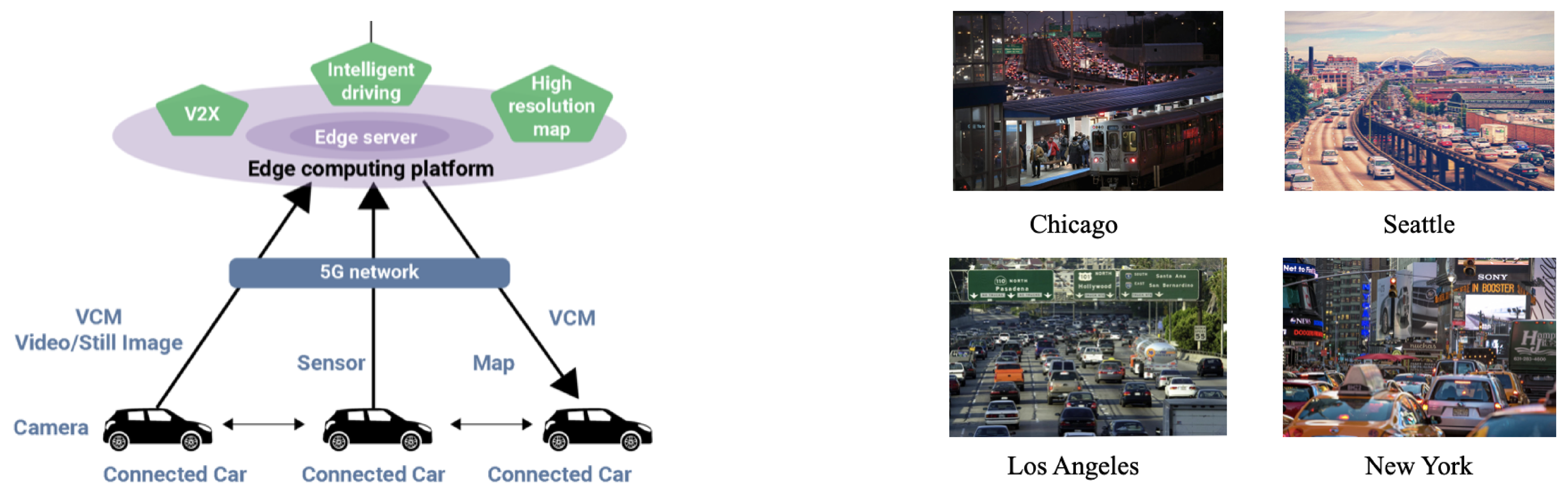
FINITE-TIME ANALYSIS OF ON-POLICY HETEROGENEOUS FEDERATED REINFORCEMENT LEARNING

Chengyu Zhang [1] Han Wang [1] Aritra Mitra [2] James Anderson [1]

[1]Columbia University [2]North Carolina State University

Motivation



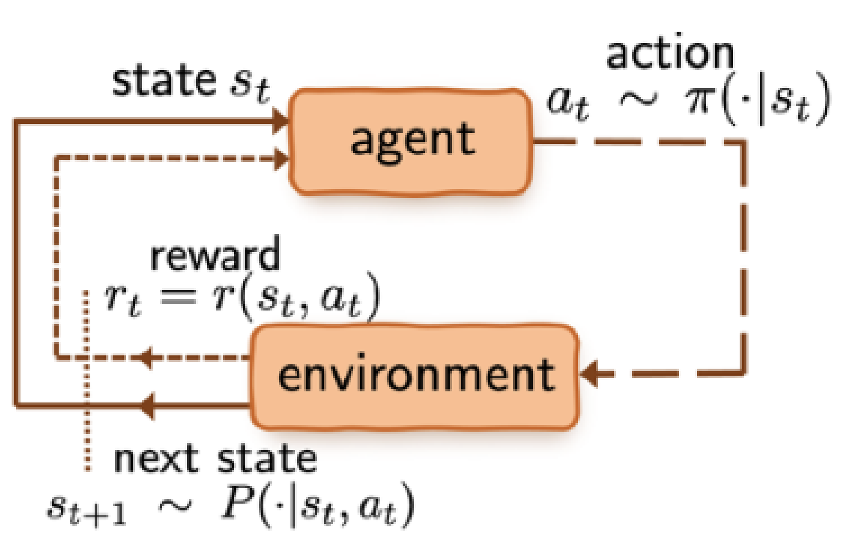
(a) Autonomous driving (b) Different cities and traffic patterns

Goal: find a universal robust strategy that minimizes the collision probability (performs well) across all environments.

Questions:

Can an agent expedite the process of learning its own near-optimal policy by leveraging information from other agents with potentially different environments?

