NC STATE UNIVERSITY
Electrical and Computer Engineering

Towards Fast Rates for Federated and Multi-Task Reinforcement Learning

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

Motivation

- For contemporary RL applications with **massive state and action spaces**, algorithm training requires lots of data samples.
- Data samples typically come from different environments.

Question. Can we use data collected from **diverse environments** to **speed up** the training process?

Goals.

1. To learn a policy that can perform well in all environments.
2. To demonstrate collaborative *speedup* in the final result, i.e., multiple agents do help **expedite the learning**.

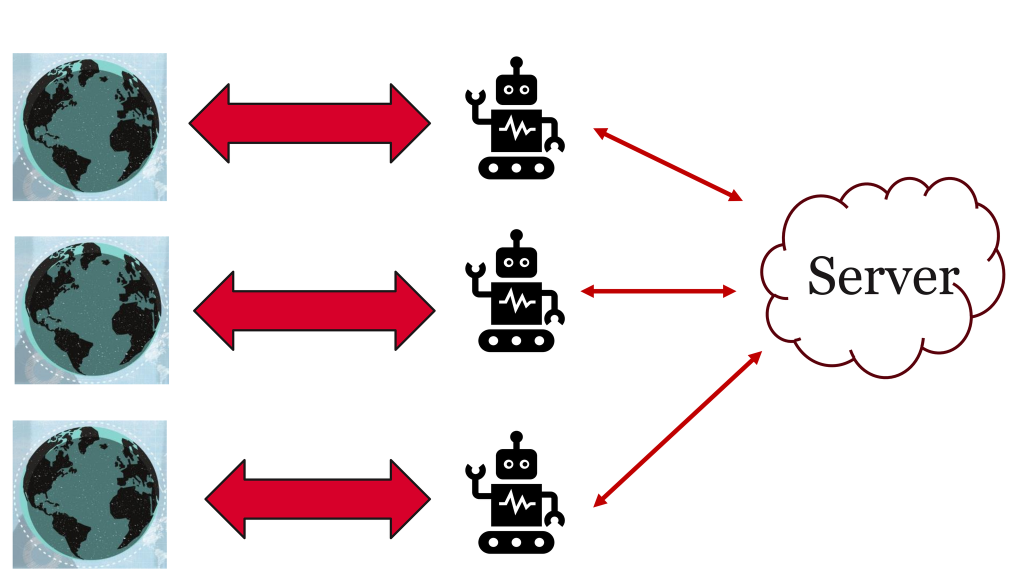
Problem Formulation

- Consider a setting with N agents, each agent i interacting with a distinct environment.
- Environment of agent i characterized by *MDP* $\mathcal{M}_i = (\mathcal{S}, \mathcal{A}, R_i, \mathcal{P}, \gamma)$.
- Agents' environments differ in reward functions (goals).
- Behavior of agent is captured by *policy* $\pi: \mathcal{S} \to \Delta(\mathcal{A})$.
- Local loss function of agent i:

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

- **Policy Gradient (PG):** Parameterize the policy to obtain π_θ, and directly optimize θ to minimize the loss function $J_i(\theta) := J_i(\pi_\theta)$.
- **Heterogeneous Federated RL:**

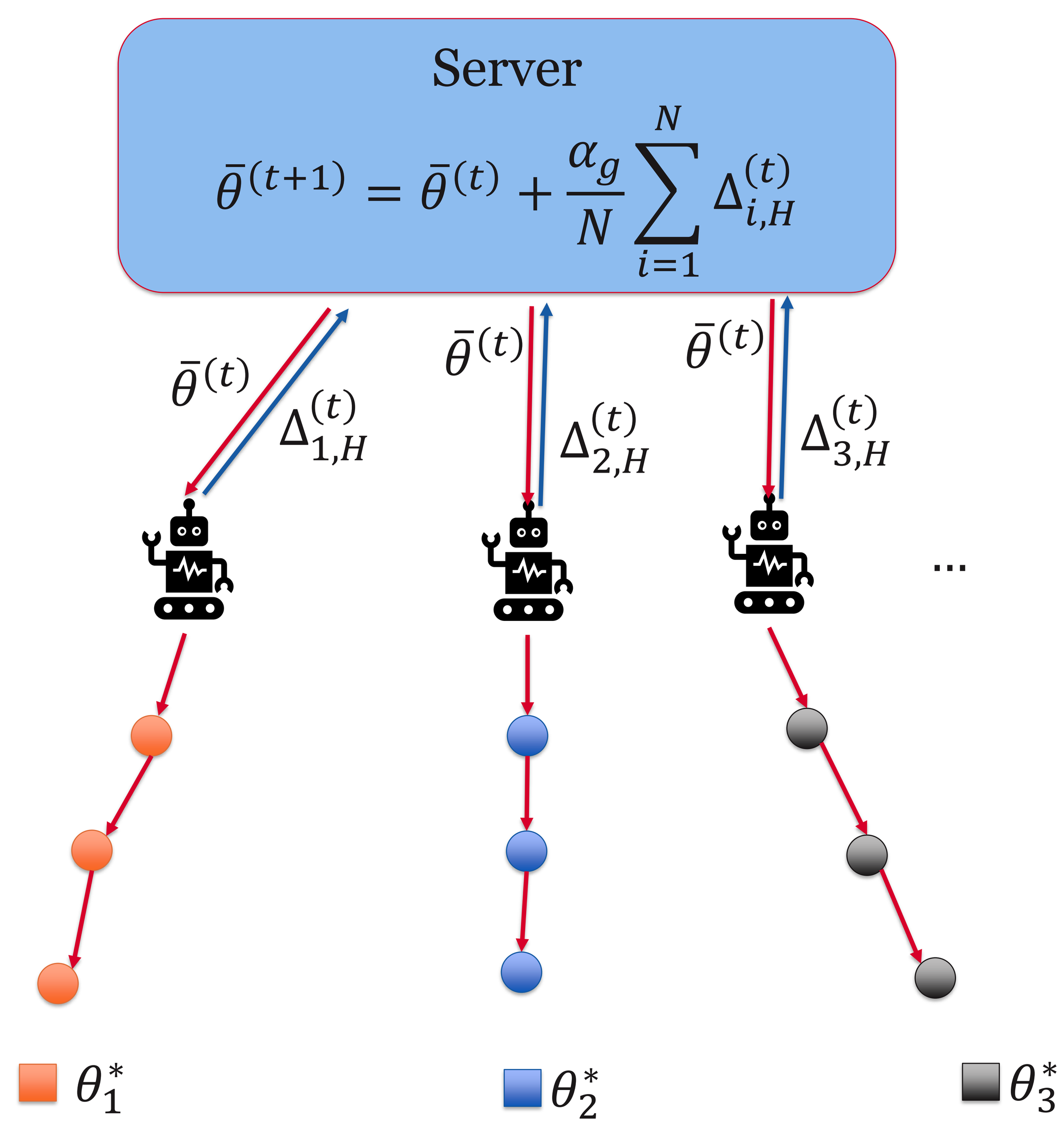
$$\min_{\theta \in \mathbb{R}^d} J(\theta) := \frac{1}{N} \sum_{i=1}^{N} J_i(\theta)$$



