Zeroth-Order Distributed and Online Learning

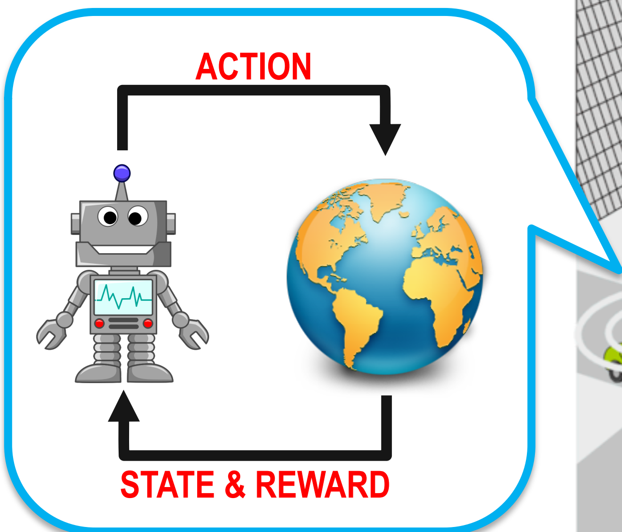
Yan Zhang and Michael M. Zavlanos

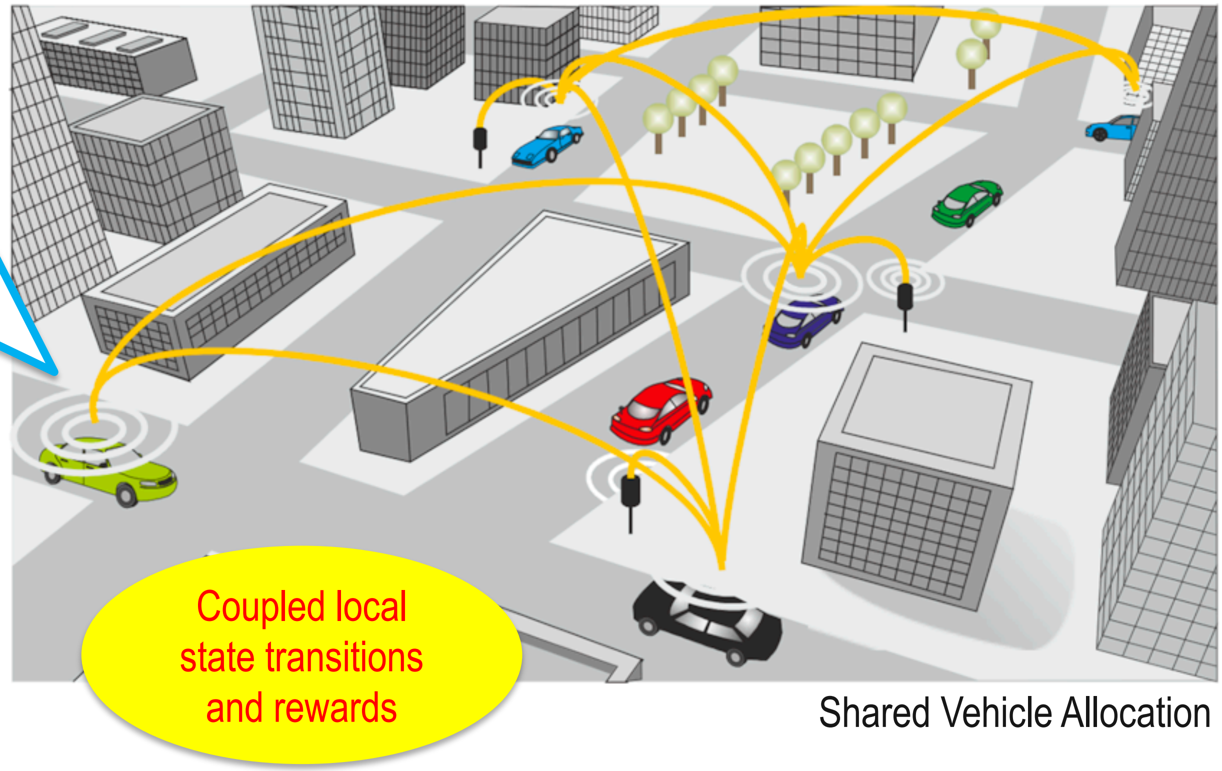
Mechanical Engineering & Materials Science
Electrical & Computer Engineering
Computer Science
Duke University

Assured Autonomy in Contested Environments (AACE)
Fall 2020 Review
October 29, 2020

Duke
UNIVERSITY

Distributed Learning for Control





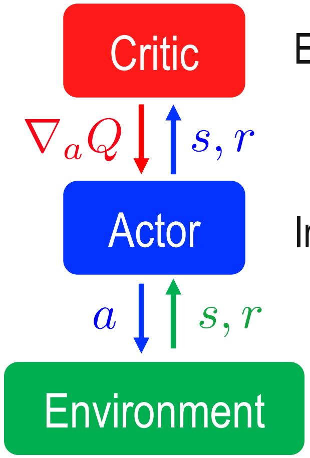
Shared Vehicle Allocation

Find the optimal policy for each agent to maximize the network-wide accumulated rewards.

