Finite-Time Analysis of Heterogeneous Federated Temporal Difference Learning

Ye Zhu , **Xiaowen Gong** and **Shiwen Mao**
Department of Electrical and Computer Engineering, Auburn University
{yzz0211, xzg0017, smao}@auburn.edu

Abstract

Federated Temporal Difference (FTD) learning has emerged as a promising framework for collaboratively evaluating policies without sharing raw data. Despite its potential, existing approaches often yield biased convergence results due to the inherent challenges of federated reinforcement learning, such as multiple local updates and environment heterogeneity. In response, we investigate federated temporal difference (TD) learning, focusing on collaborative policy evaluation with linear function approximation among agents operating in heterogeneous environments. We devise a heterogeneous federated temporal difference (HFTD) algorithm which iteratively aggregates agents' local stochastic gradients for TD learning. The HFTD algorithm involves two major novel contributions: 1) it aims to find the optimal value function model for the mixture environment which is the environment randomly drawn from agents' heterogeneous environments, using the local gradients of agents' mean squared Bellman errors (MSBEs) for their respective environments; 2) it allows agents to perform different numbers of local iterations for TD learning based on their heterogeneous computational capabilities. We analyze the finite-time convergence of the HFTD algorithm for the scenarios of IID sampling and Markovian sampling respectively. By characterizing bounds on the convergence error, we show that the HFTD algorithm can exactly converge to the optimal model and also achieves linear speedups as the number of agents increases.

1 Introduction

Federated Reinforcement Learning (FRL) has been proposed as a compelling approach, exploiting substantial computation capabilities of ubiquitous smart devices. As federated supervised learning which has been widely studied, FRL meets similar challenges, including data (environment) heterogeneity and system (computation and communication) heterogeneity. Moreover, there are some unique challenges in FRL due to its salient features. In particular, FRL involves optimizing value functions or policies that depend on environment-specific dynamics. This introduces additional complexity due to the recursive nature of the Bellman equations and the potential inconsistency of learning objectives across agents.

In contrast to the performance guarantees established in parallel reinforcement learning, where agents independently interact with identical environments, [Khodadadian *et al.*, 2022; Liu and Olshevsky, 2023; Fan *et al.*, 2021], studies on FRL with heterogeneous environments are quite limited due to the challenges mentioned above. While FRL across heterogeneous environments has been explored in several recent studies [Jin *et al.*, 2022; Wang *et al.*, 2023; Zhang *et al.*, 2024], none have established asymptotically vanishing convergence bounds for their algorithms (i.e., be made arbitrarily small by appropriately tuning hyperparameters such as the number of communication rounds and the step size). Moreover, existing works on FRL typically assume that all agents perform the *same* number of local updates in each communication round. However, in practice, agents often have heterogeneous computational capabilities, and the presence of stragglers can significantly impede overall training efficiency. In contrast, our work establishes the first asymptotically vanishing convergence bounds while explicitly accounting for such heterogeneity in computational capabilities across agents.

To tackle these challenges, we focus on federated temporal difference (TD) learning for policy evaluation with linear function approximation, where agents interact with heterogeneous environments modeled as Markov Decision Processes (MDPs) and collaboratively learn the value function for a given policy. These MDPs share identical state and action spaces but differ in their transition probability kernels and reward functions. We further allow each agent to perform a variable number of local updates per communication round, capturing both *environmental* and *computational* heterogeneity. Given this setting, we aim to address the following fundamental questions: (1) Is it possible to design a federated TD algorithm that *asymptotically converges* to the optimal value function? (2) If so, what is the *sample complexity* of such an algorithm?

We highlight the main contributions of this paper as follows.

- We study federated TD learning with linear function approximation, where multiple agents collaboratively per-

References	Heterogeneous Environments	Target Environment[1]	Markovian Sampling	Heterogeneous Local Iteration Numbers	Vanishing Convergence Error	Linear speedup
[Liu and Olshevsky, 2023]	×	Individual	×	×	√	√
[Khodadadian *et al.*, 2022]	×	Individual	√	×	√	√
[Wang *et al.*, 2023]	√	Virtual	√	×	×	√
[Mangold *et al.*, 2024]	√	Mixture	√	×	√	√
This paper	√	Mixture	√	√	√	√

Table 1: Comparison of settings and results for federated temporal difference learning

form policy evaluation via TD learning while interacting with heterogeneous environments and operating under heterogeneous computation configurations. We propose a *Heterogeneous Federated Temporal Difference* (HFTD) learning algorithm, which iteratively aggregates agents' local stochastic gradients for TD learning. Compared to existing FRL work, HFTD incorporates two key innovations: 1) The algorithm targets the optimal value function model for an environment *randomly drawn* from the distribution of agents' heterogeneous environments. To this end, it minimizes the *average* of agents' *mean squared Bellman errors* (MSBEs) using stochastic gradients specific to their individual environments. 2) The algorithm allows each agent to perform a different number of local TD updates in each communication round.

- We derive the finite-time convergence error bounds of the HFTD algorithm under both i.i.d. and Markovian sampling. Our results show that HFTD can asymptotically generate the *optimal* value function model as the number of communication rounds tends to infinity, under appropriate step size conditions. To our best knowledge, this is the *first* result in existing works on federated RL with heterogeneous environments that the convergence error can diminish to zero asymptotically. In particular, a key property of the global gradient of the average MSBE allows us to remove a *non-vanishing bias* in the convergence analysis, so that only vanishing error terms are left. We also show that the HFTD algorithm achieves a sample complexity of $\mathcal{O}\left(\frac{1}{N\epsilon}\right)$ and linear convergence speedup, which match the results of existing TD learning algorithms [Bhandari *et al.*, 2018; Khodadadian *et al.*, 2022].

