Towards Fast Rates for Federated and Multi-Task Reinforcement Learning

Feng Zhu, Robert W. Heath Jr., and Aritra Mitra

***Abstract*— We consider a setting involving N agents, where each agent interacts with an environment modeled as a Markov Decision Process (MDP). The agents' MDPs differ in their reward functions, capturing heterogeneous objectives/tasks. The collective goal of the agents is to communicate intermittently via a central server to find a policy that maximizes the average of long-term cumulative rewards across environments. The limited existing work on this topic either only provide asymptotic rates, or generate biased policies, or fail to establish any benefits of collaboration. In response, we propose `Fast-FedPG` - a novel federated policy gradient algorithm with a carefully designed bias-correction mechanism. Under a gradient-domination condition, we prove that our algorithm guarantees (i) fast linear convergence with exact gradients, and (ii) sub-linear rates that enjoy a linear speedup w.r.t. the number of agents with noisy, truncated policy gradients. Notably, in each case, the convergence is to a globally optimal policy with no heterogeneity-induced bias. In the absence of gradient-domination, we establish convergence to a first-order stationary point at a rate that continues to benefit from collaboration.**

I. Introduction

Despite the many successes of reinforcement learning (RL) in various applications (e.g., games, robotics, autonomous navigation, etc.), a large part of existing RL theory only provides asymptotic rates. Recently however, there has been a surge of interest in characterizing the finite-time behavior of model-free RL algorithms. For contemporary RL applications with massive state and action spaces, such finite-time analysis has revealed the need for lots of data samples to achieve desirable performance. Given this premise, it is natural to wonder if data collected from diverse environments can alleviate the sample-complexity bottleneck. This has prompted the emergence of a new paradigm called federated reinforcement learning (FRL), where agents interacting with potentially distinct environments collaborate with the hope of learning "good" policies with fewer samples than if they acted alone [1]. Unfortunately, existing FRL work either only provide empirical results [1], or make the unrealistic assumption of identical agent environments [2], or provide rates that exhibit a non-vanishing environmental-heterogeneity-induced bias term [3], [4]. In particular, such an additive bias term negates any potential statistical gains from collaboration. In this paper, we show for the first time that it is possible to achieve collaborative speedups in FRL even when data is collected from non-identical environments.

F. Zhu and A. Mitra are with the Dept. of Electrical and Computer Engineering, North Carolina State University. Email: {fzhu5, amitra2}@ncsu.edu. Robert W. Heath Jr. is with the Dept. of Electrical and Computer Engineering at the University of California, San Diego, USA. Email: rwheathjr@ucsd.edu. This material is based upon work supported in part by the National Science Foundation under Grant No. NSF-CCF-2225555.

Our model. We consider a sequential decision-making setting involving N agents, where each agent's environment is modeled as a Markov Decision Process (MDP). The agents' MDPs share the same state and action spaces, have identical probability transition maps, but differ in their reward functions; the non-identical reward functions help capture different goals/tasks across environments. The agents collaborate via a central server to learn a policy that can perform well in all environments by maximizing an average of the agents' long-term cumulative rewards. In this sense, our work is also related to multi-task RL, where data from different tasks is used to improve the performance on any given task [5]. As in the standard FL setting [6], to achieve communication-efficiency, the agents are allowed to communicate only once in every H iterations. Furthermore, to respect privacy, agents are not allowed to reveal their raw data (i.e., states, actions, and rewards). With this setup, we formulate a heterogeneous federated policy optimization problem. The only work we are aware of that have explored heterogeneity in the context of federated/decentralized policy gradient (PG) methods are [4], [7], [8]. While [7] only provides asymptotic rates, [4] and [8] fail to establish any provable benefits of collaboration. In this context, our **contributions** are as follows.

- **New algorithm**. We propose a novel federated PG algorithm called `Fast-FedPG` that, unlike standard "model-averaging" algorithms [2]–[4], [8], relies on a carefully constructed de-biasing/drift-mitigation mechanism using memory. Such a mechanism has not been explored earlier in FRL.
- **Key structural result**. To establish fast rates, we prove a simple, yet key structural result (Proposition 1) that relates the gradient of our objective function to the PG of an "average MDP" constructed from the agents' MDPs.
- **Fast rates and linear speedup**. Under a gradient-domination condition used to prove fast rates for centralized PG methods [9], [10], we prove that `Fast-FedPG` guarantees linear convergence to a globally optimal policy with exact gradients. With noisy, truncated policy gradients, we prove a rate of $\tilde{O}(1/(NHT))$ after T communication rounds, with H local PG steps within each round; see Theorem 2. Notably, our rates feature no heterogeneity-induced bias, and exhibit a clear N-fold speedup w.r.t. the number of agents, *thereby providing the first collaborative speedup result in FRL despite heterogeneity.* Finally, in Theorem 3, we show that in the absence of gradient-domination, `Fast-FedPG` guarantees convergence to a first-order stationary point at a rate of $\tilde{O}(1/\sqrt{NHT})$, i.e., with a $\sqrt{N}$-fold speedup.

II. Problem Formulation

We start by describing our RL setting, and then introduce the PG method to formulate our problem of interest.

