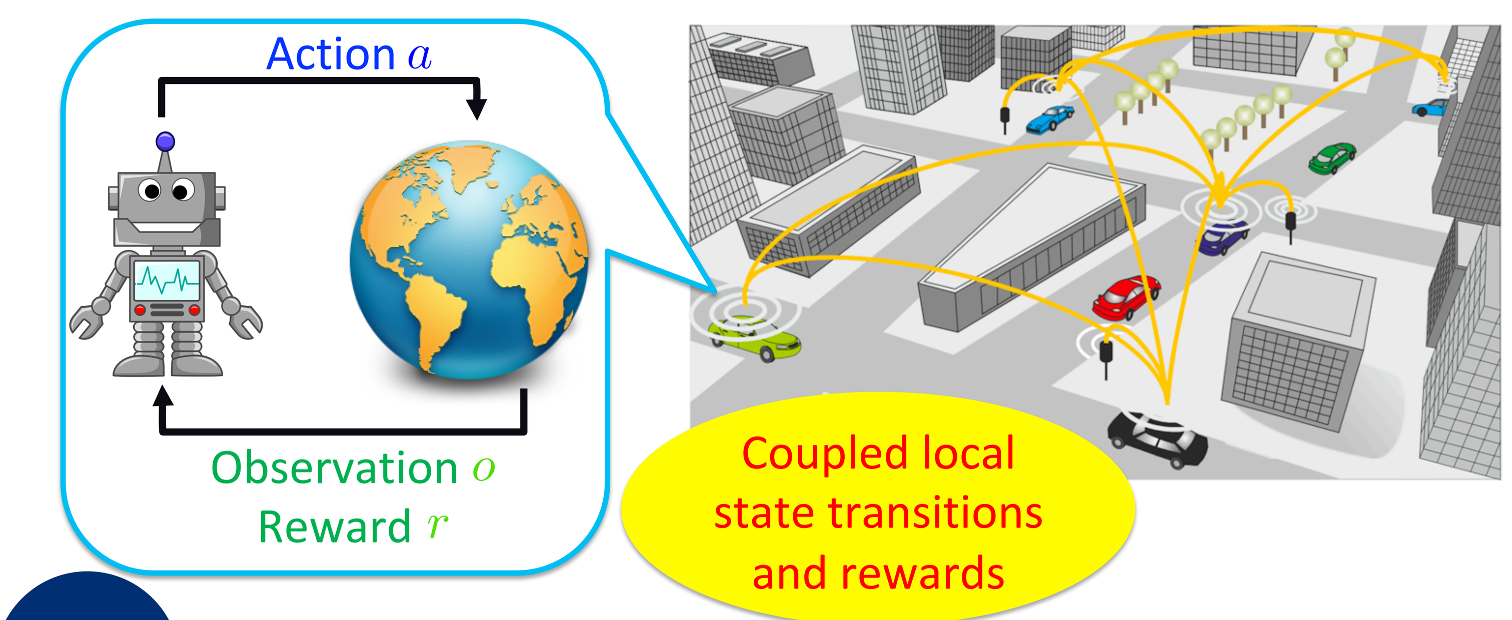
Distributed Learning for Control of Cyber-Physical Systems

Michael M. Zavlanos
Duke University

Project Overview

CPS control with Distributed Reinforcement Learning (RL) methods



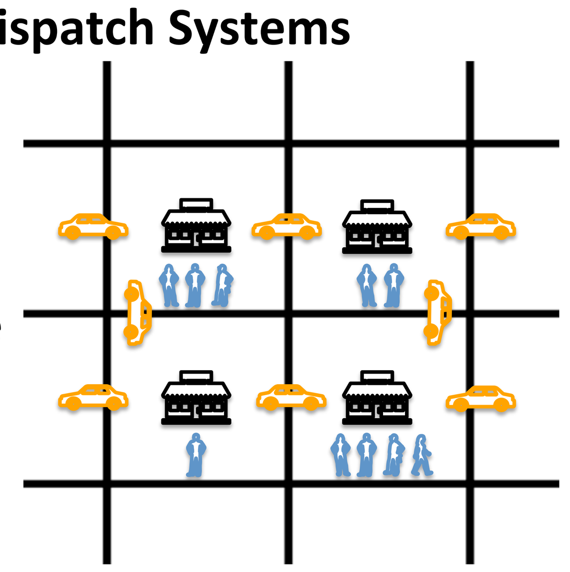
Pros
- Free of high-fidelity models
- Natural runtime adaptation

Cons
- No performance guarantees under partial observations
- Large variance during learning

Goal: To develop a novel distributed RL framework for the control of CPS, so that it has performance guarantees under partial observations.