2 Related Work

Temporal Difference Learning. Most existing works on TD learning have focused on the case of a single agent. For TD learning under IID sampling, the asymptotic convergence has been well studied in [Borkar and Meyn, 2000; Borkar, 2009], and the non-asymptotic convergence (i.e., finite-time analysis) has been studied in [Kamal, 2010; Dalal *et al.*, 2018]. For TD learning under Markovian sampling, the asymptotic convergence has been studied in [Tsitsiklis and Van Roy, 1996], and the non-asymptotic analysis has been studied in [Bhandari *et al.*, 2018; Srikant and Ying, 2019; Xu *et al.*, 2020b].

Distributed Reinforcement Learning. Distributed reinforcement learning (DRL) considers multiple agents operating in a distributed fashion. As a major class of DRL, *parallel* RL uses multiple learners to solve a large-scale single-agent RL task [Mnih *et al.*, 2016; Li and Schuurmans, 2011; Nair *et al.*, 2015], where the learners aim to learn a common policy for different instances of the *same* environment. Another major class of DRL is multi-agent reinforcement learning (MARL), where a group of agents operate in a common environment; all agents' actions influence the global state transition and MARL aims at seeking the optimal policy combining all local policies in a collaborative manner [Zeng *et al.*, 2022; Zhang *et al.*, 2018], or find local optimal policies in a non-collaborative manner [Zhang *et al.*, 2021]. These prior works of DRL are different from FRL, since 1) agents in FRL can mainly interact with heterogeneous environments and collaboratively learn a common policy in different environments; 2) FRL involves some unique features of FL [McMahan and Ramage, 2017; Bonawitz *et al.*, 2019; Stich, 2019; Li *et al.*, 2020; Wang and Ji, 2022; Guo *et al.*, 2022; Huang *et al.*, 2022; Karimireddy *et al.*, 2020], including multiple local iterations of agents, heterogeneous and time-varying computation capabilities of agents.

Federated Reinforcement Learning. The settings of FRL have significant differences from those of federated supervised learning, due to the salient features of RL, including online data sampling (especially Markovian sampling), and dynamic state transition.

Some recent works have studied FRL with *heterogeneous* environments [Jin *et al.*, 2022; Wang *et al.*, 2023; Zhu and Gong, 2023]. However, *none* of the algorithms in these works have convergence guarantee with vanishing errors. In this paper, we show that the proposed HFTD algorithm can asymptotically converge to the *optimal* value function model. The most relevant related work [Mangold *et al.*, 2024] analyzes the linear convergence of the mean squared error in heterogeneous environments, which we refer to as mixture environments. The key difference lies in our discovery: there exists an environment where the mean squared error equals the averaged mean squared error across all environments, whereas [Mangold *et al.*, 2024] employs control variates to address client drift. Furthermore, this paper allows agents to perform *heterogeneous* numbers of local iterations. We have detailed the technical distinctions in the convergence analysis of the HFTD algorithm in Section 5.3.

[1] See the definition of the target environment in Section 4.

3 Preliminaries

Policy Evaluation in a Single-Node Setting. We consider an infinite horizon Markov Decision Process with a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a transition kernel $\mathcal{P}$, a reward function $\mathcal{R}$, and a discount factor γ. We consider the problem of evaluating the value function V_μ of a given policy μ. Then for a initial state s, $\mathcal{P}_\mu(s' \,|s)$ shows the probability of transitioning form s to state s' under policy μ and $\mathcal{R}_\mu(s) = \sum_s \mathcal{P}_\mu(s' \,|s)\mathcal{R}(s,s')$ is the expected instantaneous reward. For policy μ, the expected cumulative rewards can be represented as a function of initial state s:

$$V_\mu(s) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R_\mu(s_t)\,|s_0 = s\right],$$

where $\{s_t\}$ is the sequence of states generated by the transition kernel $\mathcal{P}_\mu$. The value function satisfies the Bellman equation $T_\mu V_\mu = V_\mu$, where for any $V \in \mathbb{R}^{|S|}$,

$$(T_\mu V)(s) = R_\mu(s) + \gamma\sum_{s'} P_\mu(s,s')V(s'). \tag{1}$$

TD Learning with Linear Function Approximation. To mitigate the effect of intractable computation in face of large state spaces in policy evaluation, a common and tractable approach is to utilize linear function approximator for a representation of value functions. Let $\{\Phi_k\}_{k=1}^d$ be a set of d linearly independent basis vectors in $\mathbb{R}^n$, then the true value function V_μ is approximated as $V_\mu(s) \approx V_\theta(s) = \phi(s)^{\mathrm{T}}\theta$, where $\phi(s) \in \mathbb{R}^d$ is a fixed feature vector for state s and $\theta \in \mathbb{R}^d$ is the unknown model to be learned. For an observed tuple $O_t = (s_t, r_t, s_{t+1})$ at time t, the negative gradient of the Bellman error evaluated at the current parameter θ_t [Bhandari *et al.*, 2018] can be expressed as

$$g_t(\theta_t) = (r_t + \gamma\phi(s_{t+1})^{\mathrm{T}}\theta_t - \phi(s_t)^{\mathrm{T}}\theta_t)\phi(s_t). \tag{2}$$

Then the estimated model at time $t+1$ can be updated by the gradient descent method [Bhandari *et al.*, 2018] with step size $\alpha \in (0,1)$ as

$$\theta_{t+1} = \theta_t + \alpha g_t(\theta_t). \tag{3}$$

When state s_t sampled in tuple O_t follows the stationary distribution of the MDP, the expected negative gradient at θ is

$$\begin{aligned}\bar{g}_t(\theta) = \sum_{s_t,s_{t+1}} &\pi(s_t)P(s_{t+1}\mid s_t)\Big(R(s_t,s_{t+1}) \\ &+ \gamma\phi(s_{t+1})^\top\theta - \phi(s_t)^\top\theta\Big)\phi(s_t)\end{aligned} \tag{4}$$

where $\pi(\cdot)$ is the stationary distribution of the associated Markov chain which is assumed to be irreducible and aperiodic. Let D denote the diagonal matrix whose elements consist of the entries of $\pi(\cdot)$. In the convergence analysis of TD(0), [Tsitsiklis and Van Roy, 1996] has proved the limit point θ^* is the unique solution to the projected Bellman equation $\Phi\theta = \Pi_D T_\mu \Phi\theta$ with $\bar{g}(\theta^*) = 0$, where $\Pi_D(\cdot)$ is the projection operator defined on the subspace $\{\Phi x \,\big|x \in \mathbb{R}^d\}$.