Duke
UNIVERSITY

Learning for Control

Problem $\max_\theta J(\theta) \triangleq \mathbb{E}_{[s_0,a_0]\sim\rho^0}[r(s_0,a_0) + \sum_{t=1}^{T} \gamma^t r(s_t, \pi_\theta(s_t))]$



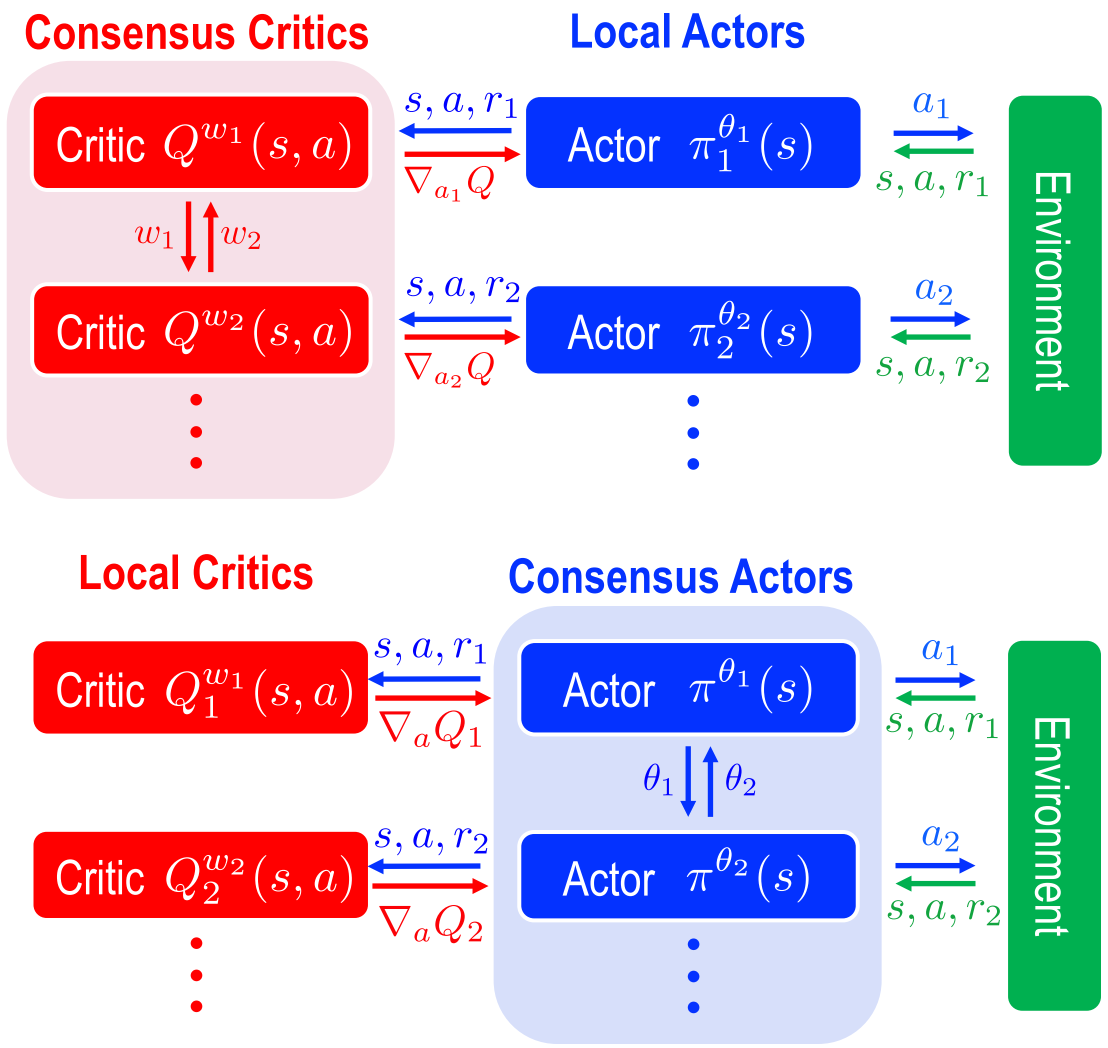
Estimate the accumulated reward given the current policy: $Q^w(s_t, a_t)$

Policy evaluation algorithms to find w, e.g., TD learning

Improve the current policy by policy gradient: $\theta(t+1) = \theta(t) + \alpha \nabla_\theta \pi^\theta(s_t) \nabla_{a_t} Q^w(s_t, a_t)$

Compute policy gradient using, e.g., backpropagation, if NN

Distributed Learning for Control



Require every agent to observe **global state and action information**

[Zhang et al., 2018] [Suttle et al., 2019] [Zhang & Zavlanos, 2019]

