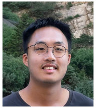
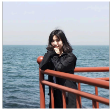
Causality Meets Locality: Provably Generalizable and Scalable Policy Learning for Networked Systems

Hao Liang

King's College London

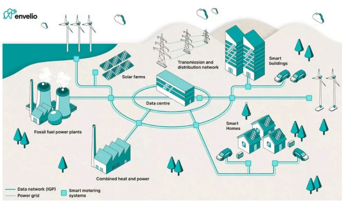
Joint work with

Shuqing Shi (KCL) **Yudi Zhang** (TU/e) **Biwei Huang** (UCSD) **Yali Du** (KCL)

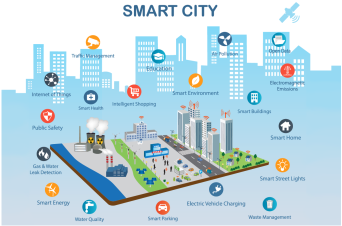
H. Liang*, S. Shi*, Y. Zhang, B. Huang, Y. Du. "Causality Meets Locality: Provably Generalizable and Scalable Policy Learning for Networked Systems." NeurIPS 2025 (Spotlight).

Motivation: Real-World Networked Systems



(a) Grid

(b) Transportation



(c) Smart city

Key challenges

- **Scalability**: Exponential state-action space growth
- **Environment changes**: traffic patterns change, user demands vary

Current methods either scale OR generalize, but rarely both

Central Research Problem

Can we design a provably **generalizable** AND **scalable** MARL algorithm for networked systems?

Our answer: Yes!

Generalizable and **S**calable **A**ctor-**C**ritic (GSAC)

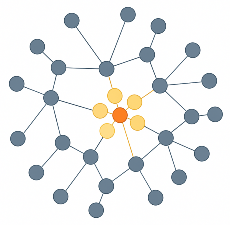
Key insights

1) **Locality + Causality** $\rightarrow$ Scalability

- **Locality**: Exploit local structure
- **Causality**: Identify minimal relevant features

2) Meta-training $\rightarrow$ Generalization

Problem Setup: Networked MARL



- **Graph:** $\mathscr{G} = (\mathscr{N}, \mathscr{E})$
- $\mathscr{N} := (1, 2, \cdots, n)$
- $\mathcal{N}_i$ is the neighbors of agent i

- Agent i observes local state $\mathbf{s}_i \in \mathcal{S}_i$
- Selects a local action $\mathbf{a}_i \in \mathcal{A}_i$
- Global state $\mathbf{s} = (\mathbf{s}_1, \ldots, \mathbf{s}_n)$
- Joint action $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)$
- **Decentralized dynamics**

$$P(\mathbf{s}(t+1) \mid \mathbf{s}(t), \mathbf{a}(t)) = \prod_{i=1}^{n} P_i\big(\mathbf{s}_i(t+1) \mid \mathbf{s}_{\mathcal{N}_i}(t), \mathbf{a}_i(t)\big),$$

where $\mathbf{s}_{\mathcal{N}_i} := (\mathbf{s}_j)_{j \in \mathcal{N}_i}$.

Agent i's next state depends only on its neighborhood states and its own action

Problem Setup: Networked MARL

- **Localized policy**: $\pi_i^{\theta_i}(\mathbf{a}_i \mid \mathbf{s}_{\mathcal{N}_i})$
- Joint policy $\pi^\theta(\mathbf{a} \mid \mathbf{s}) := \prod_{i=1}^n \pi_i^{\theta_i}(\mathbf{a}_i \mid \mathbf{s}_{\mathcal{N}_i})$
- Each agent receives a local reward $r_i(\mathbf{s}_i, \mathbf{a}_i)$
- Global reward

$$r(\mathbf{s}, \mathbf{a}) := \frac{1}{n} \sum_{i=1}^{n} r_i(\mathbf{s}_i, \mathbf{a}_i)$$

Goal: learn the optimal policy π^θ

$$\max_{\theta \in \Theta} J(\theta) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \cdot r(\mathbf{s}(t), \mathbf{a}(t))\right]. \tag{2}$$