RL setting. Our setting involves N agents, where each agent i interacts with an environment characterized by an MDP $\mathcal{M}_i = (\mathcal{S}, \mathcal{A}, R_i, \mathcal{P}, \gamma)$. Here, $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $R_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is a bounded reward function *specific to agent* i where $R_i(s, a)$ represents the immediate expected reward for taking action a in state s, $\mathcal{P}$ is a Markovian transition model where $\mathcal{P}(s'|s, a)$ represents the probability of transitioning from state s to s' under action a, and $\gamma \in [0, 1)$ is a discount factor. Therefore, agents share the same state and action spaces, are governed by the same probability transition maps, but have potentially different goals/objectives as captured by their unique reward functions. The distinction in the reward functions captures *heterogeneity* across the agents' environments.

The behavior of an agent is captured by a stochastic policy $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ is the space of probability distributions over $\mathcal{A}$. The dynamics of an agent-MDP interaction process unveils as follows. Starting from some initial state $s_i^{(0)}$, suppose an agent i interacts with its MDP $\mathcal{M}_i$ by playing a particular policy π. In particular, at each time-step $t = 0, 1, 2, \ldots$, the agent plays $a_i^{(t)} \sim \pi(\cdot|s_i^{(t)})$, observes an immediate reward $r_i^{(t)} = R_i(s_i^{(t)}, a_i^{(t)})$, and transitions to a new state $s_i^{(t+1)} \sim \mathcal{P}(\cdot|s_i^{(t)}, a_i^{(t)})$. This repeated interaction process generates a trajectory $\tau_i = \{(s_i^{(0)}, a_i^{(0)}, r_i^{(0)}), (s_i^{(1)}, a_i^{(1)}, r_i^{(1)}), \cdots\}$. In the single-agent RL setting, the typical goal of the agent i would be to find a policy π that maximizes a γ-discounted infinite-horizon expected cumulative reward, given by

$$J_i(\pi) \triangleq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_i^{(t)} \Big| s_i^{(0)} \sim \rho, \pi\right], \tag{1}$$

where ρ is an initial state distribution, and the expectation is taken w.r.t. the randomness in the initial state, the randomness induced by the stochastic policy π, and the randomness due to the state transitions prescribed by $\mathcal{P}$. For simplicity, we will assume throughout that all agents start from the same initial state distribution ρ. When the dynamics of the MDP are known, an optimal policy can be found using dynamic programming [11]. The learning aspect in our problem, however, stems from the fact that the reward functions $\{R_i\}_{i\in[N]}$ and state transition maps $\mathcal{P}$ are *unknown* to the agent. Given the fact that PG methods are easy to implement, we now describe the policy optimization approach for finding optimal policies that belong to a parameterized class.

Policy Gradient (PG) methods. Consider a class of parametric policies $\{\pi_\theta : \theta \in \mathbb{R}^d\}$, where π_θ is assumed to be differentiable w.r.t. θ. A common example of such a class is the *softmax policy*:

$$\pi_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathcal{A}} \exp(\theta_{s,a'})}, \tag{2}$$

where the parameter space is $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. For other common parametric classes (e.g., log-linear, neural softmax, etc.,), we refer the reader to [12]. Given a parameterized policy π_θ, let $J_i(\theta) \triangleq J_i(\pi_\theta)$ be agent i's local value-function associated with the parameter θ; here, $J_i(\cdot)$ is as defined in Eq. (1). PG methods operate by incrementally updating the parameter θ via gradient ascent on the value function.

Goal. Informally, we seek to find a policy π_θ that performs "well" on the set of environments $\{\mathcal{M}_i\}_{i\in[N]}$. This formulation is inspired by the federated supervised learning setting where agents with access to data from different distributions collaborate to find models with superior statistical performance relative to models trained with just individual agent-data. To formally set up our problem using the language of optimization, for each $i \in [N]$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, we reset $R_i(s, a) \leftarrow 1 - R_i(s, a)$, and interpret $R_i(\cdot, \cdot)$ as a *regret* function instead of a reward function. The collective goal of the agents then is to find a policy parameter $\theta^* \in \operatorname{argmin}_{\theta\in\mathbb{R}^d} J(\theta)$, where $J(\theta)$ is a global value-function defined as

$$J(\theta) \triangleq \frac{1}{N} \sum_{i=1}^{N} J_i(\theta). \tag{3}$$

To achieve the above objective within a federated framework, the agents can exchange information via a central server that coordinates the learning process. As in the FL setting, however, the agents need to adhere to stringent communication and privacy constraints, i.e., they are only allowed to communicate *intermittently*, and are required to keep their raw data (i.e., states, actions, and rewards) private. We now discuss the key challenges in the problem posed above.

• **Effect of reward-heterogeneity.** Since the agents have different reward functions, a locally optimal policy parameter $\theta_i^* \in \operatorname{argmin}_{\theta\in\mathbb{R}^d} J_i(\theta)$ for agent i may not coincide with the globally optimal parameter θ^*. Therefore, in the intermittent periods where the agents act locally to respect communication constraints, they will tend to drift towards their own locally optimal parameters. In this context, while *drift-mitigation* techniques have been explored for federated supervised learning, their effectiveness remains unclear in our RL setting.

• **Effect of non-convexity.** As shown in [12], the value-function $J_i(\theta)$ is non-convex w.r.t. θ for even direct and softmax parameterizations. This precludes the use of standard tools from convex optimization theory for our purpose, making it highly non-trivial, in particular, to guarantee convergence to the globally optimal parameter θ^* in our heterogeneous federated RL setting.