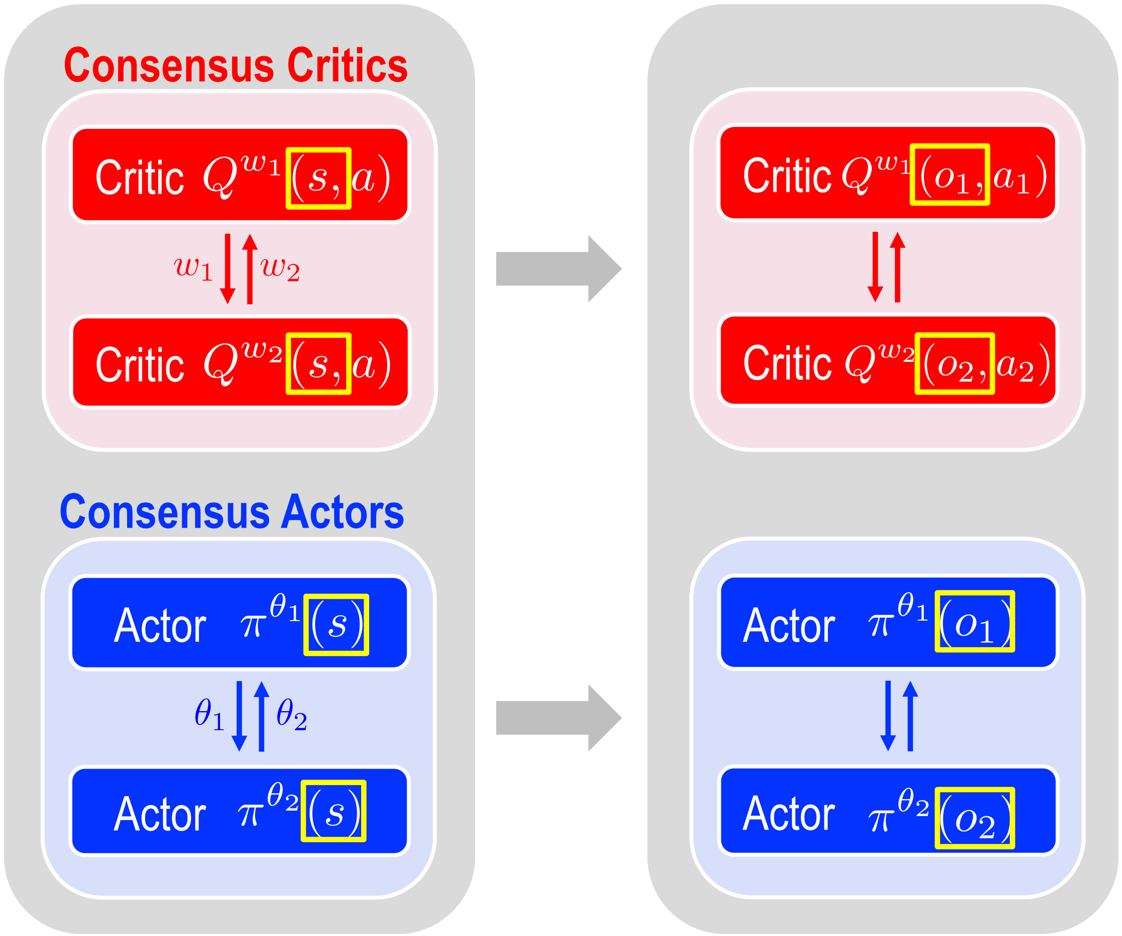
Duke UNIVERSITY

Partial Observations

Full Observations

Partial Observations

$$o_{i,t} = h([s_1, s_2, \ldots, s_N], w_{i,t})$$



Local value/policy functions are based on local observations that have different meanings, so the parameters of these local functions do not need to be equal.

Can not enforce consensus and, therefore, do not have access to the global policy and value functions!

SO WHAT CAN WE DO???

Duke UNIVERSITY

Zeroth-Order (Derivative-Free) Optimization

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

Complex or unknown models:
Gradient is unavailable,
uncomputable, private

Zeroth-order gradient estimators:

The one-point estimator $\widetilde{\nabla} f(x) = \frac{u}{\delta} F(x + \delta u, \xi)$ requires that the function $F(x, \xi)$ is bounded, it is subject to large variance and, therefore, slow convergence rate.

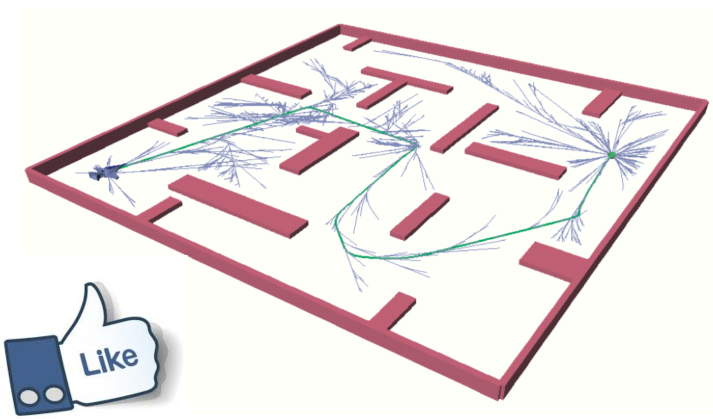
The two-point estimator $\widetilde{\nabla} f(x) = \frac{u}{\delta} \big(F(x + \delta u, \xi) - F(x, \xi) \big)$ requires that the function evaluations at x and $x + \delta u$ are subject to the same noise vector ξ. It is impossible to use if the objective function is time varying.

New residual-feedback zeroth-order gradient estimator:

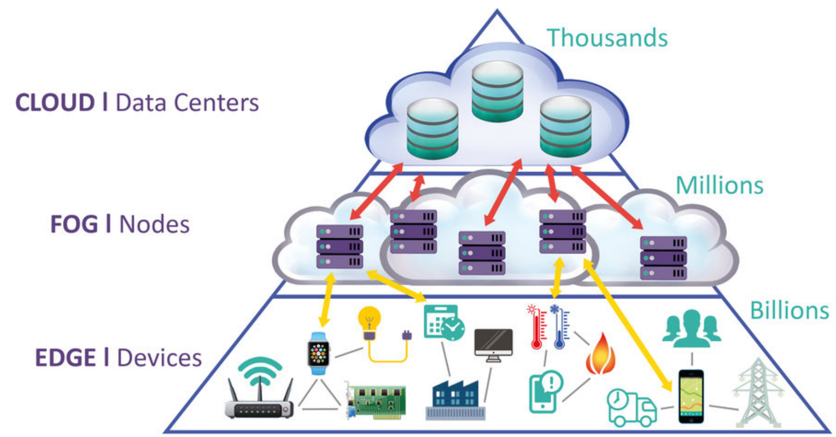
$$\widetilde{\nabla} f(x_t) := \frac{u_t}{\delta} \big(F(x_t + \delta u_t, \xi_t) - F(x_{t-1} + \delta u_{t-1}, \xi_{t-1}) \big)$$

Reduce the variance
of one-point gradient
estimator using the
previous iterate

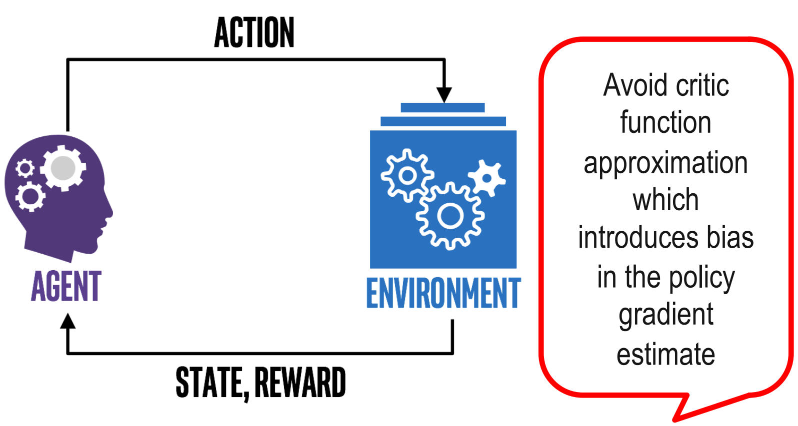
Optimization with Complex or Unknown Models