- Agent i only has access to **noisy** and **truncated** gradient $\widehat{\nabla}_K J_i(\cdot)$

Algorithm: Fast-FedPG

- *Communication constraint*: Server broadcasts $\bar{\theta}^{(t)}$ at round t, agents initialize and perform H local PG steps.



- *Client-drift effect*: Since each agent updates towards its own local optimum, a **heterogeneity bias** will occur that impedes convergence.
- **Intuition of Fast-FedPG:** Ideally, the update equation would be

$$\bar{\theta}^{(t+1)} = \bar{\theta}^{(t)} - \eta \frac{1}{N} \sum_{i=1}^{N} \widehat{\nabla}_K J_i(\bar{\theta}^{(t)})$$

- **Idea:** Use the **memory** of the global policy gradient $\widehat{\nabla}_K J(\bar{\theta}^{(t)})$ to add the correction term $\widehat{\nabla}_K J(\bar{\theta}^{(t)}) - \widehat{\nabla}_K J_i(\bar{\theta}^{(t)})$ to each local update:

$$\theta_{i,\ell+1}^{(t)} = \theta_{i,\ell}^{(t)} - \eta\left(\widehat{\nabla}_K J_i\left(\theta_{i,\ell}^{(t)}\right) + \widehat{\nabla}_K J(\bar{\theta}^{(t)}) - \widehat{\nabla}_K J_i(\bar{\theta}^{(t)})\right)$$

Main Results

Key Assumptions:

- The value function J_i for each agent $i \in [N]$ is L-smooth.
- The variance of the noisy truncated gradient $\widehat{\nabla}_K J_i(\cdot)$ is bounded by σ^2.
- The truncation error is at most $D\gamma^K$.

Main Challenges in Analysis:

- **Effect of reward-heterogeneity**. Agents tend to drift towards their own locally optimal parameters.
- **Effect of non-convexity**. The value function $J_i's$ are non-convex, precluding the use of standard convex optimization tools.
- **Effect of noise and truncation**. Agents can only access noisy and biased gradients $\widehat{\nabla}_K J_i(\cdot)$.

Theorem 1. *Under a suitable choice of step-size, Fast-FedPG guarantees*

$$\mathbb{E}\left[J(\bar{\theta}^{(T)}) - J(\theta^*)\right] \leq \tilde{O}\left(\frac{1}{NHT}\right)$$

# of agents

Main Takeaways:

- Our final result exhibits ***N*-fold speedup** and no **heterogeneity bias** is present.
- Theorem 1 bridges the gap in the literature, where no previous work has shown finite-time analysis with **linear speedup and no heterogeneity bias.**
- Key helper result: **Average of gradients from different MDPs is the gradient of the average MDP** – to allow us to use the gradient-domination condition that ensures fast rates.

Theorem 2. *Under a suitable choice of step-size, Fast-FedPG guarantees (without the gradient-domination condition):*

$$\mathbb{E}\left[J(\bar{\theta}^{(T)}) - J(\theta^*)\right] \leq \tilde{O}\left(\frac{1}{\sqrt{NHT}}\right)$$

Main Takeaways:

- Without the gradient domination condition, our result still achieves a **$\sqrt{N}$-fold speedup** with **no heterogeneity bias**.

Future work:

- Study the problem of learning personalized policies in the context of multi-task/federated RL.
- Explore clustering for multi-task RL.

This paper has been accepted to IEEE Conference on Decision and Control (CDC), 2024.