• **Effect of noise and truncation.** Policy gradients are typically *noisy* and *biased*. To see why, let us fix an agent $i \in [N]$, and note that based on the celebrated Policy Gradient Theorem [13], the ideal exact gradient $\nabla J_i(\theta)$ is given by

$$\nabla J_i(\theta) = \mathbb{E}_{\tau_i}\left[\sum_{t=0}^{\infty} \gamma^t r_i^{(t)} \sum_{k=0}^{\infty} \nabla_\theta \log \pi_\theta(a_i^{(k)}|s_i^{(k)})\right], \tag{4}$$

where the expectation is w.r.t. the random trajectory τ_i. There are two key issues that impede computing the exact gradient. First, computing the expectation in Eq. (4) would require averaging over all possible trajectories; this is infeasible.

Algorithm 1 `Fast-FedPG`

1: **Input:** Local step-size η, Global step-size α_g, Initial parameter $\bar{\theta}^{(0)} \in \mathbb{R}^d$, Initial global PG $\hat{\nabla}_K J(\bar{\theta}^{(0)})$.
2: **for** $t = 0, \ldots, T-1$ **do**
3: **for** $i = 1, \ldots, N$ **do**
4: Agent i initializes its local parameter $\theta_{i,0}^{(t)} = \bar{\theta}^{(t)}$.
5: **for** $\ell = 0, \ldots, H-1$ **do**
6: Agent i samples a truncated trajectory by playing policy $\pi_{\theta_{i,\ell}^{(t)}}$ on its MDP $\mathcal{M}_i$ over a horizon of length K. It then computes $\hat{\nabla}_K J_i(\theta_{i,\ell}^{(t)})$ as per Eq. (5).
7: Agent i updates local parameter as per Eq. (6).
8: **end for**
9: Agent i transmits $\Delta_{i,H}^{(t)} = \theta_{i,H}^{(t)} - \bar{\theta}^{(t)}$ to server.
10: **end for**
11: Server broadcasts $\bar{\theta}^{(t+1)}$ computed as per Eq. (7).
12: **for** $i = 1, \ldots, N$ **do**
13: Agent i transmits $\hat{\nabla}_K J_i(\bar{\theta}^{(t+1)})$ to server.
14: **end for**
15: Server broadcasts global PG $\hat{\nabla}_K J(\bar{\theta}^{(t+1)})$.
16: **end for**

Second, during implementation, agents do not have the luxury of rolling out/simulating a trajectory of infinite length. Therefore, complying with practice, each agent i computes an empirical estimate of $\nabla J_i(\theta)$ by sampling a truncated trajectory of length $K \in \mathbb{N}$: this is done by playing policy π_θ on MDP $\mathcal{M}_i$ over a finite roll-out horizon K. This leads to the following *noisy* and *biased* estimate of $\nabla J_i(\theta)$ that gets implemented in practice:

$$\hat{\nabla}_K J_i(\theta) = \sum_{t=0}^{K-1} \gamma^t r_i^{(t)} \sum_{k=0}^{K-1} \nabla_\theta \log \pi_\theta(a_i^{(k)}|s_i^{(k)}), \qquad (5)$$

where the noise arises due to sampling, and the bias due to truncation. For use later in the paper, let us also define the truncated gradient $\nabla_K J_i(\theta)$ as the expectation of the noisy truncated gradient, i.e., $\nabla_K J_i(\theta) \triangleq \mathbb{E}\left[\hat{\nabla}_K J_i(\theta)\right]$.

Desiderata. Despite the complex interplay between infrequent communication, client-drift effects due to reward heterogeneity, non-convex optimization landscapes, and inexact, truncated gradients, we seek to develop a federated PG method that (i) leads to *de-biased solutions*, i.e., guarantees convergence to θ^*, as opposed to θ_i^* for any $i \in [N]$; and (ii) achieves *near-optimal statistical rates* that clearly exhibit the benefit of collaboration among agents. In the next section, we will design such an algorithm.

III. Fast Federated Policy Gradient

In this section, we will develop our proposed algorithm called `Fast Federated Policy Gradient` (`Fast-FedPG`), formally described in Algorithm 1. The primary components of our algorithm involve *local policy gradient steps*, and a *de-biasing/drift mitigation strategy*. We now proceed to elaborate on these ideas.

Local policy gradient steps. The structure of `Fast-FedPG` mimics a typical FL algorithm: it operates in rounds $t = 0, 1, \ldots, T-1$, where within each round, every agent performs H local policy optimization steps in parallel by interacting with its own environment. During these local steps, there is no communication with the server. Let us denote by $\theta_{i,\ell}^{(t)}$ the policy parameter of agent i at the ℓ-th local step of the t-th communication round. At the beginning of each round t, $\theta_{i,0}^{(t)}$ is initialized from a common global policy parameter $\bar{\theta}^{(t)}$. To update $\theta_{i,\ell}^{(t)}$, agent i first samples a truncated trajectory of length K by playing the parameterized policy $\pi_{\theta_{i,\ell}^{(t)}}$ in its own MDP $\mathcal{M}_i$. Doing so enables agent i to compute the noisy truncated gradient $\hat{\nabla}_K J_i(\theta_{i,\ell}^{(t)})$ as per Eq. (5). The key question is: *How should agent i use $\hat{\nabla}_K J_i(\theta_{i,\ell}^{(t)})$ to update $\theta_{i,\ell}^{(t)}$?* Inspired by the popular FL algorithm `FedAvg` [6], one natural strategy could be to use the following update: $\theta_{i,\ell+1}^{(t)} = \theta_{i,\ell}^{(t)} - \eta \hat{\nabla}_K J_i(\theta_{i,\ell}^{(t)})$. Running this update for several local steps will however cause agent i to drift towards a locally optimal parameter θ_i^*. This bias is undesirable since our goal is to instead converge to θ^* - a minimizer of the global value function $J(\theta)$ in Eq. (3). We now describe our strategy for achieving this.