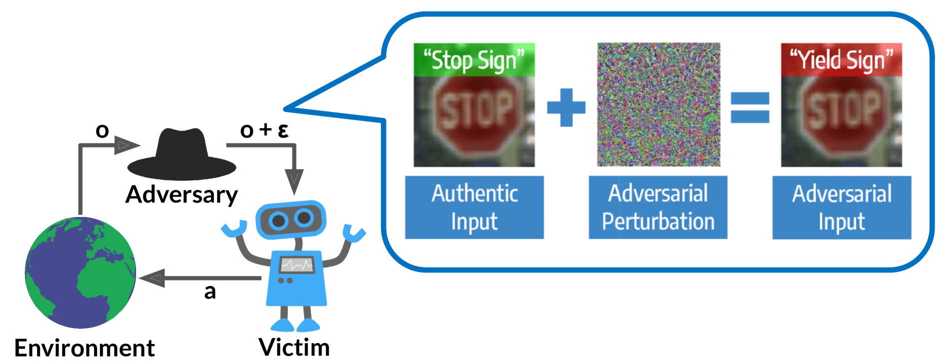
Human-in-the-loop robot planning



Computation management in IoT systems



Policy search in RL $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$



Adversarial attacks and defense in DRL

Zeroth-Order Distributed Policy Gradient Optimization

Avoid Critic Altogether

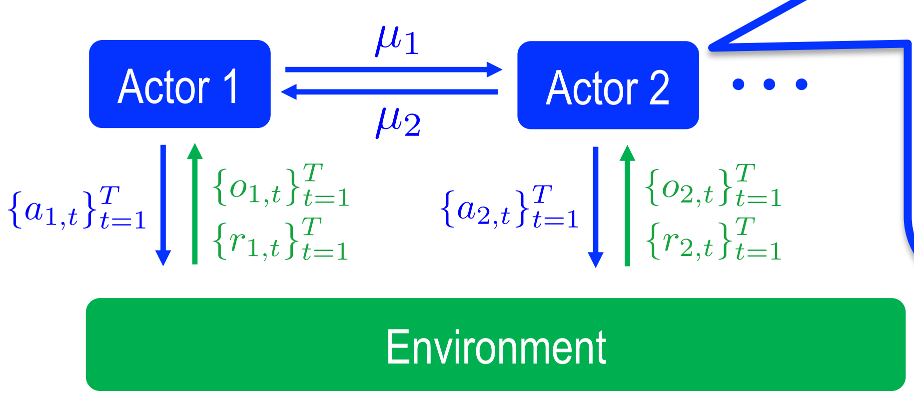
Centralized zeroth-order residual-feedback policy gradient optimization

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

$J(\theta_k + \delta u_k, \xi_k) = \sum_{i=1}^{N} J_i(\theta_k + \delta u_k, \xi_k)$ is the global return of implementing policy $\pi^{\theta_k + \delta u_k}$ at the end of episode k, which can be computed in a decentralized way using consensus.

The return in the past iteration reduces the variance of the zeroth-order policy gradient estimate, similar to the baseline effect used in the Actor Critic method.

Distributed zeroth-order policy gradient optimization



Step 1: Perturb local policy parameters, collect local rewards, and compute local return $J_i = \sum_{t=1}^{T} \gamma^{t-1} r_{i,t}$.

Step 2: Let $\mu_i^k(0) = J_i$, then run N_c local averaging steps $\mu_i^k(m+1) = \sum_{j \in \mathcal{N}_i} W_{ij} \mu_j^k(m)$.

Step 3: Update local policy parameter

$$\theta_{i,k+1} = \theta_{i,k} + \alpha \frac{\mu_i^k(N_c) - \mu_i^{k-1}(N_c)}{\delta} u_{i,k}$$

Duke UNIVERSITY

Convergence Analysis

Assumption 1: For all agents, the local policy evaluation is unbiased and subject to bounded variance. That is, $\mathbb{E}_\xi\big[J_i(\theta,\xi)\big] = J_i(\theta)$ and $\mathbb{E}\big[(J_i(\theta,\xi) - J_i(\theta))^2\big] \leq \sigma^2$ for $i = 1, 2, \ldots, N$.

Assumption 2: The local values $J_i(\theta,\xi)$ are upper and lower bounded by J_u and J_l for all $i = 1, 2, \ldots, N$ and all policy parameters θ.

Bounded bias in the local policy gradients due to consensus errors

Theorem: (Learning Rate) Let Assumptions 1 and 2 hold and define $\delta = \frac{\epsilon_J}{\sqrt{d}L_0}$, $\alpha = \frac{\epsilon_J^{1.5}}{4d^{1.5}L_0^2\sqrt{K}}$ and $N_c \geq \log(\frac{\sqrt{\epsilon}\epsilon_J}{\sqrt{2}d^{1.5}L_0(J_u - J_l)})/\log(\rho_W)$. Then, we have that

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

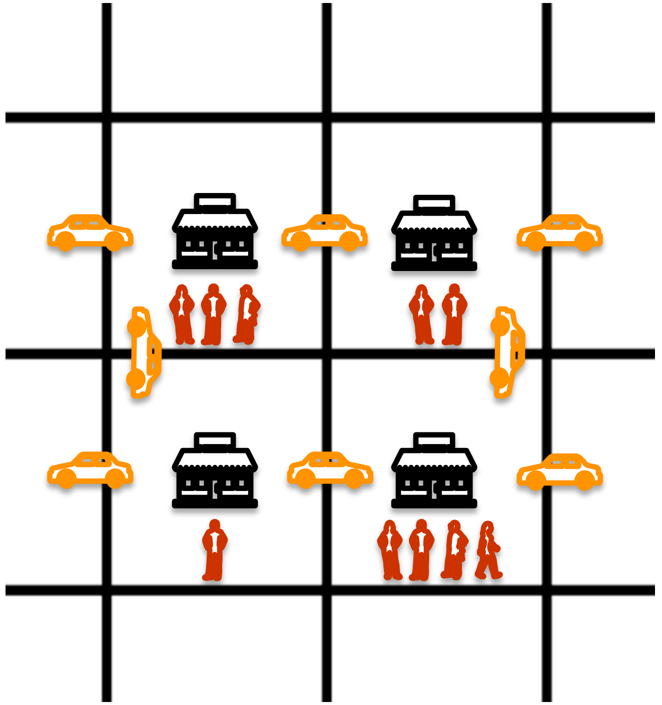
The number of consensus steps run per episode depends on the upper and lower bounds of the value functions.

