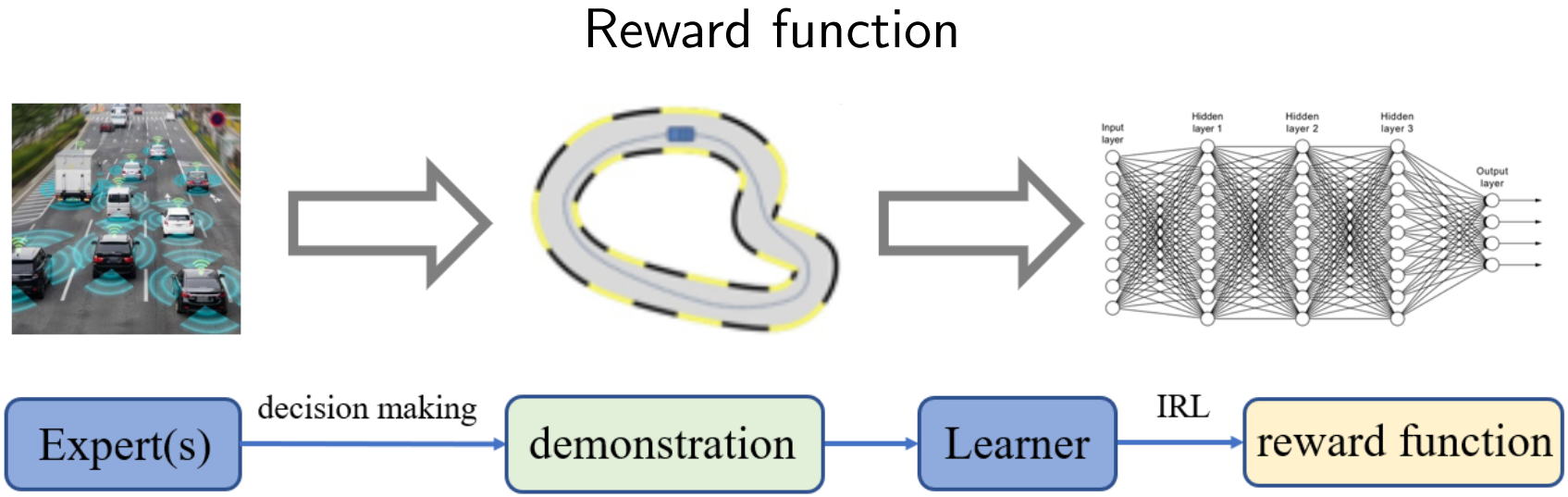
Distributed Inverse Constrained Reinforcement Learning (D-ICRL) for Multi-agent Systems (MASs)

Shicheng Liu & Minghui Zhu

The Pennsylvania State University

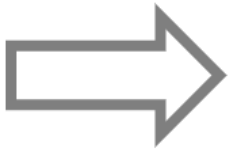
Neural Information Processing Systems 2022

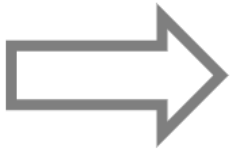
Distributed inverse constrained reinforcement learning

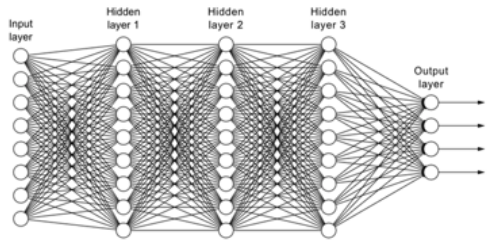


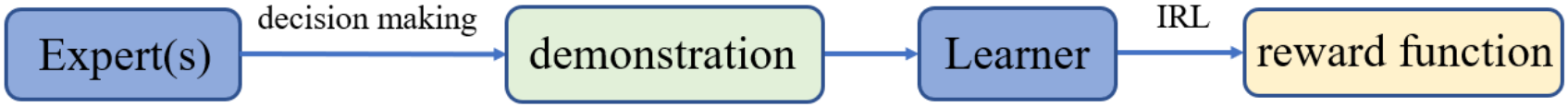
Distributed inverse constrained reinforcement learning