De-biasing/Drift mitigation. We start by observing that if the agents could in fact communicate at all time-steps, they would ideally like to implement the update rule: $\bar{\theta}^{(t+1)} = \bar{\theta}^{(t)} - \eta \hat{\nabla}_K J(\bar{\theta}^{(t)})$, where $\hat{\nabla}_K J(\theta) \triangleq (1/N) \sum_{i \in [N]} \hat{\nabla}_K J_i(\theta)$. This is not possible however, since an agent i cannot access the policy gradients of the other agents between communication rounds. The main idea behind our approach is to equip each agent with the memory of the global policy gradient direction $\hat{\nabla}_K J(\bar{\theta}^{(t)})$ from the beginning of the round. As an agent i keeps interacting with its own MDP $\mathcal{M}_i$, however, its local policy parameter $\theta_{i,\ell}^{(t)}$ evolves from its value $\bar{\theta}^{(t)}$ at the beginning of the round. To account for this staleness, agent i adds the correction term $\hat{\nabla}_K J_i(\theta_{i,\ell}^{(t)}) - \hat{\nabla}_K J_i(\bar{\theta}^{(t)})$ to the global PG guiding direction $\hat{\nabla}_K J(\bar{\theta}^{(t)})$. This leads to the update rule for `Fast-FedPG`:

$$\theta_{i,\ell+1}^{(t)} = \theta_{i,\ell}^{(t)} - \eta\Big(\hat{\nabla}_K J_i(\theta_{i,\ell}^{(t)}) - \hat{\nabla}_K J_i(\bar{\theta}^{(t)}) + \hat{\nabla}_K J(\bar{\theta}^{(t)})\Big). \quad (6)$$

At the end of H local steps, the agents transmit the change in their local parameters over the entire round to the server (line 9). The server then updates the global parameter as

$$\bar{\theta}^{(t+1)} = \bar{\theta}^{(t)} + \frac{\alpha_g}{N} \sum_{i=1}^{N} \Delta_{i,H}^{(t)}, \qquad (7)$$

where $\alpha_g \in (0, 1]$ is a global step-size. We note here that while drift-mitigation strategies similar to the one above have been studied in federated supervised learning [14], [15], it is unclear a priori whether they can yield fast rates for our RL setting. In particular, the lack of convexity and the use of noisy truncated policy gradients (in Eq. (6)) that are inherently biased leads to unique challenges in analyzing the dynamics of `Fast-FedPG`. Despite such challenges, we provide a rigorous convergence analysis of `Fast-FedPG` in this paper.

IV. Main Results

A. A key structural result

We start by establishing an important result that will serve as the key enabler for achieving fast convergence rates. To motivate the need for this result, we note that in the context of empirical risk minimization for supervised learning, one typically relies on strong-convexity of the loss function to achieve linear convergence rates. Despite the non-convexity of the policy optimization landscape, some recent work [9], [16] have shown that fast linear convergence to a globally optimal policy is still possible under a weaker (relative to strong-convexity) gradient-domination condition. This condition, however, depends on the policy parameterization and the properties of the *underlying MDP*. In our case, since we care about convergence to $\theta^* \in \operatorname{argmin}_{\theta\in\mathbb{R}^d} J(\theta)$, a gradient-domination condition on the global PG $\nabla J(\theta) \triangleq (1/N)\sum_{i\in[N]} \nabla J_i(\theta)$ is required to achieve linear convergence to θ^*. For this to happen, however, we need to link $\nabla J(\theta)$ to the policy gradient of some underlying MDP.

Given this reasoning, the subject of this section is to construct an "Average MDP" using the agents' MDPs, and establish that the PG of this average MDP is precisely equal to $\nabla J(\theta)$. Once this is achieved, a gradient-domination condition for the average MDP will immediately imply one for $\nabla J(\theta)$. With this in mind, we construct the average MDP as $\bar{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \bar{R}, P, \gamma)$, where $\bar{R}(s,a) \triangleq \frac{1}{N}\sum_{i=1}^N R_i(s,a), \forall (s,a) \in \mathcal{S}\times\mathcal{A}$. Similar to Eq. (1), we can define the value-function of this MDP for a policy π_θ as $\bar{J}(\theta) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t \bar{r}^{(t)} \middle| s^{(0)} \sim \rho, \pi_\theta\right]$, where $\bar{r}^{(t)} = \bar{R}(s^{(t)}, a^{(t)})$. We then claim the following.

Proposition 1. *For any fixed policy π_θ and initial distribution ρ, we have $\nabla \bar{J}(\theta) = \nabla J(\theta) = \frac{1}{N}\sum_{i=1}^N \nabla J_i(\theta)$, where $\nabla \bar{J}(\theta)$ is the gradient (w.r.t. θ) of the value-function $\bar{J}(\theta)$ corresponding to the average MDP $\bar{\mathcal{M}}$.*

Proof. We will prove this result in three steps by making some simple observations. To proceed, let us use the notation $\texttt{Avg}(\{c_i\}) \triangleq (1/N)\sum_{i\in[N]} c_i$ to denote the average of N scalars $c_1, \ldots, c_N$.