Given the desired solution accuracy ϵ, ϵ_f , we can select the smoothing parameter δ , the step size α and the number of consensus steps N_c per episode, so that the $\epsilon - \epsilon_f$ solution is found after K episodes.

Duke UNIVERSITY

Distributed Resource Allocation

16 agents on a 4 x 4 grid



Local demand at agent i

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

Local reward

$$r_i(s_i(t)) = \begin{cases} 0 & \text{if } m_i(t) > 0, \\ -(-m_i(t))^3 & \text{if } m_i(t) < 0. \end{cases}$$

Dynamics of local resources

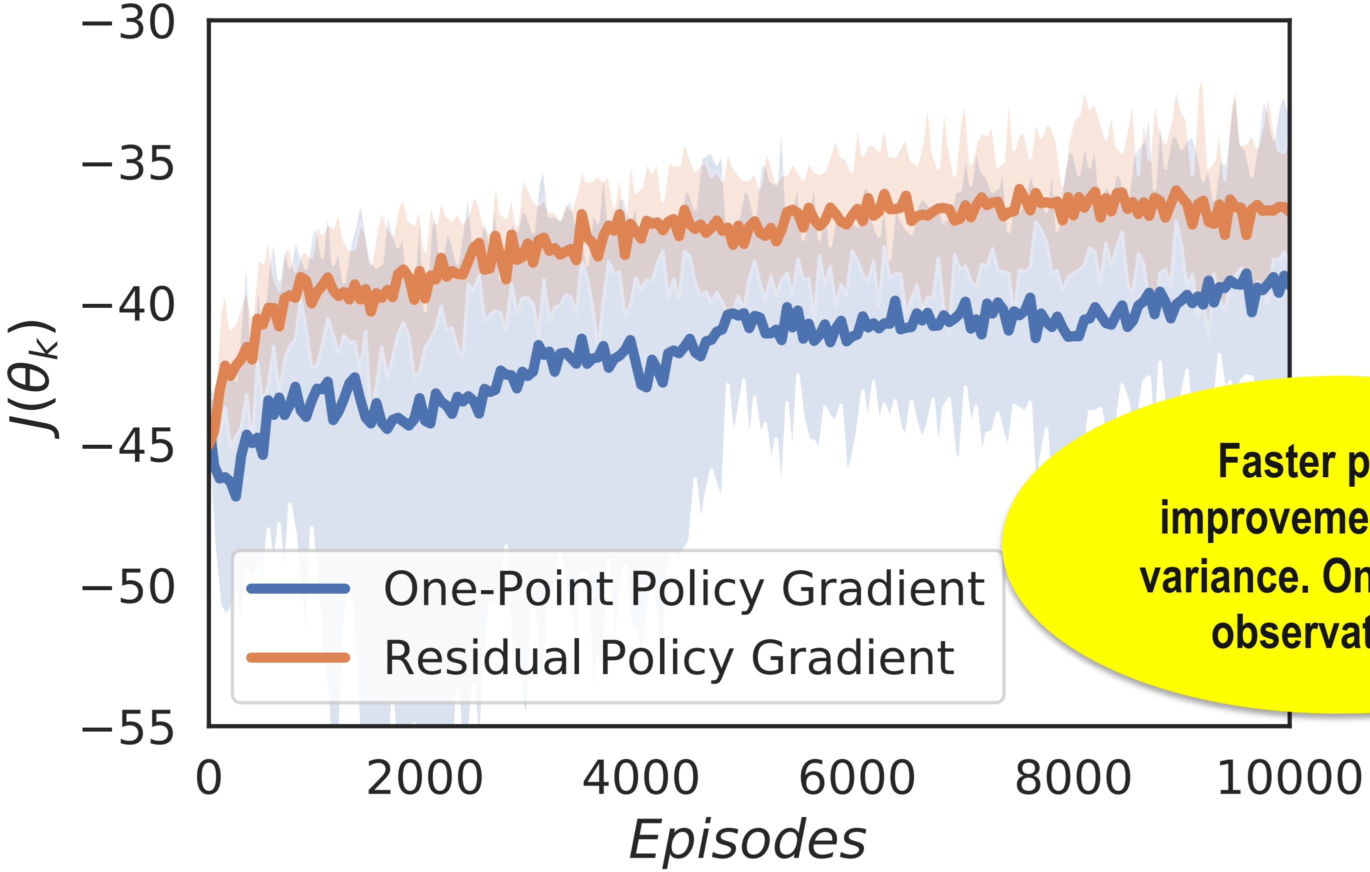
$$m_i(t+1) = m_i(t) - \sum_{j \in \mathcal{N}_i} a_{ij}(t) m_i(t) + \sum_{j \in \mathcal{N}_i} a_{ji}(t) m_j(t) - d_i(t)$$

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

Duke
UNIVERSITY

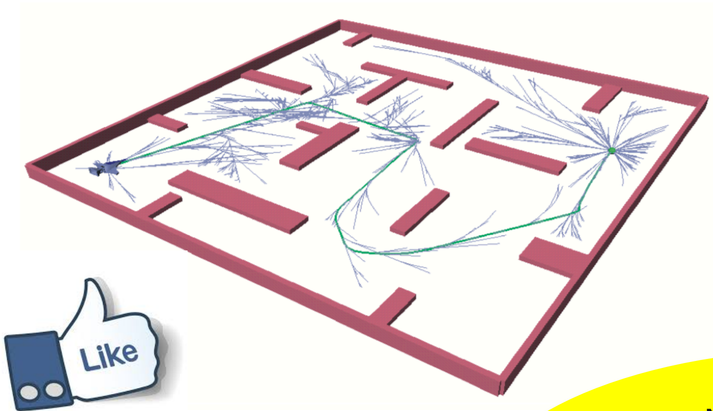
Distributed Resource Allocation

Performance improvement of distributed zeroth-order policy optimization algorithms

