 22(1):1656–1695, 2021.

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In Jennifer Dy and Andreas Krause, editors, *International Conference on Machine Learning*, volume 80, pages 4026–4035, 2018.

Matteo Pirotta, Marcello Restelli, Alessio Pecorino, and Daniele Calandriello. Safe policy iteration. In Sanjoy Dasgupta and David McAllester, editors, *International Conference on Machine Learning*, volume 28, pages 307–315, 2013.

Jiaju Qi, Qihao Zhou, Lei Lei, and Kan Zheng. Federated reinforcement learning: techniques, applications, and open challenges. *Intelligence & Robotics*, 1(1), 2021.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei, editors, *International Conference on Machine Learning*, volume 37, pages 1889–1897, 2015.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587): 484–489, 2016.

Lauren N Steimle, David L Kaufman, and Brian T Denton. Multi-model markov decision processes. *IISE Transactions*, 53(10):1124–1139, 2021.

Alexander L. Strehl and Michael L. Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Alexander L. Strehl, Lihong Li, and Michael L. Littman. Reinforcement learning in finite mdps: Pac analysis. *Journal of Machine Learning Research*, 10(84):2413–2444, 2009.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

John E. Vargas-Munoz, Shivangi Srivastava, Devis Tuia, and Alexandre X. Falcão. Openstreetmap: Challenges and opportunities in machine learning and remote sensing. *IEEE Geoscience and Remote Sensing Magazine*, 9(1):184–199, 2021.

Nino Vieillard, Olivier Pietquin, and Matthieu Geist. Deep conservative policy iteration. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 6070–6077, 2020.

Eugene Vinitsky, Aboudy Kreidieh, Luc Le Flem, Nishant Kheterpal, Kathy Jang, Cathy Wu, Fangyu Wu, Richard Liaw, Eric Liang, and Alexandre M. Bayen. Benchmarks for reinforcement learning in mixed-autonomy traffic. In *Conference on Robot Learning*, volume 87, pages 399–409, 2018.

Paul Wagner. A reinterpretation of the policy oscillation phenomenon in approximate policy iteration. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24, 2011.

Han Wang, Aritra Mitra, Hamed Hassani, George J. Pappas, and James Anderson. Federated TD learning with linear function approximation under environmental heterogeneity. *Transactions on Machine Learning Research*, 2024.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In *International Conference on Learning Representations*, 2020.

Jiin Woo, Gauri Joshi, and Yuejie Chi. The blessing of heterogeneity in federated q-learning: Linear speedup and beyond. *Journal of Machine Learning Research*, 26(26):1–85, 2025.

Yu Xianjia, Jorge Peña Queralta, Jukka Heikkonen, and Tomi Westerlund. Federated learning in robotic and autonomous systems. *Procedia Computer Science*, 191:135–142, 2021.

Zhijie Xie and Shenghui Song. Fedkl: Tackling data heterogeneity in federated reinforcement learning by penalizing kl divergence. *IEEE Journal on Selected Areas in Communications*, 41(4):1227–1242, 2023.

Chenyu Zhang, Han Wang, Aritra Mitra, and James Anderson. Finite-time analysis of on-policy heterogeneous federated reinforcement learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Representation Learning*, volume 2024, pages 501–545, 2024.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. *Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms*, pages 321–384. Springer International Publishing, Cham, 2021.

Doudou Zhou, Yufeng Zhang, Aaron Sonabend-W, Zhaoran Wang, Junwei Lu, and Tianxi Cai. Federated offline reinforcement learning. *Journal of the American Statistical Association*, 119(548):3152–3163, 2024.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, volume 8, pages 1433–1438, 2008.