Step 1. Define $R_i^{\pi_\theta}(s) \triangleq \sum_{a\in\mathcal{A}} R_i(s,a)\pi_\theta(a|s)$. For any fixed policy π_θ, we then claim that $\bar{R}^{\pi_\theta}(s) = \texttt{Avg}(\{R_i^{\pi_\theta}(s)\}), \forall s \in \mathcal{S}$. In words, this simply states that the reward function induced by a policy π_θ on the average MDP $\bar{\mathcal{M}}$ is the average of the reward functions induced by the same policy on the agents' MDPs. To see this, observe:

$$\bar{R}^{\pi_\theta}(s) = \sum_{a\in\mathcal{A}} \bar{R}(s,a)\pi_\theta(a|s) = \sum_{a\in\mathcal{A}} \frac{1}{N}\sum_{i=1}^N R_i(s,a)\pi_\theta(a|s).$$

The claim then follows by swapping the order of the summation, and using the definition of $R_i^{\pi_\theta}(s)$. Before we present the next fact, with a slight overload of notation, let us use $J_i(\theta, s)$ to represent the value-function $J_i(\theta)$ when the initial state is $s \in \mathcal{S}$ deterministically. We can define $\bar{J}(\theta, s)$ accordingly. Next, define the state-action value function as $Q_i^{\pi_\theta}(s,a) = \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r_i^{(t)} \middle| s^{(0)} = s, a^{(0)} = a, \pi_\theta\right]$.

Step 2. For any fixed policy π_θ, we claim (i) $\bar{J}(\theta, s) = \texttt{Avg}(\{J_i(\theta, s)\}), \forall s \in \mathcal{S}$, and (ii) $\bar{Q}^{\pi_\theta}(s,a) = \texttt{Avg}(\{Q_i^{\pi_\theta}(s,a)\})$, where $\bar{J}(\theta,s)$ and $\bar{Q}^{\pi_\theta}(s,a)$ are the value-function and state-action value function induced by the policy π_θ on the average MDP $\bar{\mathcal{M}}$. To prove this claim, we will exploit the fact that the policy π_θ induces the same Markov transition matrix $\mathbf{P}^\theta$ on each MDP $\mathcal{M}_i, i \in [N]$, as well as on $\bar{\mathcal{M}}$, since they all share the same transition kernels. From the policy-specific Bellman fixed-point equation [11], we then have:

$$\mathbf{J}_i^\theta = (\mathbf{I} - \gamma\mathbf{P}^\theta)^{-1}\mathbf{R}_i^\theta, \forall i \in [N], \bar{\mathbf{J}}^\theta = (\mathbf{I} - \gamma\mathbf{P}^\theta)^{-1}\bar{\mathbf{R}}^\theta, \quad (8)$$

where we stacked up $R_i^{\pi_\theta}(s), \bar{R}^{\pi_\theta}(s), J_i(\theta, s)$, and $\bar{J}(\theta, s)$ for different states into the vectors $\mathbf{R}_i^\theta, \bar{\mathbf{R}}^\theta$, $\mathbf{J}_i^\theta$, and $\bar{\mathbf{J}}^\theta$. The claim that $\bar{J}(\theta,s) = \texttt{Avg}(\{J_i(\theta,s)\}), \forall s \in \mathcal{S}$, then immediately follows from Eq. (8) and Step 1. Next, observe

$$\begin{aligned}\bar{Q}^{\pi_\theta}(s,a) &= \bar{R}(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[\bar{J}(\theta,s')\right] \\ &\overset{(a)}{=} \frac{1}{N}\sum_{i=1}^N R_i(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[\frac{1}{N}\sum_{i=1}^N J_i(\theta,s')\right] \\ &= \frac{1}{N}\sum_{i=1}^N \Bigg(\underbrace{R_i(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}\left[J_i(\theta,s')\right]}_{Q_i^{\pi_\theta}(s,a)}\Bigg),\end{aligned}$$

where for (a), we used $\bar{J}(\theta,s) = \texttt{Avg}(\{J_i(\theta,s)\})$.

Step 3. To complete the last step, recall the definition of *state occupancy measure* from [17]:

$$d_{s^{(0)}}^{\pi_\theta}(s) = (1-\gamma)\sum_{t=0}^\infty \gamma^t \mathbb{P}(s_t = s | s^{(0)}, \pi_\theta),$$

where $\mathbb{P}(s_t = s|s^{(0)}, \pi)$ denotes the probability of starting from $s^{(0)}$ and ending up in s at round t by playing policy π_θ. From [17, Theorem 11.4], we then know that

$$\nabla J_i(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi_\theta}}\left[\sum_{a\in\mathcal{A}} \nabla\log\pi_\theta(a|s) Q_i^{\pi_\theta}(s,a)\pi_\theta(a|s)\right], \quad (9)$$

where $d_\rho^{\pi_\theta}(s) = \mathbb{E}_{s^{(0)}\sim\rho}\left[d_{s^{(0)}}^{\pi_\theta}(s)\right]$. For the average MDP $\bar{\mathcal{M}}_i$, $\nabla\bar{J}(\theta)$ can be computed exactly as in Eq. (9), with just $Q_i^{\pi_\theta}(s,a)$ replaced by $\bar{Q}^{\pi_\theta}(s,a)$. This is because identical transition kernels imply identical occupancy measures across the agents' MDPs and the average MDP. The claim in Proposition 1 then follows immediately from Step 2 where we showed that $\bar{Q}^{\pi_\theta}(s,a) = \texttt{Avg}(\{Q_i^{\pi_\theta}(s,a)\})$. □

B. Assumptions and main convergence results

To obtain our main results, we need to make a few standard assumptions that we state and describe below.