Although the gradient steps $g_t(\theta)$ do not correspond to minimizing any fixed objective, it has been studied in the light of the stability of a dynamical system described by an ordinary difference equation (ODE). First we rewrite the stochastic gradient as $g_t(\theta) = A(O_t)\theta + b(O_t)$, where $A(O_t) = \phi(s)(\gamma\phi(s')^{\mathrm{T}} - \phi(s)^{\mathrm{T}})$, $b(O_t) = r(s)\phi(s)$. We define $\bar{A} = \mathbb{E}_\pi[A(O_t)]$ and $\bar{b} = \mathbb{E}_\pi[b(O_t)]$, then the expected negative gradient (4) can be established as $\bar{g}(\theta) = \bar{A}\theta + \bar{b}$, corresponding to the following ODE:

$$\dot{\theta} = \bar{g}(\theta) = \bar{A}\theta + \bar{b} \tag{5}$$

Under mild technical assumptions, it was shown in [Tsitsiklis and Van Roy, 1996] and [Sun *et al.*, 2020] that ODE as (5) admits a globally, asymptotically stable equilibrium point θ^* where $\theta^* = -\bar{A}^{-1}\bar{b}$ when the matrix $\bar{A}$ is non-singular.

4 Heterogeneous Federated Temporal Difference Learning with Linear Function Approximation

In this section, we first describe the problem statement about the federated policy evaluation in heterogeneous environments where agents collectively seek to find a global model to approximate the value function under a given policy μ. As we discussed above, in the process of policy evaluation for a single agent i, the local loss function F_i is usually defined as expected Bellman error squared [Bhandari *et al.*, 2018; Srikant and Ying, 2019]. Accordingly, the optimization problem for federated value evaluation can be formulated as

$$\min_{\theta\in\mathbb{R}^d}\left[F(\theta) = \frac{1}{N}\sum_{i=1}^{N} F_i(\theta)\right] \tag{6}$$

where $F_i(\theta) = \mathbb{E}_{O_t\sim d_i}\left[\frac{1}{2}(r_t + \gamma V_\theta(s_{t+1}) - V_\theta(s_t))^2\right]$ is the local objective function of i-th agent, i.e., the expected squared Bellman error with respect to the model parameter θ. Here, d_i is the stationary distribution of the state transition Markov chain in i-th environment, and $O_t^i = (s_t^i, r_t^i, s_{t+1}^i)$ represents a data sample from the environment i. We assume that each agent collects samples by interacting with its own environment independently. The MDP of agent i can be expressed by: $\mathcal{M}_i \triangleq \{\mathcal{S},\mathcal{A},\mathcal{P}_i,\mathcal{R}_i,\gamma\}$. We assume that all agents have the same state space and action space while the reward functions and transition probability functions may differ across various agents.

Comparisons with Virtual Environment. Prior works on FRL with heterogeneous environments [Jin *et al.*, 2022; Wang *et al.*, 2023] considered the objective of optimizing the value function model for a *single virtual* environment. That is to say, agents cooperate to build a model measured by a virtual environment (as a target environment). This virtual environment is constructed as an MDP $\bar{\mathcal{M}} \triangleq \{\mathcal{S},\mathcal{A},\bar{\mathcal{P}},\bar{\mathcal{R}},\gamma\}$ by directly averaging the transition kernels and reward functions of each agent's environment, given by $\bar{\mathcal{P}} = (1/N)\sum_{i=1}^N \mathcal{P}_i$ and $\bar{\mathcal{R}} = (1/N)\sum_{i=1}^N \mathcal{R}_i$. However, such an "averaged" environment may not coincide with any agent's individual environment. Intuitively, from the perspective of an individual agent, the objective function may not encourage more agents to participate in the federation. Motivated by this observation, in this paper, we consider a *mixture* environment defined as the environment *randomly drawn* from agents' heterogeneous environments. Note that this mixture environment is different from the virtual environment defined in [Jin *et al.*, 2022; Wang *et al.*, 2023]. Thus our goal is to find the optimal value function model for this mixture environment of agents. This goal is *similar* in spirit to that of federated supervised learning, since the latter aims to find the optimal model that minimizes the average training loss for all data samples of all clients.

Towards this goal, the global objective function of our federated TD learning problem is the *average MSBE* of all agents for their respective environments, as the MSBE quantifies the error of a value function model for an environment. To minimize the average MSBE, we devise the HFTD algorithm which updates the value function model via federated TD learning. The algorithm aims to iteratively estimate the gradient of the global objective function (i.e., the average MSBE), given by

$$\bar{g}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \bar{g}_i(\theta) \tag{7}$$

where $\bar{g}_i(\theta)$ is the gradient of agent i's local objective function (i.e., the MSBE for agent i's environment). Note that the condition (7) is substantially different from the estimation in [Wang *et al.*, 2023] (7) does not hold in [Wang *et al.*, 2023]), and is also a key property that allow us to show that our HFTD algorithm can asymptotically converge to the optimal model θ^* (rather than to a neighborhood of θ^* where the radius of the convergence error is determined by some non-vanishing bias error as in [Wang *et al.*, 2023], see Section 5.3 for detailed discussions of the technical differences). The optimal value function model θ^* that minimizes the average MSBE of agents satisfies $\bar{g}(\theta^*) = 0$. Note that the gradient in TD learning is different from that of the standard gradient descent, as $\bar{g}_i(\theta)$ or $\bar{g}(\theta)$ is not the gradient of any *fixed* objective function. To estimate the gradient $g(\theta)$, the HFTD algorithm computes a stochastic gradient $g(\theta_t)$ given by $g(\theta_t) = \frac{1}{N}\sum_{i=1}^{N} g_i(\theta_t)$, where $g_i(\theta_t)$ is the stochastic gradient of $g_i(\theta)$. Note that $g_i(\theta)$ is the expectation of stochastic gradient $g_i(\theta_t)$ following the stationary distribution of environment i.