Networked MARL under Domain Generalization

Inspired by *single-agent* domain generalization [HFL$^+$22][1]
Domain-specific dynamics for each agent i: $\mathbf{s}_i = (s_{i,1}, \ldots, s_{i,d_i^{\mathbf{s}}})$

$$s_{i,j}(t+1) = f_{i,j}\left(\mathbf{c}^{\mathbf{s}}_{\mathcal{N}_i,j} \odot \mathbf{s}_{\mathcal{N}_i}(t), \mathbf{c}^{\mathbf{a}}_{i,j} \odot \mathbf{a}_i(t), \mathbf{c}^{\boldsymbol{\omega}}_{i,j} \odot \boldsymbol{\omega}_i, \epsilon^{\mathbf{s}}_{i,j}(t)\right), \tag{3}$$

- $\boldsymbol{\omega}_i$: domain factor, encodes environment changes/shifts
- $\mathbf{c}$: causal masks, **invariant** across domains
- $f_{i,j}$: transition function, **invariant**
- $\epsilon_{i,j}$: noise

Train on M i.i.d. source domains $\langle \mathcal{E}_1, \ldots, \mathcal{E}_M \rangle$, adapt to a target domain $\mathcal{E}_{M+1}$

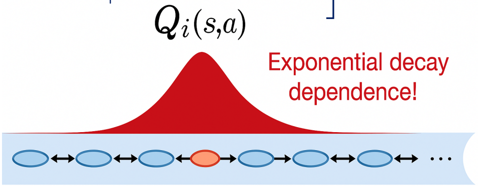
[1] Biwei Huang, et al. "Adarl: What, where, and how to adapt in transfer reinforcement learning". ICLR 2022.

Challenge 1: Scalability

- Curse of dimensionality $\#(\mathbf{s}, \mathbf{a}) = |\mathcal{S}_i|^n \times |\mathcal{A}_i|^n$
- Local Q-function depends on the global $(\mathbf{s}, \mathbf{a}) = (\mathbf{s}_1, \ldots, \mathbf{s}_n; \mathbf{a}_1, \ldots, \mathbf{a}_n)$

$$Q_i^{\pi}(\mathbf{s}, \mathbf{a}) := \mathbb{E}_{\mathbf{a}(t) \sim \pi^{\theta}(\cdot|\mathbf{s}(t))} \left[\sum_{t=0}^{\infty} \gamma^t r_i(\mathbf{s}_i(t), \mathbf{a}_i(t)) \,\middle|\, \mathbf{s}(0) = \mathbf{s},\ \mathbf{a}(0) = \mathbf{a} \right]$$

- **Exponential decay** property [QWL22][2]
- $\mathcal{N}_i^{\kappa}$: κ-hop neighborhood of agent i
- κ**-hop truncation** as efficient approximation



$$\left| Q_i^{\pi}(\mathbf{s}_{\mathcal{N}_i^{\kappa}}, \mathbf{a}_{\mathcal{N}_i^{\kappa}}) - Q_i^{\pi}(\mathbf{s}, \mathbf{a}) \right| \leq \mathcal{O}(\gamma^{\kappa})$$

---

[2] Qu, Guannan, Adam Wierman, and Na Li. "Scalable reinforcement learning for multiagent networked systems." *Operations Research*, 70(6): 3601–3628, 2022.

Challenge 1: Scalability

Without truncation,

- dimension $= \dim(\mathcal{S} \times \mathcal{A}) = \sum_{j \in \mathcal{N}} (d_j^{\mathbf{s}} + d_j^{\mathbf{a}})$, size **exponential** in n!

κ-hop truncation

- Input: $(\mathbf{s}_{\mathcal{N}_i^\kappa}, \mathbf{a}_{\mathcal{N}_i^\kappa})$
- dimension $= \dim(\mathcal{S}_{\mathcal{N}_i^\kappa} \times \mathcal{A}_{\mathcal{N}_i^\kappa}) = \sum_{j \in \mathcal{N}_i^\kappa} (d_j^{\mathbf{s}} + d_i^{\mathbf{a}})$, still **LINEAR** in neighborhood!