Assumption 1 (Smoothness)**.** *There exists a constant $L \geq 1$ such that for each agent $i \in [N]$, $J_i(\cdot)$ is L-smooth, i.e.,*

$$\|\nabla J_i(\theta_1) - \nabla J_i(\theta_2)\| \leq L\|\theta_1 - \theta_2\|, \forall \theta_1, \theta_2 \in \mathbb{R}^d,$$

where $\nabla J_i(\cdot)$ is the exact gradient of $J_i(\cdot)$ as defined in (4).

The smoothness of local objective functions immediately implies that of the global objective function, yielding:

$$\|\nabla J(\theta_1) - \nabla J(\theta_2)\| \leq L\|\theta_1 - \theta_2\|, \forall \theta_1, \theta_2 \in \mathbb{R}^d.$$

Almost all papers on PG methods we are aware of rely on smoothness [9], [10], [12], [18]. The next assumption follows directly from the definition of $\nabla_K J_i(\cdot)$.

Assumption 2 (Unbiasedness)**.** *For each agent $i \in [N]$, $\hat{\nabla}_K J_i(\cdot)$ is an unbiased estimate of $\nabla_K J_i(\cdot)$.*

Next, we make a bounded variance assumption that is typical in the literature on stochastic optimization.

Assumption 3 (Bounded variance)**.** *There exists a constant $\sigma \geq 1$ such that*

$$\mathbb{E}\left[\left\|\hat{\nabla}_K J_i(\theta) - \nabla_K J_i(\theta)\right\|^2\right] \leq \sigma^2, \forall i \in [N], \forall \theta \in \mathbb{R}^d.$$

The term σ captures the variance in the noisy gradients. Our next assumption will help to control the effect of truncating the gradients [10].

Assumption 4 (Truncation)**.** *There exists a constant $D \geq 1$ such that for each agent $i \in [N]$, the following bound holds:*

$$\|\nabla_K J_i(\theta) - \nabla J_i(\theta)\| \leq D\gamma^K, \forall \theta \in \mathbb{R}^d. \tag{10}$$

Finally, we will assume that the trajectories across agents are statistically independent, as is done in FRL [2]–[4].

Assumption 5 (Independence)**.** *We assume that the sampled trajectories $\tau_i, i \in [N]$ are independent across agents.*

Given the above assumptions, our first main result characterizes `Fast-FedPG`'s progress in each round.

Theorem 1. *Suppose Assumptions 1 - 5 hold. Define $\alpha = H\eta\alpha_g$ as the effective step-size. Then there exists a universal constant $C \geq 1$, such that with $\alpha_g = 1$ and η chosen to satisfy $\eta \leq 1/(4CLH)$, `Fast-FedPG` guarantees $\forall t \geq 0$:*

$$\mathbb{E}\left[J(\bar{\theta}^{(t+1)})\right] \leq \mathbb{E}\left[J(\bar{\theta}^{(t)})\right] - \frac{\alpha}{4}\mathbb{E}\left[\left\|\nabla J(\bar{\theta}^{(t)})\right\|^2\right] + \mathcal{O}\left(\frac{\alpha^2 L\sigma^2}{NH} + \alpha^3 L^2\sigma^2\right) + \mathcal{O}(\alpha) D^2\gamma^{2K}. \tag{11}$$

Due to space constraints, the detailed proofs of Theorem 1 and all our subsequent results are omitted here, but available in the extended version [19]. For now, let us see how Theorem 1 yields fast rates under gradient-domination.

Theorem 2. ***(Fast rates)*** *Suppose all the conditions in Theorem 1 hold. Additionally, suppose the following gradient-domination condition is satisfied by the average MDP:*

$$\mu(\bar{J}(\theta) - \bar{J}(\theta^*)) \leq \left\|\nabla \bar{J}(\theta)\right\|^2, \forall \theta \in \mathbb{R}^d, \tag{12}$$

for some $\mu > 0$. Then, `Fast-FedPG` guarantees $\forall T \geq 0$:

$$\mathbb{E}\left[J(\bar{\theta}^{(T)}) - J(\theta^*)\right] \leq \left(1 - \frac{\alpha\mu}{4}\right)^T \left(J(\bar{\theta}^{(0)}) - J(\theta^*)\right) + \mathcal{O}\left(\frac{\alpha L\sigma^2}{\mu NH} + \frac{\alpha^2 L^2\sigma^2}{\mu}\right) + \mathcal{O}\left(\frac{D^2\gamma^{2K}}{\mu}\right). \tag{13}$$

Proof. The statement and proof of Proposition 1 tell us that $\nabla \bar{J}(\theta) = \nabla J(\theta)$ and $\bar{J}(\theta) = J(\theta), \forall \theta \in \mathbb{R}^d$. Combining this with Eq. (12), we get $\mu(J(\theta) - J(\theta^*)) \leq \|\nabla J(\theta)\|^2, \forall \theta \in \mathbb{R}^d$. Plugging this bound into Eq. (11) and unrolling the resulting inequality leads to the desired claim. □

Discussion. To parse Theorem 2, we note that in the absence of noise (i.e., $\sigma = 0$) and truncation errors (i.e., $D = 0$), `Fast-FedPG` guarantees linear convergence of $J(\bar{\theta}^{(T)})$ to the globally optimal value $J(\theta^*)$. This is consistent with recent findings in the centralized PG literature [9], [10] that achieve similar linear rates under gradient-domination.