Details of HFTD The detailed design of the HFTD algorithm is described as below (as summarized in Algorithm 1). In each round $t \in \{1, \ldots, T\}$, the central server first broadcasts the global model θ_t to all agents and each agent $i \in \{1, \ldots, N\}$ independently performs K_t^i local iterations starting from the current global model $\theta_i^{t,0}$ to optimize its local objective. K_t^i may vary across agents since agents have *heterogeneous* computation capabilities. Following the same policy μ, agent i observes the tuple $O_{t,k}^i = (s_{t,k}^i, r_{t,k}^i, s_{t,k+1}^i)$ at each local iteration k of the round t which is generated by its own MDP characterized by $\{\mathcal{S}, \mathcal{A}, \mathcal{P}_i, \mathcal{R}_i, \gamma\}$. Using the observation $O_{t,k}^i$, agent i can compute the stochastic gradient by (2) and update its local model. At the end of each round, agents send the gradients directly to the server. The server then aggregates the gradients, updates the global model and starts round $t+1$ of federation. Note that no exchange of raw samples is required, hence the privacy of local environments can be well protected. The update rule can be expressed as

$$\theta_{t+1} = \Pi_{2,\mathcal{H}} \left(\theta_t + \alpha \left(\frac{1}{N} \sum_{i=1}^{N} K_t^i \right) \cdot \frac{1}{N} \sum_{i=1}^{N} d_t^i \right) \tag{8}$$

where d_t^i is the normalized stochastic gradient in the t-th round at agent i as $d_i^t = \frac{1}{K_t^i} \sum_{k=0}^{K_t^i - 1} g_i(\theta_{t,k}^i)$; K_t^i is the number of local updates in round t at agent i. Here we consider agents have heterogeneous number of local updates while the number of local updates are identical and fixed in [Khodadadian *et al.*, 2022; Wang *et al.*, 2023; Jin *et al.*, 2022]. Note that cumulative local gradients are normalized when averaging, and this is a necessary technique when dealing with heterogeneous number of local updates in analysis [refer to Section 5.3 for details]. Besides, we use $\Pi_{2,\mathcal{H}}(\cdot)$ to denote the standard Euclidean projection on to a convex compact subset $\mathcal{H} \subset \mathbb{R}^d$ that is assumed to contain θ^*. Such a projection step is commonly adopted in RL [Bhandari *et al.*, 2018; Doan *et al.*, 2019] which ensures that the global models do not blow up. Note that the subset does not need to contain each θ_i^*.

Algorithm 1 Heterogeneous Federated TD (HFTD) Learning

```
Input: number of rounds $T$, step size $\alpha$, initial model $\theta_0$
for $t = 1$ to $T$ do
    $\theta_{t,0}^i \leftarrow \theta_t$ for all agents $i$
    for each agent $i = 1, 2, ..., N$ do
        for $k = 0$ to $K_t^i - 1$ do
            Observe a tuple $O_{t,k}^i = (s_{t,k}^i, r_{t,k}^i, s_{t,k+1}^i)$ and calculate the gradient by (2)
            Update the local model by (3)
        end for
    end for
    Agents send the normalized gradient $d_t^i = \frac{\theta_{t,K_t^i-1}^i - \theta_{t,0}^i}{K_t^i}$ to the server
    Server computes the global model by (8)
end for
Output: $\{\theta_t\}_{t=1}^T$
```

5 Theoretical Analysis of HFTD

5.1 IID Sampling

First, we start from the scenario where the random tuples are independently and identically sampled from the stationary distribution π_i of the Markov reward process for each agent i. That is to say, samples for updating the local model are independently drawn across iterations and rounds for each agent. We make the following assumptions, which are commonly imposed in federated reinforcement learning [Khodadadian *et al.*, 2022; Wang *et al.*, 2023; Fan *et al.*, 2021].

Assumption 1. (Bounded Gradient Variance) *For each agent i, there is a constant σ such that $\mathbb{E}\left\|g_i(\theta)-\bar{g}_i(\theta)\right\| \leq \sigma^2$ for all $\theta \in \mathbb{R}^d$.*

Assumption 2. (Exploration) *For each agent i, the matrix $\bar{A}_i$ defined in (5) is negative definite and its maximum eigenvalue can be bounded by $-\lambda$.*

Assumption 2 is commonly employed in analyzing TD learning with linear function approximation.

Assumption 3. (Bounded Gradient Heterogeneity) *For any set of weights satisfying convex combination, i.e., $\{p_i \geq 0\}_{i=1}^N$ and $\sum_{i=1}^N p_i = 1$, there exist constants $\beta^2 \geq 1$, $\kappa^2 \geq 0$ such that $\sum_i p_i \left\|\bar{g}_i(\theta)\right\|_2^2 \leq \beta^2 \left\|\sum_i p_i \bar{g}_i(\theta)\right\|_2^2 + \kappa^2$. If agents execute in identical environments, then $\beta^2 = 1$, $\kappa^2 = 0$.*

Assumption 3 is commonly used in the federated learning literature to capture the dissimilarities of local objectives [Wang *et al.*, 2020; Yang *et al.*, 2024]. In this work, it is necessary for getting rid of $\theta_i^* \in \mathcal{H}$ [Wang *et al.*, 2023]. We use Assumption 3 to measure the heterogeneity of environments instead of the heterogeneity of $\mathcal{P}_i$ and $\mathcal{R}_i$ used in [Jin *et al.*, 2022; Wang *et al.*, 2023; Zhang *et al.*, 2024]. Due to Lemma 2, we can establish an unbiased model. Now we are ready to present the convergence guarantees using the HFTD algorithm:

Theorem 1. (**HFTD with IID Sampling**) *Under Assumptions 1, 2 and 3, let $K_M = \max_{i,t}\{K_t^i\}$, $K_{\max} = \max_t\{K_t\}$ and $K_{\min} = \min_t\{K_t\}$. If $\alpha \leq \min_t\{\alpha_t\}$, Then output of Algorithm 1 can be represented as*

$$\mathbb{E}\left\|\theta_t-\theta^*\right\|_2^2 \leq e^{-\frac{\lambda\alpha K_{\min}T}{8}}\left\|\theta_0-\theta^*\right\|_2^2 + \frac{16\alpha K_{\max}^2\sigma^2\hat{K}_{\max}}{N\lambda K_{\min}} + \frac{256\alpha^2 K_{\max}\left(\sigma^2(K_{\max}-1)+2\kappa^2 K_M(K_M-1)\right)}{\lambda^2 K_{\min}}, \tag{9}$$

where $\hat{K}_t^{-1} = \frac{1}{N}\sum_{i=1}^N \frac{1}{K_t^i}$ and $\hat{K}_{\max}^{-1} = \max_t\{\hat{K}_t^{-1}\}$.

Remark 1. Theorem 1 provides a bound on the convergence error of the HFTD algorithm when agents operates in heterogeneous environments using heterogeneous local iterations. The error bound consists of three terms. As $\alpha > 0$, the 1st term converges to zero as T increases. Moreover, it achieves an exponential decay rate which matches the results of existing RL algorithms [Bhandari *et al.*, 2018; Xu *et al.*, 2020a; Kumar *et al.*, 2023]. The last two terms are all caused by the variances of stochastic gradients.

Remark 2. We note that the 2rd term shrinks at rate $\frac{1}{N}$ as N increases. Also note that the 3rd term becomes zero when each agent's local iteration number K_t^i is 1 (i.e., perfect synchronization for all agents). Moreover, if agents interact with the same environments ($\kappa^2 = 0$), the second part of the 3rd term of the convergence bound vanishes.

Corollary 1. *Suppose a constant local update number K for each agent, the convergence rate of HFTD with IID sampling is:*

$$\mathbb{E}\left\|\bar{\theta}^T-\theta^*\right\|_2^2 \leq e^{-\frac{\lambda\alpha KT}{8}}\left\|\bar{\theta}^0-\theta^*\right\|_2^2 + \frac{16\sigma^2\alpha}{N\lambda} + \frac{256\alpha^2\left(\sigma^2(K-1)+2\kappa^2 K(K-1)\right)}{\lambda^2} \tag{10}$$

Remark 3. When the communication rounds T is sufficiently large, then the convergence of HFTD will be dominated by the second term. Then we can conclude that the total complexity which can achieve an ϵ-accurate optimal solution $\mathbb{E}\left\|\theta_t-\theta^*\right\|_2^2 \leq \epsilon$ is $KT = \mathcal{O}\left(\frac{1}{N\epsilon}\right)$. When $K = 1$ and $N = 1$, the sample complexity will match the results in Theorem 2(b) in [Bhandari *et al.*, 2018].

5.2 Markovian Sampling

The case of IID sampling for RL can be hard to achieve in practical scenario. A more realistic setting is Markovian sampling, where the observed tuples used by TD are gathered from a single trajectory of the Markov chain. Different from the setting of IID sampling, Markovian sampling brings more challenges since samples are highly correlated. Specifically, in IID case, $\mathbb{E}[g(\theta)-\bar{g}(\theta)] = 0$ since $g(\theta)$ is the unbiased estimate of $\bar{g}(\theta)$. However, in the Markovian setting, the samples for calculating $g(\theta)$ are not sampled from the stationary distribution. To put it another way, θ and the sample observed at time t, O_t, are not independent. Hence, $\mathbb{E}[g(\theta)-\bar{g}(\theta)] \neq 0$, indicating bias exists in the gradient evaluation for the analysis of a single agent. Federated temporal learning introduces more intricate time correlations, which complicate theoretical analysis.

In the following analysis, we first introduce the geometric mixing property of finite-state, aperiodic and irreducible Markov chains as follows.

$$\sup\left\|\mathcal{P}_i(x_k \in \cdot|x_0)-\pi_i(\cdot)\right\|_{TV} \leq \eta_i\rho_i^k \tag{11}$$

where $\pi_i(\cdot)$ is the stationary distribution of the MDP i; $\eta_i > 0$ and $\rho_i \in [0,1]$ for all $i \in [N]$.

Assumption 4. (**Irreducibility and Aperiodicity**) *For each $i \in [N]$, the Markov chain induced by policy μ, corresponding to the state transition matrix $\mathcal{P}_i$, is aperiodic and irreducible.*

Theorem 2. (**HFTD with Markovian Sampling**) *Under Assumptions 2, 3, and 4, if we choose $\alpha \leq \min_t\{\alpha_t\}$, then the output of Algorithm 1 satisfies*

$$\mathbb{E}\left\|\theta_T-\theta^*\right\|_2^2 \leq e^{-\frac{\lambda\alpha K_{\min}T}{4}}\mathbb{E}\left\|\bar{\theta}_0-\theta^*\right\|_2^2 + C_1\alpha^3 + C_2\alpha^2 + C_3\frac{\alpha}{N} + C_4\alpha \tag{12}$$

where λ, C_1, C_2, C_3, and C_4 are positive, problem-dependent constants, with their detailed definitions provided in the supplementary material. Note that when $K_t^i = 1$ for all i and t, C_1 and C_2 will be zero. Only C_4 depends on the level of heterogeneity.

Remark 4. Theorem 2 characterizes the convergence of the HFTD algorithm where each agent's sampling follows a Markov chain. As in the setting of IID sampling, we can make similar observations here on the sampling complexity and the impacts of various system parameters on the convergence error. Specifically, the last four terms are quadratically or linearly amplified by K. This requires a sufficiently small α to mitigate the variance between two communication rounds. Different from IID setting, we note that the fourth term comes from the variance reduction in the Markovian setting. If t is sufficiently large, it will diminish to zero. This is because the Markov chain geometrically converges to its stationary distribution as t evolves.