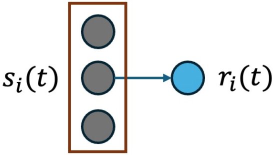
Our contribution: **A**pproximately **C**ompact **R**epresentations (ACRs)

Further reduce to subsets of $\mathbf{s}_{\mathcal{N}_i^\kappa}$ while maintaining approximation accuracy

More Scalable via Approximately Compact Representations

Core idea: Identify minimal variables within $\mathbf{s}_{\mathcal{N}_i^\kappa}$ that influence κ-step rewards

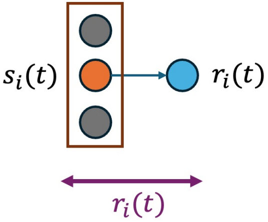
Recursion



More Scalable via ACRs

Core idea: Identify minimal variables within $\mathbf{s}_{\mathcal{N}_i^\kappa}$ that influence κ-step rewards

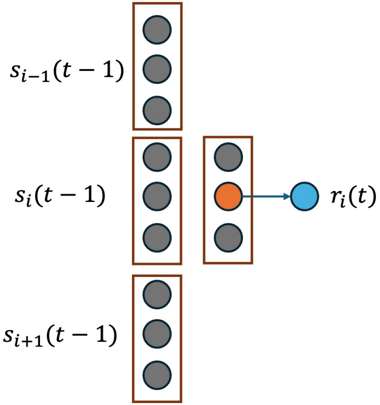
Recursion



More Scalable via ACRs

Core idea: Identify minimal variables within $\mathbf{s}_{\mathcal{N}_i^\kappa}$ that influence κ-step rewards

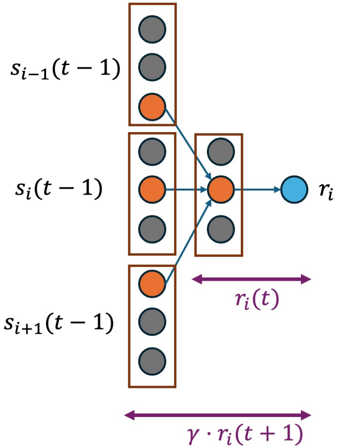
Recursion



More Scalable via ACRs

Core idea: Identify minimal variables within $\mathbf{s}_{\mathcal{N}_i^\kappa}$ that influence κ-step rewards

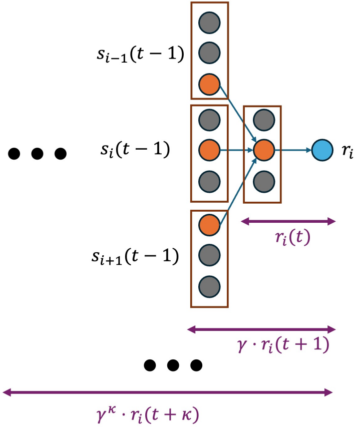
Recursion



More Scalable via ACRs

Core idea: Identify minimal variables within $\mathbf{s}_{\mathcal{N}_i^\kappa}$ that influence κ-step rewards

Output: $\mathbf{s}^{\circ}_{\mathcal{N}^\kappa}$



GSAC Algorithm Overview

Four Sequential Phases

- **Phase 1: Causal Discovery and Domain Estimation**
  - — Estimate causal masks $\mathbf{c}_i$ and domain factors $\hat{\boldsymbol{\omega}}$ per domain
- **Phase 2: ACR Construction**
  - — Build ACR using causal masks
- **Phase 3: Meta Actor-Critic Learning**
  - — Train domain-shared policy $\pi_i^{\theta_i}$ across M source domains
  - — Condition on ACR inputs: $(\mathbf{s}^{\circ}_{\mathcal{N}_i}, \hat{\boldsymbol{\omega}}^{\circ}_{\mathcal{N}_i})$
  - — Output: $\bar{\theta}_i$ for each agent i's policy
- **Phase 4: Fast Adaptation**
  - — Collect T_a trajectories in new domain to estimate $\hat{\boldsymbol{\omega}}^{M+1}$
  - — Deploy $\pi_i^{\bar{\theta}_i}(\cdot|\mathbf{s}^{\circ}_{\mathcal{N}_i}, \hat{\boldsymbol{\omega}}^{M+1}_{\mathcal{N}_i})$, no tuning of θ