Linear speedup. We now discuss how under a suitable selection of the local step-size η, the number of communication rounds T, and the roll-out horizon K, one can achieve a linear speedup result from Theorem 2. To that end, suppose

$$\eta = \frac{4}{\mu H}\frac{\log(NHT)}{T}, \quad T \geq \frac{L}{\mu}\max\{16C\log(NHT), NH\}.$$

Note that T can always be chosen large enough to meet the above condition, and the above choices of η and T respect the criterion $\eta \leq 1/(4CLH)$ needed for Theorem 1 to hold. Next, let the roll-out horizon K be picked to satisfy: $K \geq \log(NHT)/(2\log(1/\gamma))$. Substituting the above choices of parameters into Eq. (13), and simplifying, we obtain:

$$\mathbb{E}\left[J(\bar{\theta}^{(T)}) - J(\theta^*)\right] \leq \tilde{O}\left(\left(G + \frac{L\sigma^2}{\mu^2} + \frac{D^2}{\mu}\right)\frac{1}{NHT}\right),$$

where $G = \left(J(\bar{\theta}^{(0)}) - J(\theta^*)\right)$. We note that despite noisy, biased policy gradients and reward-heterogeneity, `Fast-FedPG` guarantees convergence (in expectation) to a globally optimal policy parameter θ^* at the rate $\tilde{O}(1/(NHT))$. There are two important takeaways here. First, unlike [3] and [4], our final rate exhibits no heterogeneity-induced bias. Second, the $\tilde{O}(1/(NHT))$ rate is essentially the best one can hope for since the total amount of data (i.e., trajectories) across agents over T rounds is precisely NHT. Notably, our results clearly exhibit an N-fold speedup w.r.t. the number of agents (relative to the centralized setting), demonstrating the benefit of federation. These results are the first of their kind in the context of multi-task/federated policy gradients, and significantly improve upon those in [7] that only come with asymptotic rates, and those in [8] that exhibit no linear speedup.

Finally, suppose the gradient-domination condition no longer holds. Moreover, suppose the transition kernels across the agents are potentially non-identical. The proof of Theorem 1 in [19] reveals that Theorem 1 continues to hold. An immediate consequence of this result is the following guarantee on convergence to a first-order stationary point.

Theorem 3. *Suppose all the conditions in Theorem 1 hold. Then, `Fast-FedPG` guarantees:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla J(\bar{\theta}^{(t)})\right\|^2\right] \leq \frac{4\mathbb{E}\left[J(\bar{\theta}^{(0)}) - J(\bar{\theta}^{(T)})\right]}{\alpha T} + \mathcal{O}\left(\frac{\alpha L\sigma^2}{NH} + \alpha^2 L^2\sigma^2\right) + \mathcal{O}\left(D^2\gamma^{2K}\right).$$

With $\eta = \frac{4}{H}\sqrt{\frac{NH}{T}}$, $\quad T \geq L^2\max\{256C^2NH, N^3H^3\}$, and K chosen as before, we obtain a final convergence rate

of $\tilde{O}(1/\sqrt{NHT})$ in this case. Once again, there is a clear benefit of collaboration captured by the inverse scaling of this bound w.r.t. $\sqrt{N}$.

V. Analysis

The goal of this section is to provide a sketch of the proof of Theorem 1. Our first main step is to exploit smoothness of the local objective functions to establish a one-round progress bound for `Fast-FedPG`.

Lemma 1. *Suppose Assumptions 1 - 5 hold. Let* $\Delta_{i,\ell}^{(t)} = \theta_{i,\ell}^{(t)} - \bar{\theta}^{(t)}$. *Then, the following is true for* `Fast-FedPG`*:*

$$\mathbb{E}\left[J(\bar{\theta}^{(t+1)})\right] \leq \mathbb{E}\left[J(\bar{\theta}^{(t)})\right] - \frac{\alpha}{2}(1-8\alpha L)\,\mathbb{E}\left[\left\|\nabla J(\bar{\theta}^{(t)})\right\|^2\right] + \alpha L\left(\frac{L+4\alpha L^2}{NH}\right)\sum_{i=1}^{N}\sum_{\ell=0}^{H-1}\mathbb{E}\left[\left\|\Delta_{i,\ell}^{(t)}\right\|^2\right] + \left(\alpha+2\alpha^2 L\right)D^2\gamma^{2K} + \frac{2\alpha^2 L\sigma^2}{NH}.$$

The above lemma relates the progress made in a particular round t to the magnitude of the policy gradient $\left\|\nabla J(\bar{\theta}^{(t)})\right\|$. Notably, the progress is not controlled by the policy gradients of the agents' individual MDPs, but rather by the policy gradient of the global objective function. This is precisely what we want to ensure that progress is made towards θ^*, not θ_i^* for any agent i. The object that impedes the progress is the client-drift term $\sum_{i=1}^{N}\sum_{\ell=0}^{H-1}\mathbb{E}\left[\left\|\Delta_{i,\ell}^{(t)}\right\|^2\right]$. Therefore, to further refine the bound in Eq. (14), we need to control this drift effect. To that end, we have the following lemma.