Corollary 2. *If $N = 1$ and $K = 1$, then we have:*

$$\mathbb{E}\left\|\theta_t - \theta^*\right\|_2^2 \leq e^{-\frac{\lambda\alpha T}{4}} \left\|\bar{\theta}_0 - \theta^*\right\|_2^2 \\ + 2\alpha\lambda^{-1}\left(2c'H^2 + q'H + 4\beta^2\tau^2 cH^2 + 4\left(3 + 2\tau^2\right)\left[q'\right]^2\right)$$

Then we will clarify the differences of Markovian sampling and IID sampling in results and make comparisons with prior works in FRL.

Results: Markovian Vs. IID Sampling. Although IID sampling is a special case of Markovian sampling in single-agent reinforcement learning, in FRL, theoretical results under the Markovian sampling setting do not necessarily generalize to those under IID sampling due to differences in assumptions and convergence analyses [Khodadadian *et al.*, 2022; Wang *et al.*, 2023].

Comparison with Prior Works in FRL. Our proposed method, `HFTD`, improves upon existing results in federated reinforcement learning (FRL) with heterogeneous environments in terms of convergence. Specifically, Theorem 2 in [Wang *et al.*, 2023] only guarantees inexact convergence to a suboptimal solution, with the accuracy depending on the level of heterogeneity among the N agents. In contrast, `HFTD` provably converges to the exact optimal solution of the target problem. In addition, prior works such as [Khodadadian *et al.*, 2022; Liu and Olshevsky, 2023; Wang *et al.*, 2023] focus exclusively on homogeneous environments and homogeneous local iteration numbers, and thus do not address the challenges arising from heterogeneity.

5.3 Technical Differences of Convergence Analysis

In this subsection, we will highlight the key technical differences in the convergence analysis of HFTD (i.e., the proofs of Theorems 1 and 2), compared to prior works.

Similar to the convergence analysis of federated temporal difference learning [Khodadadian *et al.*, 2022; Wang *et al.*, 2023; Zhang *et al.*, 2024], the contraction property of the Bellman equation is utilized to produce a descent direction for the critic error. The informal decomposition can be expressed as:

$$\mathbb{E}\left\|\theta_{t+1} - \theta^*\right\| \leq \text{recursion} + \text{descent direction} \\ + \text{client drift} + \text{gradient variance} + \text{gradient norm}.$$

In the convergence analysis of this paper, in order to bound $\mathbb{E}\left\|\theta_{t+1} - \theta^*\right\|$, we need to bound an inner product term, which can be decomposed into three terms. As the objective of the HFTD algorithm is to minimize the average MSBE, the term B can be *canceled* (after the double summation before the inner product) due to the condition (7). In contrast, in [Wang *et al.*, 2023], this term B cannot be cancelled and becomes a *non-vanishing bias* in the convergence error.

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{K_t^i}\sum_{k=0}^{K_t^i-1}\mathbb{E}\left\langle \bar{g}_i(\theta_{t,k}^i), \theta_t - \theta^*\right\rangle \\ = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K_t^i}\sum_{k=0}^{K_t^i-1}\mathbb{E}\Big\langle \theta_t - \theta, \\ \bar{g}_i(\theta_{t,k}^i) - \bar{g}_i(\theta_t) + \underbrace{\bar{g}_i(\theta_t) - \bar{g}(\theta_t)}_{\text{B}} + \underbrace{\bar{g}(\theta_t)}_{\text{Lemma 2}}\Big\rangle \quad (13)$$

Similarly, in the convergence analysis of the Markovian setting, when dealing with such an inner product term, the term B can also be canceled.

In the convergence analysis, we also need to bound the error between the total accumulated local gradients $\frac{1}{N}\sum_i \frac{1}{K_t^i}\sum_{k=0}^{K_t^i-1}\bar{g}_i(\theta_{t,k}^i)$ and the global gradient $\sum_i \bar{g}_i(\theta_t)$ (as in (14)). By *normalizing* each agent's accumulated local gradients with the agent's local iteration number K_t^i, we are allowed to decompose the error into the sum of multiple error terms $\left\|\bar{g}_i(\theta_{t,k}^i) - \bar{g}_i(\theta_t)\right\|$, each involving *only one* agent's local gradients. Then each of these error terms can be further bounded using the the smoothness condition of the local gradient as

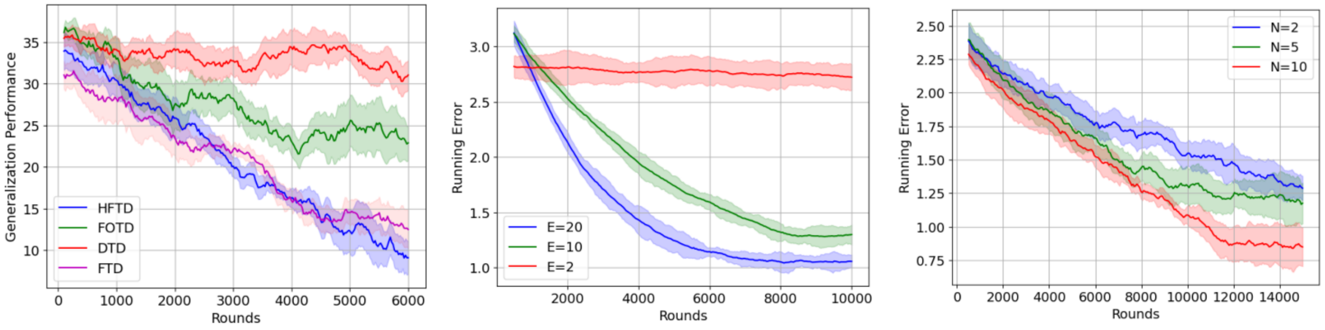
$$\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{1}{K_t^i}\sum_{k=0}^{K_t^i-1}\left(\bar{g}_i(\theta_{t,k}^i) - \bar{g}_i(\theta_t)\right)\right\|_2^2 \\ \leq \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|\frac{1}{K_t^i}\sum_{k=0}^{K_t^i-1}\left(\bar{g}_i(\theta_{t,k}^i) - \bar{g}_i(\theta_t)\right)\right\|_2^2 \\ \leq \frac{1}{N}\sum_{i=1}^{N}\frac{1}{K_t^i}\sum_{k=0}^{K_t^i-1}\mathbb{E}\left\|\bar{g}_i(\theta_{t,k}^i) - \bar{g}_i(\theta_t)\right\|_2^2. \quad (14)$$