Reward function







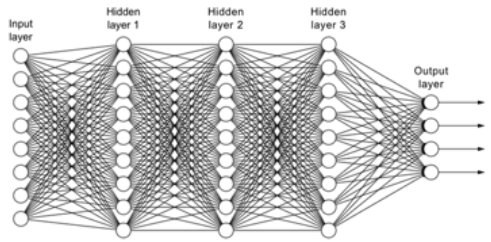


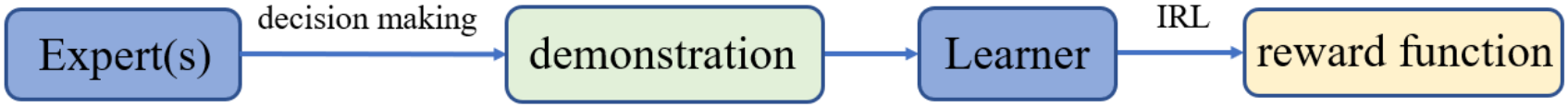
Constraints

Distributed demonstrations

Distributed inverse constrained reinforcement learning

Reward function



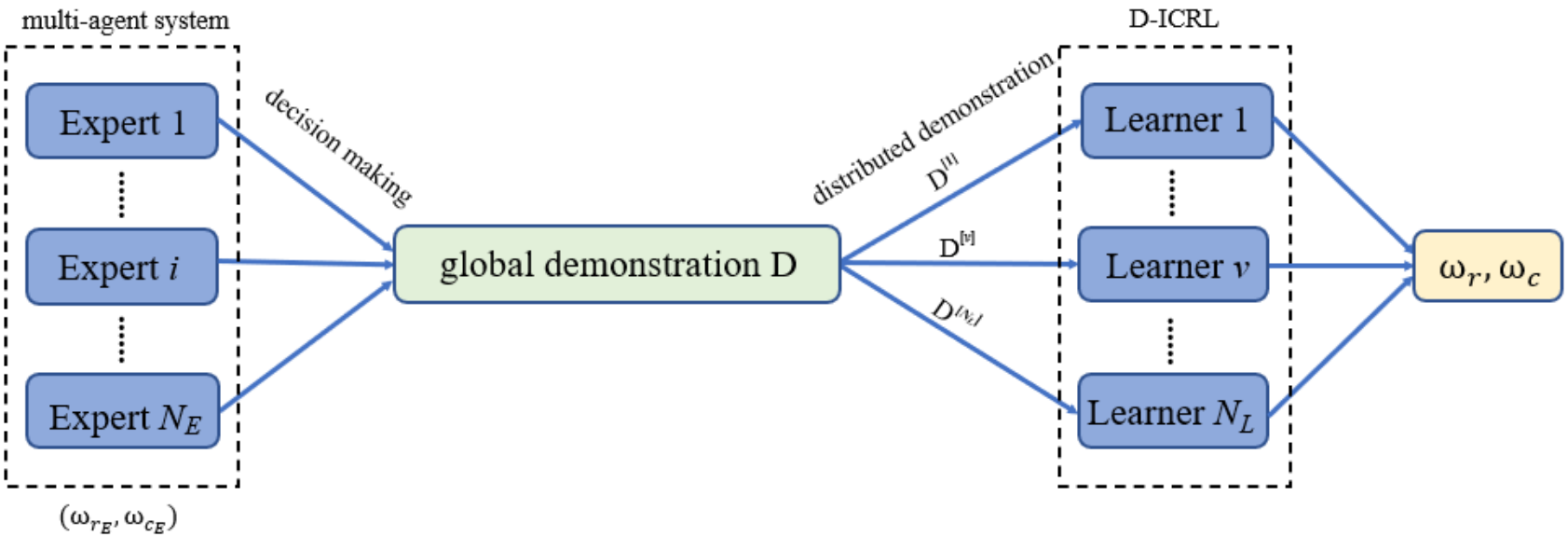


Constraints

Distributed demonstrations

D-ICRL: Solve the three challenges at once.

Model: Multiple experts & multiple learners



N_E cooperative experts: $fr_E = !^{>}_{r_E}\ _r; c_E = !^{>}_{c_E}\ _c g\ )\ D = fD^{[v]}g^{N_L}_{v=1}$

N_L collaborative learners: $fD^{[v]} = f\ ^j g^{m^{[v]}}_{j=1}; \ _r; \ _c g\ )\ f!_r; !_c g$

Distributed bi-level optimization formulation

Distributed bi-level optimization formulation

$$\max_{!_c 2 \Omega_c} \quad F(!_c; \quad (!_c)) = \sum_{v=1}^{N_L} F^{[v]}(!_c; \quad (!_c)); \qquad \text{(outer level)}$$

$$\text{s.t.} \quad (!_c) = \arg\min \sum_{v=1}^{N_L} m^{[v]} G^{[v]}(\ ; !_c): \quad \text{(inner level)}$$

- The outer level learns constraints by maximizing the log likelihood $\sum_{v=1}^{N_L} F^{[v]}$ of the demonstrations.
- Given a constraint estimate $!_c$, the inner level learns the corresponding reward function and policy by minimizing the dual function $\sum_{v=1}^{N_L} m^{[v]} G^{[v]}$ of maximum causal entropy (MCE).