Lemma 2. *Suppose Assumptions 1 - 4 hold. Let the local step-size* η *satisfy* $3\eta LH \leq 1$. *Then, the following holds for the expected client-drift* $\forall i \in [N], \forall \ell \in \{0,\cdots,H-1\}$*:*

$$\mathbb{E}\left[\left\|\Delta_{i,\ell}^{(t)}\right\|^2\right] \leq 32\eta^2H^2\underbrace{\left(\mathbb{E}\left[\left\|\nabla J(\bar{\theta}^{(t)})\right\|^2\right] + 18\sigma^2 + 18D^2\gamma^{2K}\right)}_{\mathcal{G}^{(t)}}.$$

To gain some intuition about the above result, suppose that there is no noise, i.e., $\sigma = 0$, and no truncation, i.e., $D = 0$. In other words, suppose all policy gradients are exact. Lemma 2 then tells us that the drift over the round t is caused due to an $O(\eta^2H^2\left\|\nabla J(\bar{\theta}^{(t)})\right\|^2)$ perturbation. We immediately observe that if $\bar{\theta}^{(t)} = \theta^*$, i.e., the parameter at the beginning of the round is where we eventually want it to be, then there will be no drift. This is again precisely what we desire, and aligns with the design strategy behind our algorithm `Fast-FedPG`.

To summarize the discussion, up to noise- and truncation-induced errors, the "good" term that contributes to progress is on the order of $\alpha\left\|\nabla J(\bar{\theta}^{(t)})\right\|^2$, while the "bad" term that impedes progress is $O(\eta^2H^2\left\|\nabla J(\bar{\theta}^{(t)})\right\|^2)$. Since the bad term is a higher-order term in the step-size, by tuning the local and global step-sizes appropriately, one can hope to achieve overall progress. Making the above informal argument precise takes quite a bit of work. The details of this analysis are available in [19].

VI. Conclusion

We studied the problem of finding an optimal policy that performs well on average across multiple heterogeneous environments. To find such an optimal policy, we formulated a federated policy optimization problem, and developed the first communication-efficient policy gradient algorithm that (i) achieves fast linear rates; (ii) provides a linear speedup in sample-complexity w.r.t. the number of agents; and (iii) incurs no heterogeneity-induced bias. As future work, we plan to study the problem of learning personalized policies in the context of multi-task RL.

References

[1] J. Qi, Q. Zhou, L. Lei, and K. Zheng, "Federated reinforcement learning: Techniques, applications, and open challenges," *arXiv:2108.11887*, 2021.

[2] S. Khodadadian, P. Sharma, G. Joshi, and S. T. Maguluri, "Federated reinforcement learning: Linear speedup under Markovian sampling," in *Int. Conf. on Machine Learning*. PMLR, 2022, pp. 10 997–11 057.

[3] H. Wang, A. Mitra, H. Hassani, G. J. Pappas, and J. Anderson, "Federated temporal difference learning with linear function approximation under environmental heterogeneity," *arXiv:2302.02212*, 2023.

[4] H. Jin, Y. Peng, W. Yang, S. Wang, and Z. Zhang, "Federated reinforcement learning with environment heterogeneity," in *International Conf. on Artificial Intelligence and Stat.* PMLR, 2022, pp. 18–37.

[5] S. Sodhani, A. Zhang, and J. Pineau, "Multi-task reinforcement learning with context-based representations," in *International Conference on Machine Learning*. PMLR, 2021, pp. 9767–9779.

[6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *AISTATS*. PMLR, 2017, pp. 1273–1282.

[7] Z. Xie and S. Song, "FedKL: Tackling data heterogeneity in federated reinforcement learning by penalizing KL divergence," *IEEE Journal on Selected Areas in Comm.*, vol. 41, no. 4, pp. 1227–1242, 2023.

[8] S. Zeng, M. A. Anwar, T. T. Doan, A. Raychowdhury, and J. Romberg, "A decentralized policy gradient approach to multi-task reinforcement learning," in *Uncertainty in Artificial Intelligence*. PMLR, 2021.

[9] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans, "On the global convergence rates of softmax policy gradient methods," in *Int. Conf. on Machine Learning*. PMLR, 2020, pp. 6820–6829.

[10] R. Yuan, R. M. Gower, and A. Lazaric, "A general sample complexity analysis of vanilla policy gradient," in *Int. Conf. on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 3332–3380.

[11] M. L. Puterman, "Markov decision processes," *Handbooks in Operations Research and Management Science*, vol. 2, pp. 331–434, 1990.

[12] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "On the theory of policy gradient methods: Optimality, approximation, and distribution shift," *Journal of Machine Learning Research*, vol. 22, no. 98, pp. 1–76, 2021.

[13] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in Neural Information Processing Systems*, vol. 12, 1999.

[14] A. Mitra, R. Jaafar, G. J. Pappas, and H. Hassani, "Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients," *Advances in Neural Information Processing Systems*, vol. 34, pp. 14 606–14 619, 2021.

[15] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5132–5143.

[16] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator," in *Int. Conf. on Machine Learning*. PMLR, 2018, pp. 1467–1476.

[17] A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun, "Reinforcement learning: Theory and algorithms," *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, vol. 32, 2019.

[18] L. Xiao, "On the convergence rates of policy gradient methods," *Journal of Machine Learning Research*, vol. 23, no. 282, pp. 1–36, 2022.

[19] F. Zhu, R. W. Heath, and A. Mitra, "Towards fast rates for federated and multi-task reinforcement learning," *arXiv preprint*, 2024.