6 Simulations

In this section, we present comprehensive experimental evaluations of HFTD on the RL task Gridword. We compare our proposed algorithm with the following baseline methods:

- *Federated On-policy Temporal Difference* (`FOTD`) [Khodadadian *et al.*, 2022], an FRL algorithm that combines `FedAvg` with TD;
- *Federated Temporal Difference* (`FTD`) [Wang *et al.*, 2023], an FRL algorithm conducting in heterogeneous environments;
- *Distributed Temporal Difference* (`DTD`) [Liu and Olshevsky, 2023], a distributed TD algorithm with almost no communication.

The experiments aim to demonstrate the efficacy of HFTD We provide numerical results under IID sampling setting and Markovian sampling setting on the platform. We first verify our theoretical results in a small-scale problem; see examples in [Sutton *et al.*, 1999]. Each experiment is conducted 10



(a) Comparisons with baselines

(b) Results on heterogeneity level

(c) Impact of number of agents

Figure 1: Training performance of HFTD with different settings

times. We plot the mean and standard deviation across the 10 runs.

In the simulations, the agent is initially placed in one corner of the maze and selects an action to move to the next cell with a certain probability. In the policy evaluation process, in order to avoid low learning efficiencies due to sparse rewards, the agent will receive a reward 0 if it reaches the desired goal and $-\frac{1}{2}\left[\nu_i(x-3)^2+\delta_i(y-3)^2\right]$ otherwise where (x, y) is the position of the agent and it is also the current state. Here the state space size is 16 and the action can be selected from up, down, left, right directions. The goal of the agents is to learn a common model to approximate the value function under the given policy.

Empirical results in Figure 1 reveal that the system parameters affect HFTD in different ways. Training performance refers to the running error between current model and optimal model. Generalization performance refers to the averaged performance in N environments with newly generated state-transitions [Jin *et al.*, 2022]. Fig. 1(a) shows that FTD and HFTD outperform the other two algorithms when facing heterogeneous environments.

To check the impact of environment heterogeneity on HFTD, we construct tasks of HFAC with various h, which controls how different the state transitions are. A higher h represents a larger heterogeneity level. Fig. 1(b) shows that, when we keep increasing h, the performance decreases. This result validates the theoretical results.

To verify the advantages due to the federation, we conduct the experiments on the impact of the number of agents of HFTD. As shown in Fig. 1(c), with a certain level of environment heterogeneity, increasing the number of participated agents accelerate the training. However, due to environmental heterogeneity, a gap to the optimal solution persists unless a smaller step size is used. This aligns with our theoretical insights and highlights the practical performance benefits of involving more agents.

Additionally, we conduct experiments to analyze how step sizes and the number of local iterations affect the convergence of HFTD. Due to space limitations in the main text, we provide brief explanations here, with detailed results presented in the supplementary material. From the experiments, we observe that increasing the number of local updates accelerates convergence. Moreover, while a larger step size results in faster convergence, it may cause fluctuations near the optimal solution. Consistent with Theorem 1, a smaller step size ensures that the convergence error approaches zero.

Compared with the simulation results of IID sampling, the learning process under Markovian sampling shows more instability, shown in the supplementary material. This is because in Markovian sampling, the next state depends only on the current state under a fixed policy so that some states are not visited enough. Hence, it is difficult to approximate the value function well for the entire state space.

7 Conclusion and Future Work

In this paper, we have developed a HFTD algorithm for federated TD learning with linear function approximation under environment heterogeneity and computation heterogeneity. We have shown that aggregated model using the HFTD algorithm can asymptotically converge to the optimal value function model, which is the first such result in existing works on FRL with heterogeneous environments. The HFTD algorithm also achieves sampling complexity of $\mathcal{O}\left(\frac{1}{N\epsilon}\right)$ and linear speedup that match the results of existing RL algorithms.

For future work, we will explore FRL algorithms that involve both policy evaluation and policy improvement, such as the actor-critic algorithms. Also, the assumptions used in this paper are common in existing theoretical studies on RL (including FRL). Indeed, some of these assumptions would not hold in practice for real-world applications. For example, linear function approximation (also a common assumption in many existing works) assumes that the true value function can be well approximated by a linear function, using a fixed feature vector $\phi(s)$ for state s. However, in real-world settings, it is hard to find a good approximate feature matrix for all states. Besides, we assume that each agent participates in FRL in each round in a synchronous manner. In real-world settings, it can be more efficient for agents to participate in some but not all rounds of FRL in an asynchronous manner. In future, we will also explore the problem of this paper in more general settings by relaxing some of the assumptions used in this paper, such as using non-linear function approximation, and considering partial and asynchronous participation of agents.

Acknowledgments

This work was supported by U.S. NSF under Grants ECCS-2121215 and CNS-2319342.

References

[Bhandari *et al.*, 2018] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory (COLT)*, 2018.

[Bonawitz *et al.*, 2019] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, Brendan McMahan, et al. Towards federated learning at scale: System design. In *Proceedings of Machine Learning and Systems*, pages 374–388, 2019.

[Borkar and Meyn, 2000] Vivek S Borkar and Sean P Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

[Borkar, 2009] Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

[Dalal *et al.*, 2018] Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for td (0) with function approximation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[Doan *et al.*, 2019] Thinh Doan, Siva Maguluri, and Justin Romberg. Finite-time analysis of distributed td (0) with linear function approximation on multi-agent reinforcement learning. In *International Conference on Machine Learning (ICML)*, pages 1626–1635, 2019.