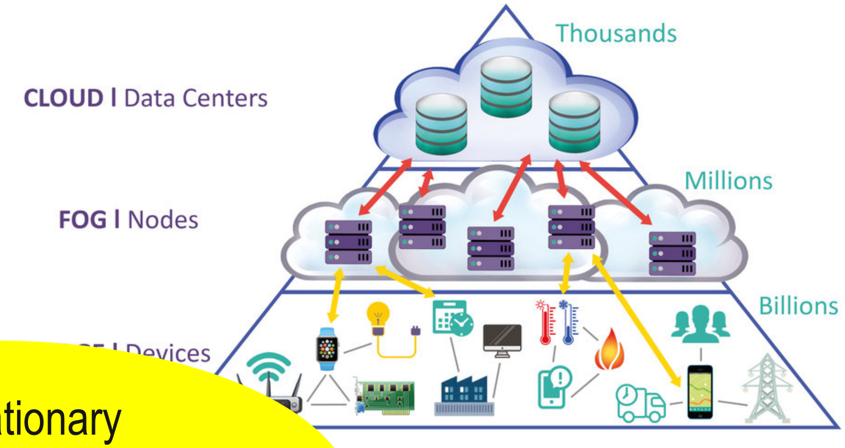


Faster policy improvement. Less variance. Only partial observations.

Zeroth-Order Online Learning for Control

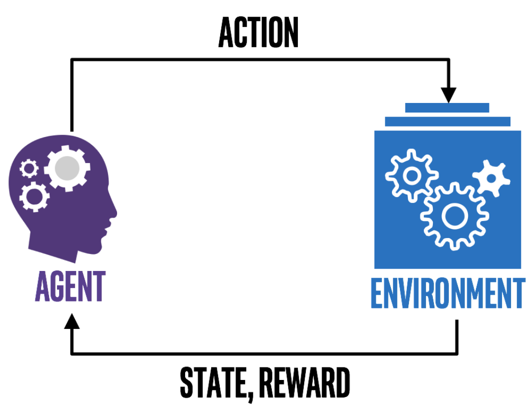


Human-in-the-loop robot p



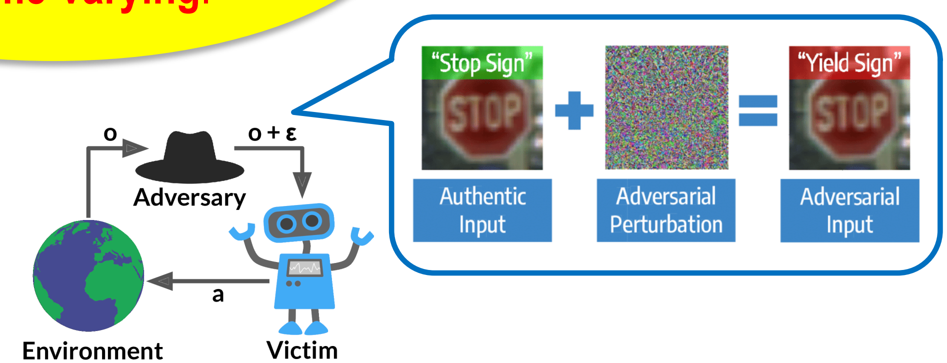
management in IoT systems

Non-stationary environments: **The reward and dynamic functions are time-varying.**



Avoid cri
function approximation which introduces bias in the policy gradient estimate

Policy search in RL $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$



Adversarial attacks and defense in DRL

Online Optimization

Time-varying non-convex optimization

$$\min_{\{x_t\}} \sum_{t=0}^{T-1} f_t(x_t)$$

Performance measure

Gradient Size Regret:

$$R_g^T := \sum_{t=0}^{T-1} \mathbb{E}[\|\nabla f_t(x_t)\|^2]$$

→ Tracking the time-varying stationary points

Online zeroth-order gradient estimators

Two-point estimator:

$$\frac{u}{\delta}\Big(f_t(x+\delta u) - f_t(x)\Big)$$

Impractical to use because f_t can only be evaluated once.

Traditional one-point estimator:

$$\frac{u}{\delta} f_t(x+\delta u)$$

Does not track the non-stationary points well because of large variance.

Duke UNIVERSITY

Residual-Feedback Online Optimization

$$x_{t+1} = x_t - \eta \boxed{\frac{u_t}{\delta}\big(f_t(x_t + \delta u_t) - f_{t-1}(x_{t-1} + \delta u_{t-1})\big)}$$

Assumption: (Bounded Regularity) There exist constants $W_T, \widetilde{W}_T > 0$ such that the sequence of functions $\{f_t\}_{t=0,\ldots,T-1}$ satisfies the following two conditions.

$$1.\ \sum_{t=1}^{T} \mathbb{E}[f_t(x) - f_{t-1}(x)] \leq W_T; \quad 2.\ \sum_{t=1}^{T} \mathbb{E}[|f_t(x) - f_{t-1}(x)|^2] \leq \widetilde{W}_T \ \text{ for all } t \text{ and } x.$$

$W_T, \widetilde{W}_T$ measure the total variation of the objective value at any fixed policy.