Meta Actor-Critic

Outer loop (iteration $k = 1, 2, \cdots, K$)

- Sample domain index
- **Inner loop** (iteration $t = 1, 2, \cdots, T$)
  — **Critic** update: TD learning on ACR inputs

$$\hat{Q}_i^t(\mathbf{s}^{\circ}_{\mathcal{N}_i^\kappa}, \mathbf{a}_{\mathcal{N}_i^\kappa}, \hat{\boldsymbol{\omega}}^{\circ}_{\mathcal{N}_i^\kappa}) \leftarrow (1 - \alpha_{t-1})\hat{Q}_i^{t-1} + \alpha_{t-1}\left[r_i(t) + \gamma \hat{Q}_i^{t-1}(\text{next})\right]$$

- **Actor** update: policy gradient

$$\hat{g}_i(k) \leftarrow \sum_{t=0}^{T} \gamma^t \cdot \frac{1}{n} \sum_{j \in \mathcal{N}_i^\kappa} \hat{Q}_j^T(\text{ACR}) \cdot \nabla_{\theta_i} \log \pi_i^{\theta_i(k)}$$

$$\theta_i(k+1) \leftarrow \theta_i(k) + \eta_k \cdot \hat{g}_i(k)$$

Output: $\bar{\theta}_i = \theta_i(K)$
Key: All computation uses compact ACR inputs!

Convergence

Theorem (Critic error bound)

With high probability, after T inner iterations:

$$|Q_i(\mathbf{s},\mathbf{a},\omega) - \hat{Q}_i^T(\mathbf{s}_{\mathcal{N}_i^\kappa},\mathbf{a}_{\mathcal{N}_i^\kappa},\hat{\omega}_{\mathcal{N}_i^\kappa})| \leq \mathcal{O}\left(\underbrace{\frac{1}{\sqrt{T+t_0}}}_{\text{TD error}} + \underbrace{\frac{2c\rho^{\kappa+1}}{(1-\gamma)^2}}_{\text{ACR error}} + \underbrace{\sqrt{\frac{1}{T_e}}}_{\text{Domain estimation error}}\right),$$

Theorem (Policy gradient convergence)

$$\frac{\sum_{k=0}^{K-1}\eta_k\|\nabla J(\theta(k))\|^2}{\sum_{k=0}^{K-1}\eta_k} \leq \tilde{\mathcal{O}}\left(\underbrace{\frac{1}{\sqrt{K+1}}}_{\text{optimization error}} + \rho^{\kappa+1} + \sqrt{\frac{1}{T_e}} + \underbrace{\sqrt{\frac{1}{M}}}_{\text{domain generalization}}\right).$$

Adaptation Guarantee

Theorem (Adaptation gap)

For new domain, after collecting T_a adaptation trajectories:

$$J_{source} - J(\pi^{\theta^K}(\cdot|\hat{\omega}^{M+1})) \geq \Theta\left(\frac{1}{T_a}\right).$$

- Meta-training on M domains provides good initialization/zero-shot performance
- The expected return is close to the meta-policy's average performance on source domains
- Adaptation gap decreases at a rate of $\Theta\left(\frac{1}{T_a}\right)$

Benefits of ACR

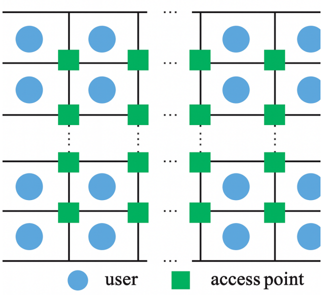
Method	Input	Dimension	Approx. Error	Size
Full State	$\mathbf{s}$	$\sum_{j=1}^{n} d_j^s$	0	all agents ✗
Truncation [QWL22]	$\mathbf{s}_{\mathcal{N}_i^\kappa}$	$\sum_{j \in \mathcal{N}_i^\kappa} d_j^s$	$\mathcal{O}(\gamma^\kappa)$	κ-hop neighbors ▲
GSAC (ACR)	$\mathbf{s}^{\circ}_{\mathcal{N}_i^\kappa}$	$< \sum_{j \in \mathcal{N}_i^\kappa} d_j^s$	$\mathcal{O}(\gamma^\kappa)$	**Much Lower ✓**