[Fan *et al.*, 2021] Xiaofeng Fan, Yining Ma, Zhongxiang Dai, Wei Jing, Cheston Tan, and Bryan Kian Hsiang Low. Fault-tolerant federated reinforcement learning with theoretical guarantee. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1007–1021, 2021.

[Guo *et al.*, 2022] Yuanxiong Guo, Ying Sun, Rui Hu, and Yanmin Gong. Hybrid local SGD for federated learning with heterogeneous communications. In *International Conference on Learning Representations (ICLR)*, 2022.

[Huang *et al.*, 2022] Yan Huang, Ying Sun, Zehan Zhu, Changzhi Yan, and Jinming Xu. Tackling data heterogeneity: A new unified framework for decentralized sgd with sample-induced topology. In *International Conferene on Machine Learning (ICML)*, pages 9310–9345, 2022.

[Jin *et al.*, 2022] Hao Jin, Yang Peng, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Federated reinforcement learning with environment heterogeneity. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 18–37, 2022.

[Kamal, 2010] Sameer Kamal. On the convergence, lock-in probability, and sample complexity of stochastic approximation. *SIAM Journal on Control and Optimization*, 48(8):5178–5192, 2010.

[Karimireddy *et al.*, 2020] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning (ICML)*, pages 5132–5143, 2020.

[Khodadadian *et al.*, 2022] Sajad Khodadadian, Pranay Sharma, Gauri Joshi, and Siva Theja Maguluri. Federated reinforcement learning: Linear speedup under markovian sampling. In *International Conference on Machine Learning (ICML)*, pages 10997–11057, 2022.

[Kumar *et al.*, 2023] Harshat Kumar, Alec Koppel, and Alejandro Ribeiro. On the sample complexity of actor-critic method for reinforcement learning with function approximation. *Machine Learning*, 112(7):2433–2467, 2023.

[Li and Schuurmans, 2011] Yuxi Li and Dale Schuurmans. Mapreduce for parallel reinforcement learning. In *European Workshop on Reinforcement Learning*, pages 309–320. Springer, 2011.

[Li *et al.*, 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 (ICLR)*, 2020.

[Liu and Olshevsky, 2023] Rui Liu and Alex Olshevsky. Distributed td (0) with almost no communication. *IEEE Control Systems Letters*, 2023.

[Mangold *et al.*, 2024] Paul Mangold, Sergey Samsonov, Safwan Labbi, Ilya Levin, Reda Alami, Alexey Naumov, and Eric Moulines. Scafflsa: Taming heterogeneity in federated linear stochastic approximation and td learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[McMahan and Ramage, 2017] Brendan McMahan and Daniel Ramage. Federated learning: Collaborative machine learning without centralized training data. *Google Research Blog*, 3, 2017.

[Mnih *et al.*, 2016] 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 *International Conference on Machine Learning (ICML)*, pages 1928–1937, 2016.

[Nair *et al.*, 2015] Arun Nair, Praveen Srinivasan, Sam Blackwell, Cagdas Alcicek, Rory Fearon, Alessandro De Maria, Vedavyas Panneershelvam, Mustafa Suleyman, Charles Beattie, Stig Petersen, et al. Massively parallel methods for deep reinforcement learning. *arXiv preprint arXiv:1507.04296*, 2015.

[Srikant and Ying, 2019] Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation andtd learning. In *Conference on Learning Theory (COLT)*, 2019.

[Stich, 2019] Sebastian U Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations (ICLR)*, 2019.

[Sun *et al.*, 2020] Jun Sun, Gang Wang, Georgios B Giannakis, Qinmin Yang, and Zaiyue Yang. Finite-time analysis of decentralized temporal-difference learning with linear function approximation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[Sutton *et al.*, 1999] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1057–1063, 1999.

[Tsitsiklis and Van Roy, 1996] John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems (NeurIPS)*, 9, 1996.

[Wang and Ji, 2022] Shiqiang Wang and Mingyue Ji. A unified analysis of federated learning with arbitrary client participation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 19124–19137, 2022.

[Wang *et al.*, 2020] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[Wang *et al.*, 2023] Han Wang, Aritra Mitra, Hamed Hassani, George J Pappas, and James Anderson. Federated temporal difference learning with linear function approximation under environmental heterogeneity. *arXiv preprint arXiv:2302.02212*, 2023.

[Xu *et al.*, 2020a] Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[Xu *et al.*, 2020b] Tengyu Xu, Zhe Wang, Yi Zhou, and Yingbin Liang. Reanalysis of variance reduced temporal difference learning. In *International Conference on Learning Representations (ICLR)*, 2020.

[Yang *et al.*, 2024] Yifan Yang, Peiyao Xiao, and Kaiyi Ji. Simfbo: Towards simple, flexible and communication-efficient federated bilevel learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[Zeng *et al.*, 2022] Siliang Zeng, Tianyi Chen, Alfredo Garcia, and Mingyi Hong. Learning to coordinate in multi-agent systems: A coordinated actor-critic algorithm and finite-time guarantees. In *Learning for Dynamics and Control Conference*, pages 278–290, 2022.

[Zhang *et al.*, 2018] Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning (ICML)*, pages 5872–5881, 2018.

[Zhang *et al.*, 2021] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Decentralized multi-agent reinforcement learning with networked agents: Recent advances. *Frontiers of Information Technology & Electronic Engineering*, 22(6):802–814, 2021.

[Zhang *et al.*, 2024] Chenyu Zhang, Han Wang, Aritra Mitra, and James Anderson. Federated temporal difference learning with linear function approximation under environmental heterogeneity. In *International Conference on Learning Representations (ICLR)*, 2024.

[Zhu and Gong, 2023] Ye Zhu and Xiaowen Gong. Distributed policy gradient with heterogeneous computations for federated reinforcement learning. In *2023 57th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2023.