Impact:
- Guarantees on sample complexity of distributed RL for control of CPS under partial observations
- Preservation of local user's privacy and robustness to single-node failure
- Applications to many domains, e.g., smart city, health care, etc.
- K-12, undergraduate, and graduate education
- Diversity

CPS Case Study: Distributed Shared Vehicle Dispatch Systems



- 16 dispatch centers on a 4 x 4 grid
- Local uncertain demand

$$d_i(t) = A_i sin(\omega_i t + \phi_i) + w_i(t)$$

- Local reward penalizing resource shortage

$$r_i(t) = \begin{cases} 0 & \text{if } m_i(t) > 0, \\ -m_i(t)^2 & \text{if } m_i(t) < 0 \end{cases}$$

- Transition of local resource

$$m_i(t+1) = m_i(t) - \sum_{j \in \mathcal{N}_i} a_{ij}(t) m_i(t) + \sum_{j \in \mathcal{N}_i} a_{ji}(t) m_j(t) - d_i(t)$$

- Local observation

$$o_i(t) = [m_i(t), d_i(t)]$$

Zeroth-Order Gradient Estimators

Consider the optimization problem

$$\min_{x \in \mathbb{R}^d} f(x) = \mathbb{E}_\xi [F(x, \xi)]$$

where ξ is the objective function evaluation noise.

ZO Estimator	One-Point	Two-Point
$\tilde{\nabla}\mathbf{f}(\mathbf{x})$	$\frac{u}{\delta} F(x + \delta u, \xi)$	$\frac{u}{\delta}(F(x + \delta u, \xi) - F(x, \xi))$
Drawback	Subject to large variance and slow convergence	Each update requires multiple-point evaluation, difficult to implement in distributed or non-stationary environment.

A New One-Point ZO Residual Feedback Oracle

Proposed One-Point Residual-Feedback Estimator:

$$\tilde{\nabla} f(x_t) := \frac{u_t}{\delta}(F(x_t + \delta u_t, \xi_t) - F(x_{t-1} + \delta u_{t-1}, \xi_{t-1}))$$

It solves static optimization problems with iteration complexity:

Complexity		Convex $C^{0,0}$	Convex $C^{1,1}$	Nonconvex $C^{0,0}$	Nonconvex $C^{1,1}$
One-point	Gasnikov et al.(2017)	$d^2\epsilon^{-4}$	$d\epsilon^{-3}$	–	–
Two-point	Duchi et al. (2015)	$d\log(d)\epsilon^{-2}$	$d\epsilon^{-2}$	–	–
	Shamir (2017)	$d\epsilon^{-2}$	–	–	–
	Nesterov & Spokoiny (2017)	$d^2\epsilon^{-2}$	$d\epsilon^{-1}$	$d^3\epsilon_f^{-1}\epsilon^{-2}$	$d\epsilon^{-1}$
	Bach & Perchet (2016)	–	$d^2\epsilon^{-3}$ (UN)	–	–
Residual One-point	Deterministic	$d^2\epsilon^{-2}$	$d^3\epsilon^{-1.5}$	$d^4\epsilon_f^{-1}\epsilon^{-2}$	$d^3\epsilon^{-1.5}$
	Stochastic	$d^2\epsilon^{-4}$	$d^2\epsilon^{-3}$	$d^3\epsilon_f^{-3}\epsilon^{-2}$	$d^4\epsilon^{-3}$

where ϵ is suboptimality in the function value for convex problems, or is the squared norm of the gradient at the final iterate. The one-point residual feedback estimator enjoys almost the same convergence speed as the two-point estimator, but only require one-point evaluation per update.

Distributed RL under Partial Observations

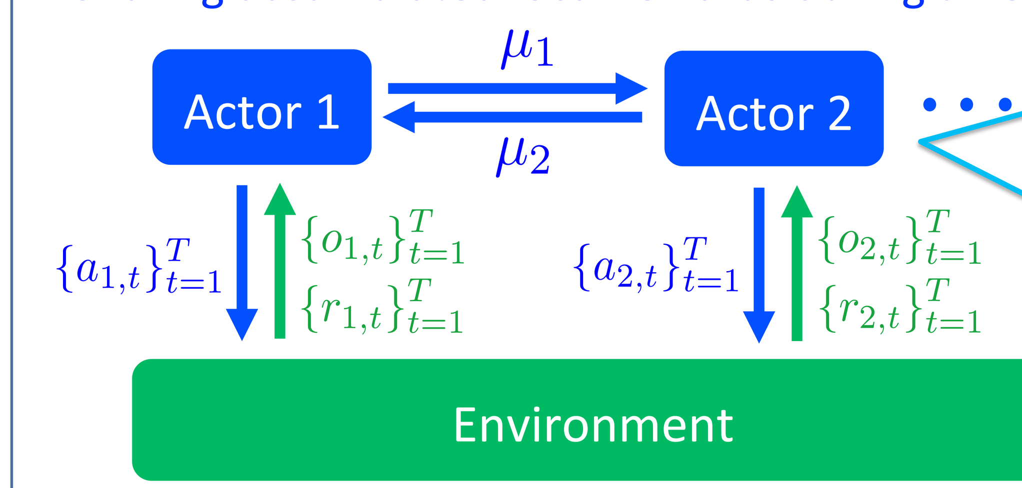
Proposed distributed RL problem:

$$\min_\theta J(\theta) := \sum_{i=1}^N J_i(\theta)$$

where $\theta = [\theta_1^T, \theta_2^T, \ldots, \theta_N^T]^T$ and θ_i parameterizes agent i 's local policy function $\pi_i : \mathcal{O}_i \to \mathcal{A}_i$.