Background



Markov Decision Process (MDP)

- $\mathcal{S}$: state space (continuous)
- $\mathcal{A}$: the action space (continuous)
- $r : \mathcal{S} \times \mathcal{A} \to [0, R]$
- $\gamma \in (0, 1)$: discounted factor
- P: Markov transition kernel
- $P_a(s, s')$: probability of transiting from state s to s' following action a.

SARSA with Linear Function Approximation

SARSA: on-policy algorithms may potentially yield more reliable convergence performance. For a given $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, we approximate the Q-value function as $Q_\theta(s, a) = \phi(s, a)^T \theta$.

Algorithm 1 SARSA

```
1: Initialization:
2: θ_0, x_0, R, φ_i, for i = 1, 2, ..., d
3: Method:
4: π_{θ_0} ← Γ(φ^T θ_0)
5: Choose a_0 according to π_{θ_0}
6: for t = 1, 2, ... do
7:     Observe x_t and r(x_{t-1}, a_{t-1})
8:     Choose a_t according to π_{θ_{t-1}}
9:     θ_t ← proj_{2,R}(θ_{t-1} + α_t − g_{t-1}(θ_{t-1}))
10:    Policy improvement: π_{θ_t} ← Γ(φ^T θ_t)
11: end for
```

- $g_t(\theta_t) = \phi(x_t, a_t)\Delta_t$, where $\Delta_t = r(x_t, a_t) + \phi^T(x_{t+1}, a_{t+1})\theta_t - \phi^T(x_t, a_t)\theta_t$.
- The projection step

$$\text{proj}_{2,R}(\theta) := \arg \min_{\theta' : \|\theta'\|_2 \le R} \|\theta - \theta'\|_2.$$

which is to control the norm of the gradient $g_t(\theta_t)$.

- Γ is the policy improvement operator, which satisfies the Lipchitz continous condition such as the softmax function.

Assumption: The behavior policy $\pi_\theta = \Gamma(\phi^T \theta)$ is Lipschitz with respect to any θ, which is

$$|\pi_{\theta_1}(a \mid x) - \pi_{\theta_2}(a \mid x)| \le C\|\theta_1 - \theta_2\|_2$$

holds for all $(x, a) \in \mathcal{X} \times \mathcal{A}$ and C is a Lipschitz constant.

Our heterogeneous FRL problem

Problem Setup: N agents whose environment is

$$\mathcal{M}^{(i)} = \left(\mathcal{S}, \mathcal{A}, r^{(i)}, P^{(i)}, \gamma\right).$$

Environmental Heterogeneity:

- Markov kernel:

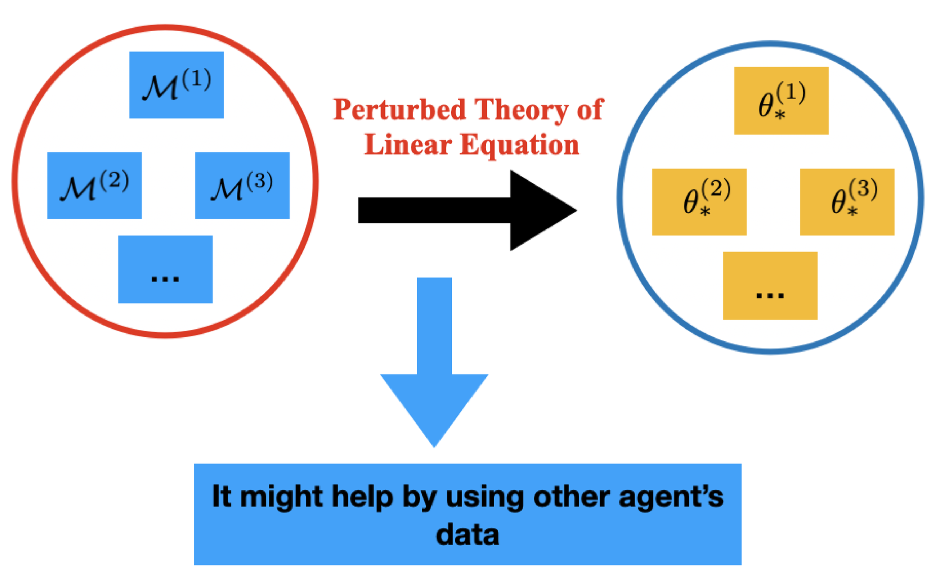
$$\max_{i,j \in [N]} \|P^{(i)} - P^{(j)}\|_{\text{TV}} \le \epsilon_p$$