A perspective of double-loop learning

Double-loop communication: sharing reward and cost function parameters

Inner communication (faster): $W^{[vv^0]}(k)$ and $N^{[v]}(k)$.

Outer communication (slower): $W^{[vv^0]}(n)$ and $N^{[v]}(n)$.

Inner process

Receives $^{[v^0]}(k)$ from neighbor$v^0 2$ N $^{[v]}(k)$.

$$^{[v]}(k+1) = \sum_{v^0=1}^{N_L} W^{[vv^0]}(k) \; ^{[v^0]}(k) \qquad (k)m^{[v]}r \quad G^{[v]}(\; ^{[v]}(k); ! \; _c)$$

RunsK iterations

Outer process

Di culties

Local gradientr $F^{[v]}(!_c; \quad (!_c))$ inaccessible.
global gradientr $F(!_c; \quad (!_c))$ inaccessible.
$F(!_c; \quad (!_c))$ non-convex.

Outer process

Difficulties

- Local gradient $\nabla F^{[v]}(\omega_c; \theta(\omega_c))$ inaccessible.
- global gradient $\nabla F(\omega_c; \theta(\omega_c))$ inaccessible.
- $F(\omega_c; \theta(\omega_c))$ non-convex.

Our solutions

- Local gradient approximation $\nabla F^{[v]}(\omega_c; \theta^{[v]}(\omega_c))$.
- Global gradient tracking $\gamma^{[v]}(n) = \sum_{v=1}^{N_L} W^{[vv']}(n)\gamma^{[v']}(n-1) + \nabla F^{[v]}(\omega_c^{[v]}(n); \theta^{[v]}(\omega_c^{[v]}(n))) - \nabla F^{[v]}(\omega_c^{[v]}(n-1); \theta^{[v]}(\omega_c^{[v]}(n-1)))$.
- Successive convex approximation $\omega_c^{[v]}(n) = \text{Project}_{\mathcal{C}}(\omega_c^{[v]}(n) + N_L\gamma^{[v]}(n))$.

Outer process

Difficulties

- Local gradient $\nabla F^{[v]}(\theta_c; \quad (\theta_c))$ inaccessible.
- global gradient $\nabla F(\theta_c; \quad (\theta_c))$ inaccessible.
- $F(\theta_c; \quad (\theta_c))$ non-convex.

Our solutions

- Local gradient approximation $\bar{\nabla} F^{[v]}(\theta_c; {}^{-[v]}(\theta_c))$.
- Global gradient tracking $\bar{\nabla}^{[v]}(n) = \sum_{v=1}^{N_L} W^{[vv^0]}(n)\, \bar{\nabla}^{[v^0]}(n \quad 1)$
  $+ \bar{\nabla} F^{[v]}(\theta_c^{[v]}(n); {}^{-[v]}(\theta_c^{[v]}(n))) \quad \bar{\nabla} F^{[v]}(\theta_c^{[v]}(n \quad 1); {}^{-[v]}(\theta_c^{[v]}(n \quad 1)))$.
- Successive convex approximation
  $\tau_c^{[v]}(n) = \text{Project}_{\Omega_c}(\theta_c^{[v]}(n) + N_L\, \bar{\nabla}^{[v]}(n))$.

Update rule: $\theta_c^{[v]}(n+1) = \sum_{v^0=1}^{N_L} \quad (n)\, \tau_c^{[v^0]}(n) + (1 \quad (n))\, \theta_c^{[v^0]}(n)$