- Faster convergence
- Lower memory
- Better generalization

Experimental Setup

- **Benchmark**: Wireless communication network [Zoc19][a]
- n users, each with packet queue d_i
- Packet arrival $\sim$ Ber(p_i)
- $\mathbf{s}_i$: que status
- $\mathbf{a}_i$ = AP/null
- Success if no collision, Reward + 1

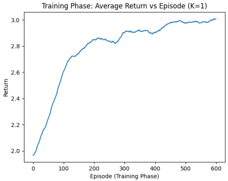
[a] Zocca, Alessandro. "Temporal starvation in multi-channel csma networks: an analytical framework." ACM SIGMETRICS Performance Evaluation Review (2019).



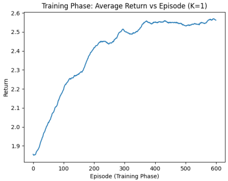
- **Interaction graph**: Users sharing APs are neighbors

Training Performance

- 3 source domains: $p \in \{0.2, 0.5, 0.8\}$
- Consistent improvement across all grid sizes
- Scalability: 3×3 (9 agents) $\rightarrow$ 4×4 (16 agents)



(a) grid size 3

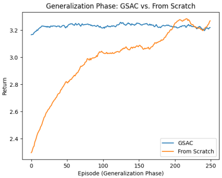


(b) grid size 4

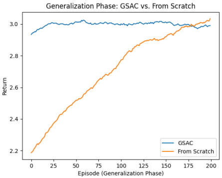
Figure: GSAC Training for different grid sizes.

Adaptation Performance

GSAC vs. Learning From Scratch (LFS) for different settings

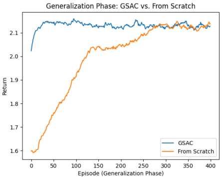


(a) grid size 3

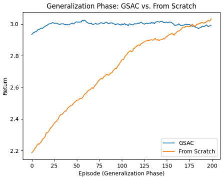


(b) grid size 4

Figure: Adaptation performance comparison for different grid sizes.



(a) $p_{\text{target}} = 0.4$



(b) $p_{\text{target}} = 0.6$

Figure: Adaptation performance comparison for different target domains

Comparison to Prior Work

Method	Scalability	Generalization	Theory
SAC [QWL22]	Truncation ✓	✗	Convergence ✓
AdaRL [HFL+22]	Single-agent ✗	✓	Causality ✓
GSAC (Ours)	**ACRs** ✓	✓	**All phases** ✓

- First to combine causality with networked MARL
- First end-to-end guarantees for generalization + scalability
- ACR framework

Future Research Directions

- **Continuous spaces**: continuous state/action space, function approximation w/ ACRs
- **Large-scale networked systems**:
  - — Traffic networks
  - — power grids
  - — robotics swarms
- **Partial observability**

Key Takeaways

$$\text{Causality + Locality} \implies \text{Scalable \& Generalizable Networked MARL}$$

- **GSAC**: First provably generalizable + scalable MARL for networked systems
- Technical innovation
  - — ACRs via causal structure
  - — Meta actor-critic with domain-conditioned policies
- Theoretical guarantees
  - — Approximation error
  - — Convergence
  - — Adaptation
- Empirical validation
  - — Scalability
  - — Fast adaptation

Thank you!

References I

- Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, and Kun Zhang, *Adarl: What, where, and how to adapt in transfer reinforcement learning*, International Conference on Learning Representations, 2022.
- Guannan Qu, Adam Wierman, and Na Li, *Scalable reinforcement learning for multiagent networked systems*, Operations Research **70** (2022), no. 6, 3601–3628.
- Alessandro Zocca, *Temporal starvation in multi-channel csma networks: an analytical framework*, ACM SIGMETRICS Performance Evaluation Review **46** (2019), no. 3, 52–53.