- Reward:

$$\max_{i,j \in [N]} \frac{\|r^{(i)} - r^{(j)}\|_\infty}{R} \le \epsilon_r,$$

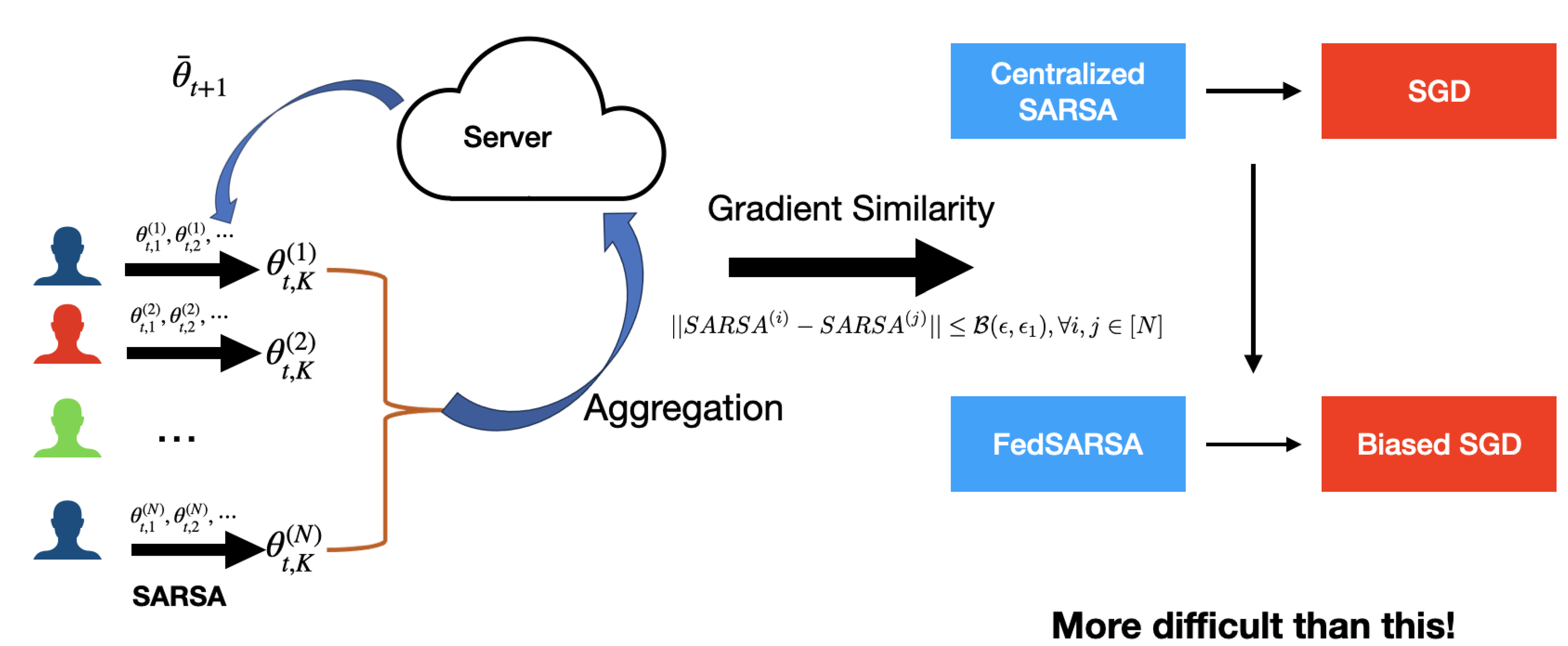
where $\|r\|_\infty = \sup_{s,a \in \mathcal{S} \times \mathcal{A}} |r(s, a)|$.

"Simulation Lemma":



Our proposed algorithm FedSARSA

FedSARSA Algorithm:



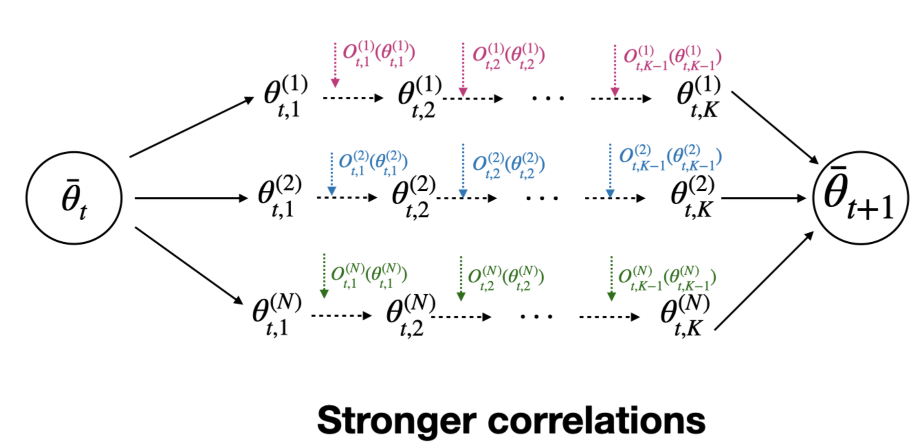
Difficulties

- We propose an on-policy heterogeneous FRL algorithm called FedSARSA.