Theoretical guarantee

Convergence rate of inner problem

Suppose $(k) = \frac{}{k+1}$ where is a positive constant, it holds for any learner v and $!_c \ 2 \ \Omega_c$ that

$$jj^{-[v]}(!_c) \qquad (!_c)jj \quad O(p\frac{1}{\log K})$$

Asymptotic convergence of outer problem

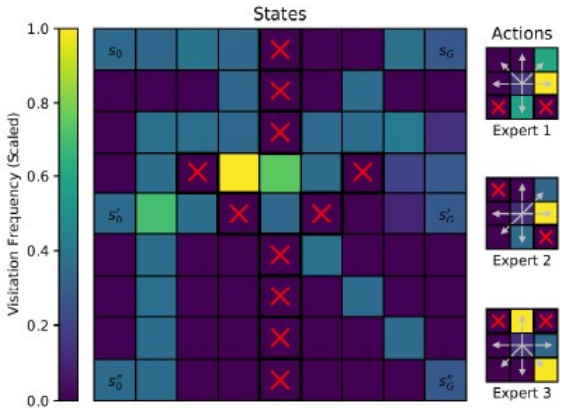
Suppose $(n) \ 2 \ (0;1)$, $\sum_{n=0}^{1} (n) = +1$, and $\sum_{n=0}^{1} (n)^2 < +1$, it holds for any learner v that

$$\lim_{n \ ! \ 1} \max_{v;v^{\theta}} jj\, !_c^{[v]}(n) \quad !_c^{[v^{\theta}]}(n)jj = 0;$$

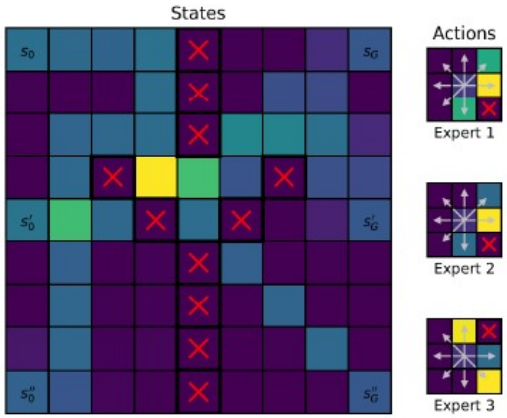
$$\limsup_{n \ ! \ 1}(r\, F(!_c^{[v]}(n); \quad (!_c^{[v]}(n)))^{>}(!_c \quad !_c^{[v]}(n)) \quad O(p\frac{1}{\log K})$$

Simulations

Discrete environment

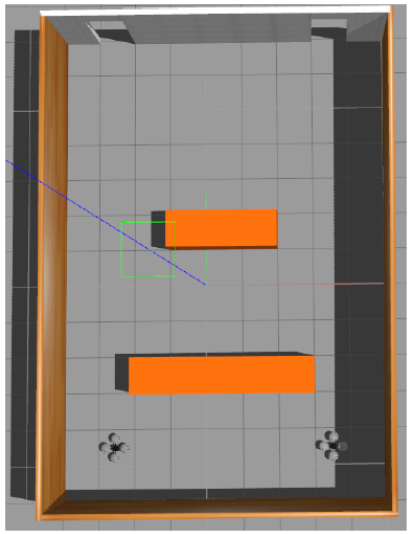


Ground truth environment

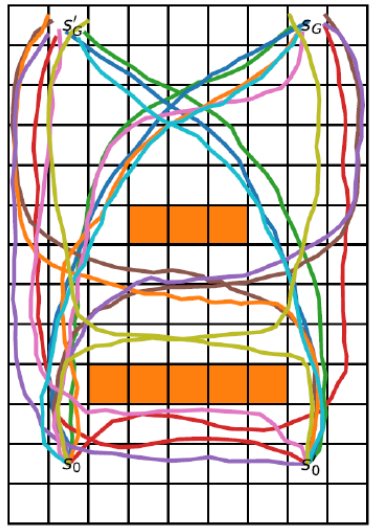


Learned environment

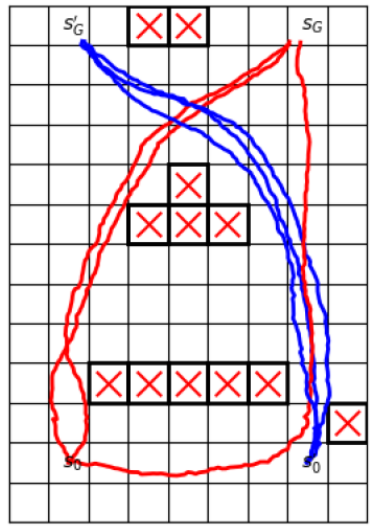
Continuous environment



Simulator environment



Demonstrated trajectories



Learned trajectories

D-ICRL can successfully imitate the experts and recover the constraints.

Conclusion

Solve three challenges at once: Reward function, constraints, and distributed data.

Formulate as a distributed bi-level optimization problem.

D-ICRL: Theoretical framework effective to continuous and discrete environments empirically.