Theorem: (Regret for Smooth Nonconvex Problems) Assume that $f_t(x) \in C^{0,0} \cap C^{1,1}$ with Lipschitz constant L_0 and smoothness constant L_1 and that f_t is bounded below by f_t^* for all t . Run ZO with residual feedback for T iterations with $\eta = (2\sqrt{2}L_0 d^{\frac{4}{3}} T^{\frac{1}{2}})^{-1}$ and $\delta = (d^{\frac{5}{6}} T^{\frac{1}{4}})^{-1}$. Then,

$$R_g^T = \mathcal{O}(d^{\frac{4}{3}} L_0 W_T T^{\frac{1}{2}} + d^{\frac{4}{3}} L_1 {L_0}^{-1} \widetilde{W}_T).$$

The algorithm tracks the path of the non-stationary points within a neighborhood, the size of which is given by the bound on the variation of the objective function.

Duke UNIVERSITY

Non-Stationary LQR

Dynamical system: $$x_{k+1} = A_t x_k + B_t u_k + w_k$$

Dynamical matrices change over each episode t

Policy function: $$u_k = K_t x_k$$

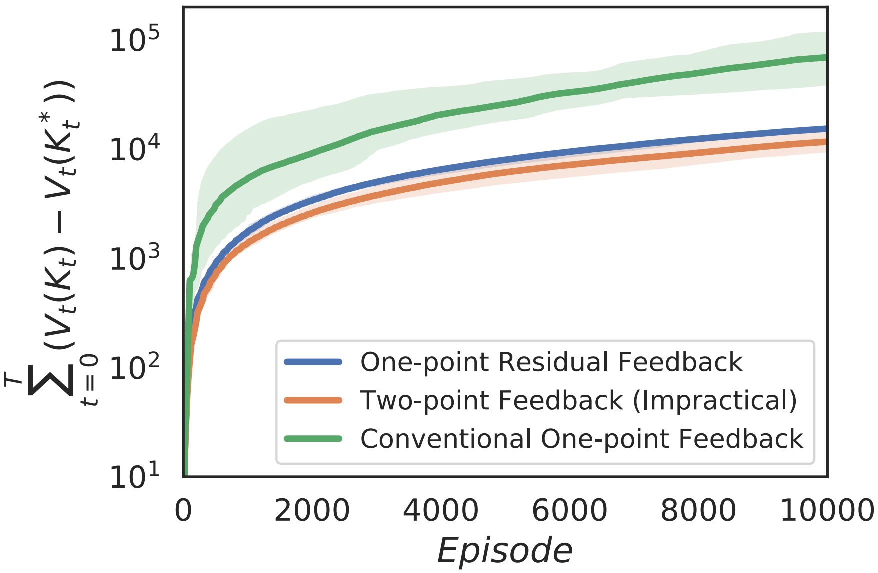
Policy parameter applied during episode t

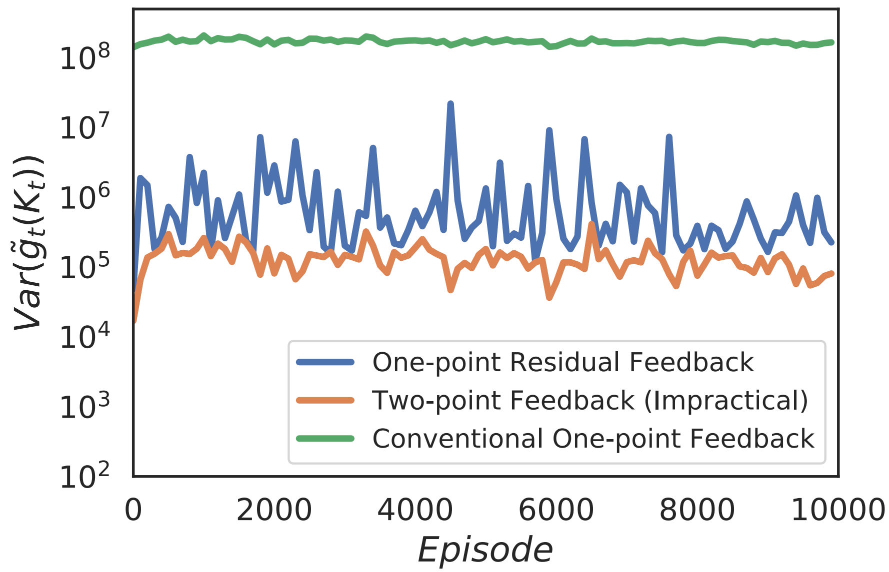
Objective function: $$V_t(K) := \mathbb{E}[\sum_{k=0}^{H-1} \gamma^k (x_k^T Q x_k + u_k^T R u_k)]$$

Objective function at episode t

Duke UNIVERSITY

Non-Stationary LQR





Applying the one-point residual feedback estimator achieves the same level of accumulated suboptimality as the impractical two-point feedback, both much lower than that of the conventional one-point feedback scheme.

Duke
UNIVERSITY

Non-Stationary Resource Allocation

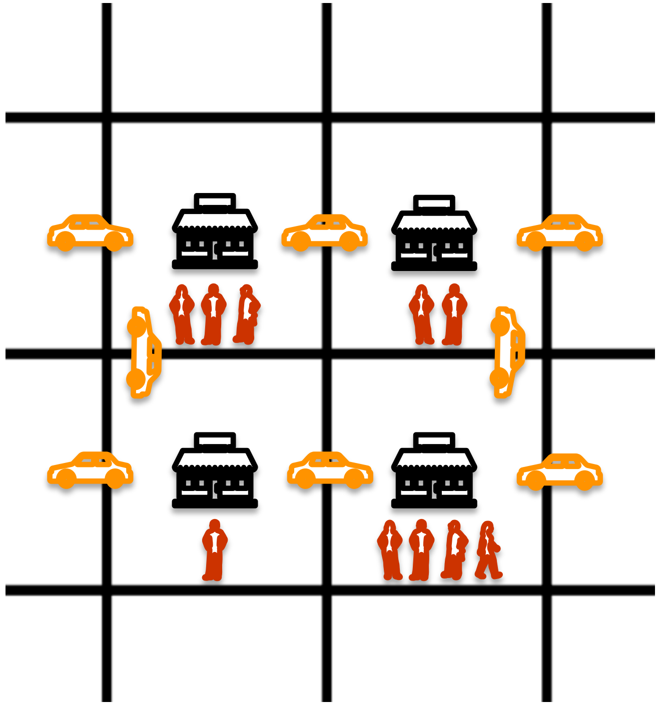
Dynamical system:

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

Reward function:

$$r_{i,t}(k) = \begin{cases} 0, & \text{when } m_i(k) \geq 0, \\ \zeta_t m_i(k)^2, & \text{when } m_i(k) < 0. \end{cases}$$

Sensitivity to the shortage of resources change over each episode t.