Proposed Framework:

Sharing accumulated local rewards during an episode



Update local policy with **zeroth-order policy gradient (1)**, instead of distributed Actor-Critic based policy gradient which requires full observation.

Distributed ZO policy gradient estimator with residual feedback:

$$\theta_{i,k+1} = \theta_{i,k} + \alpha \frac{\tilde{J}(\theta_k + \delta u_k, \xi_k) - \tilde{J}(\theta_{k-1} + \delta u_{k-1}, \xi_{k-1})}{\delta} u_{i,k} \qquad (1)$$

where $\tilde{J}(\theta_k + \delta u_k, \xi_k) = \sum_{j \in \mathcal{N}_i} [W^{N_c}]_{ij} \mu_j^k$, W and N_c represent the communication matrix and the number of consensus steps per episode.

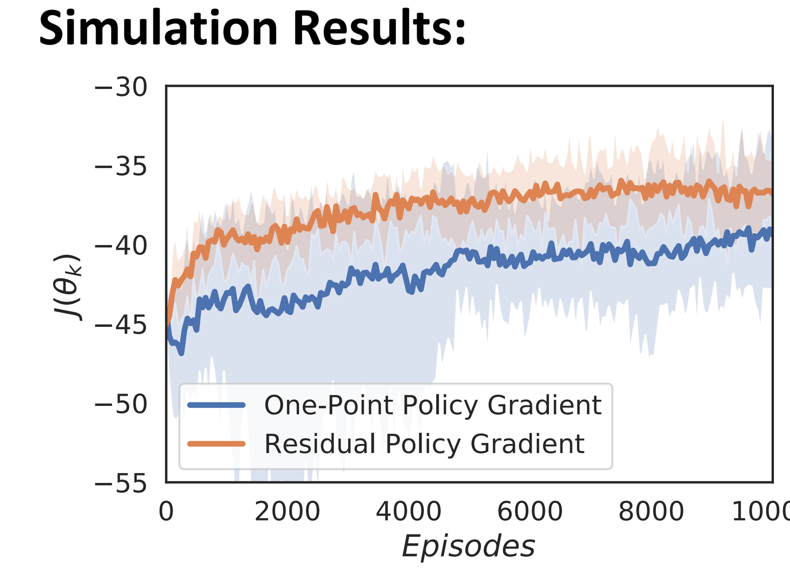
Theorem: Assume that the value function $J_i(\theta, \xi)$ belongs to the interval $[J_l, J_u]$ and select the number of consensus steps as

$$N_c \geq \log(\frac{\sqrt{\epsilon}\epsilon_J}{\sqrt{2} d^{1.5} L_0 (J_u - J_l)}) / \log(\rho_W)$$

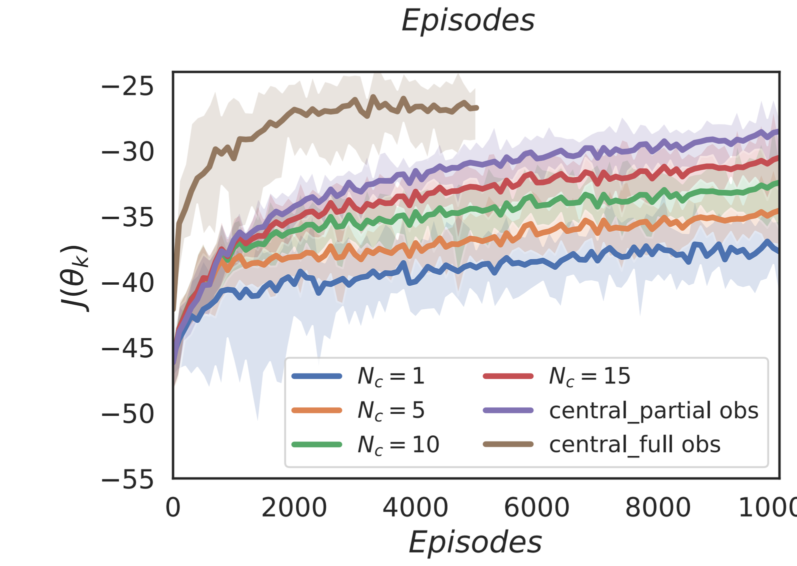
Then, we have that

$$\frac{1}{K} \sum_{k=1}^{K-1} \mathbb{E}[\|\nabla J_\delta(\theta_k)\|^2] \leq \mathcal{O}(d^{1.5} \epsilon_J^{-1.5} K^{-0.5}) + \frac{\epsilon}{2}.$$

Simulation Results:



The proposed distributed residual-feedback zeroth-order policy gradient enjoys faster convergence speed and lower variance during learning, compared to conventional one-point policy gradient.



The performance of the proposed distributed RL algorithm gets closer to that of the centralized algorithm under partial observation scenario as N_c increases .

2021 NSF Cyber-Physical Systems Principal Investigators' Meeting
June 2-4, 2021

Award ID#: CNS #1932011