TD: a fixed policy **VS** SARSA: time-varying polices

Difficulties:

- Linear Function Approximation
- Markov Sampling
- Multiple Local Updates
- Environmental Heterogeneity
- Time-varying behavior polices



Main Results

Q: Does more data from heterogeneous MDPs help or hurt?

Theorem: For each agent i,

$$\mathbb{E}\left\|\widetilde{\theta}_T - \theta_*^{(i)}\right\|^2 = \tilde{\mathcal{O}}\left(\frac{K^2 + \tau^5}{(1-\gamma)^2 T^2} + \underbrace{\frac{\tau}{NT}}_{\text{FL Gain}} + \underbrace{\frac{\Lambda^2(\epsilon_p, \epsilon_r)}{H^2}}_{\text{FL Cost}}\right).$$

Dorminant term: $\frac{1}{NT}$

Unavoidable!

Where K is the number of local updates, T is the number of total iterations.

Main Takeaways: In a low-heterogeneity regime, there is a clear benefit of collaboration.

Simulations

Experiments: Synthetic MDPs with $|S| = 100$, an action space of size $|A| = 100$, a feature space of dimension $d = 25$, and set $\gamma = 0.2$ and $R = 10$. The synchronization period is set to $K = 10$.

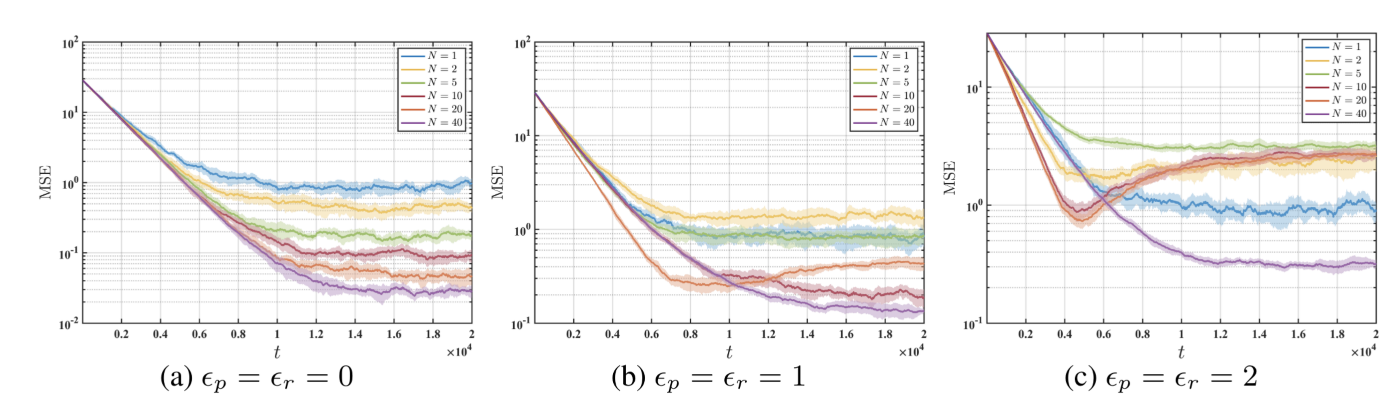


Figure 1: Performance of FedSARSA under Markovian sampling.

Main Takeaways: N times faster than independent training!

Comparison

Table 1: Comparison of finite-time analysis for value-based FRL methods. LSP and LFA represent linear speedup and linear function approximation under the Markovian sampling setting; Pred and Plan represent prediction (policy evaluation) and planning (policy optimization) tasks, respectively.

Work	Hetero-geneity	LSP	LFA	Markovian Sampling	Task	Behavior Policy
Doan et al. (2019)	✗	✗	✓	✗	Pred	Fixed
Jin et al. (2022)	✓	✗	✗	✗	Plan	Fixed
Khodadadian et al. (2022)	✗	✓	✓	✓	Pred & Plan	Fixed
Shen et al. (2023)	✗	✓[1]	✓	✓	Plan	Adaptive
Wang et al. (2023a)	✓	✓	✓	✓	Pred	Fixed
Woo et al. (2023)	✗	✓	✗	✓	Plan	Fixed
Our work	✓	✓	✓	✓	Pred & Plan	Adaptive

https://arxiv.org/pdf/2401.15273

The Twelfth International Conference on Learning Representations, Vienna