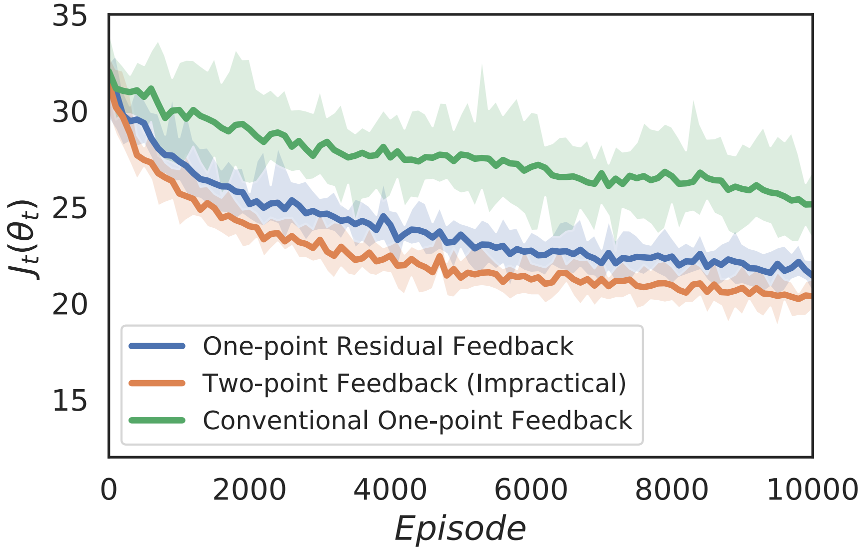


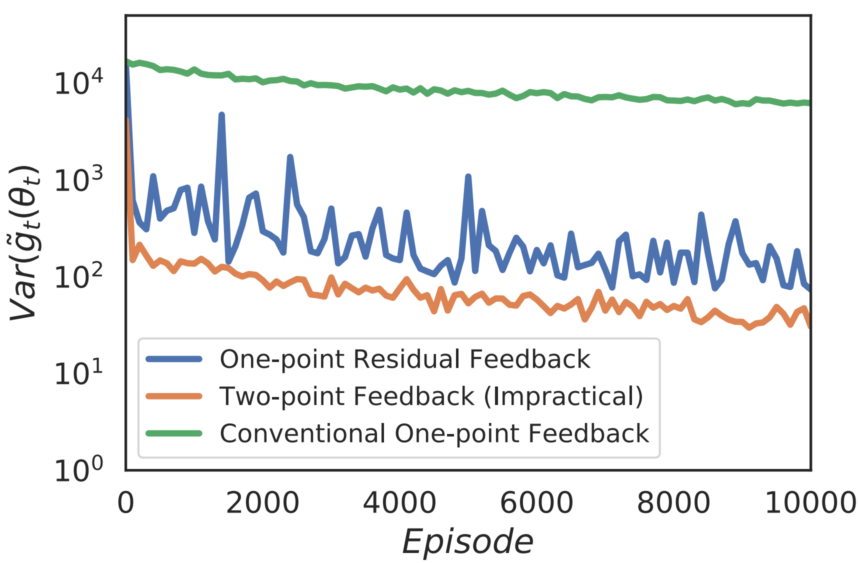
Policy function: $\pi_{i,t}(o_i; \theta_{i,t}) : \mathcal{O}_i \to [0,1]^{|\mathcal{N}_i|}$

Objective function: $J_t(\theta_t) = \sum_{i=1}^{N} \sum_{k=0}^{H} \gamma^k r_{i,t}(k)$

Duke UNIVERSITY

Non-Stationary Resource Allocation





Applying the one-point residual feedback estimator can maintain low costs in non-stationary environments as well as the impractical two-point feedback, both much lower than that of the conventional one-point feedback scheme.

Duke
UNIVERSITY

Acknowledgements

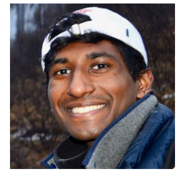
Current Group Members

Reza Khodayi
Postdoc

Luke Calkins
PhD ME

Yan Zhang
PhD ME

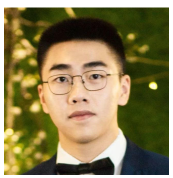
Xusheng Luo
PhD ME



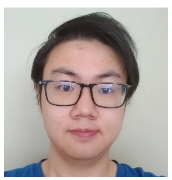
Kavin Sivakumar
PhD ECE

Yi Shen
PhD ME

Panagiotis Vlantis
Postdoc



Jayson Zhou
MSc ME



Shiqi Sun
MSc ME

Jim Turner
MSc CS

Chenyu Liu
MSc ME

Cong Li
MSc ME

Amik Mandal
Undergrad CS

Pratik Mulpury
Undergrad CS

Kenneth Marenco
Undergrad ME

Support

Thank You

Distributed Zeroth-Order Learning for Control

- Y. Zhang, Y. Zhou, K. Ji, and M. M. Zavlanos, "Boosting One-Point Derivative-Free Online Optimization via Residual Feedback," 9th International Conference on Learning Representations (ICLR), 2021, under review.
- Y. Zhang and M. M. Zavlanos, "Cooperative Multi-Agent Reinforcement Learning with Partial Observations," Journal of Machine Learning Research, under review.
- Y. Zhang, Y. Zhou, K. Ji, and M. M. Zavlanos, "Improving the Convergence Rate of One-Point Zeroth-Order Optimization using Residual Feedback," IEEE Transactions on Automatic Control, under review.

Duke UNIVERSITY