

Abstract

- This work presents a model-free inverse reinforcement learning (IRL) framework for robotic systems, enabling the recovery of control objectives from expert demonstrations.
- A model-based inverse optimal control formulation is developed to recover quadratic cost functions using exact Riccati-based gradient information.
- A data-driven off-policy inverse Q-learning method is introduced to infer equivalent feedback controllers using only observed state–action data.
- Results on leader–follower robotic manipulators demonstrate effective objective recovery with interpretable controllers and satisfactory control performance.

Objectives

Inferring expert intent from demonstrations is a key challenge in learning-based control.

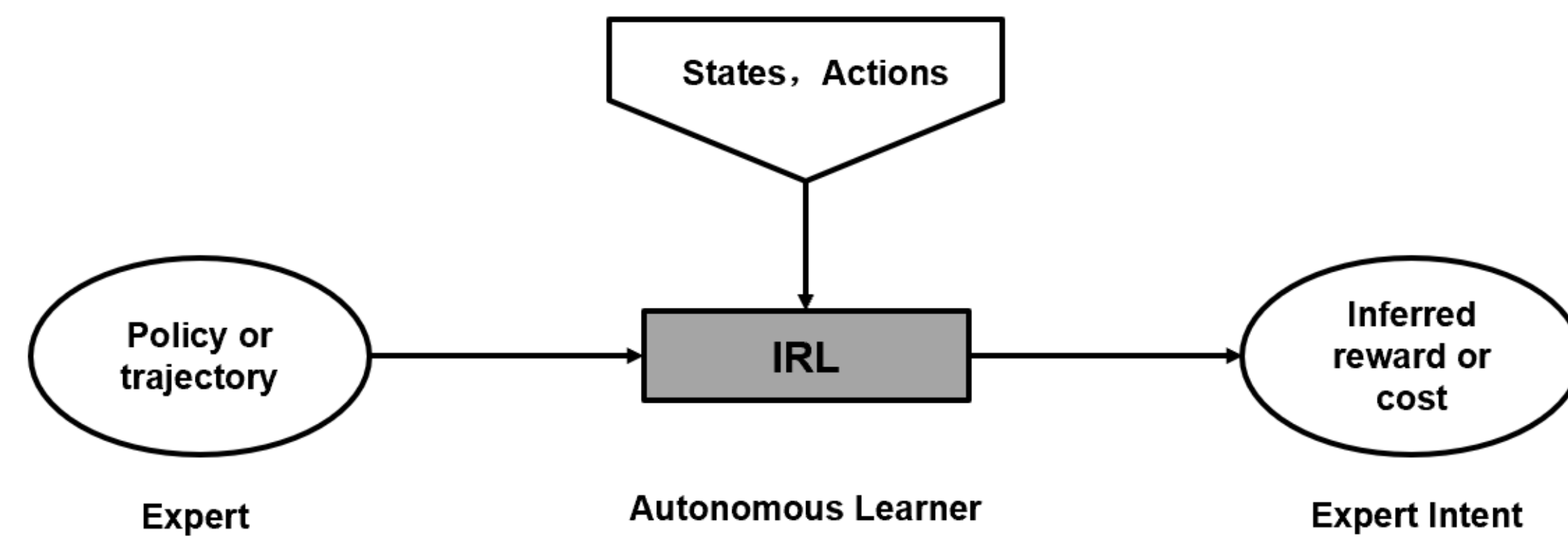


Fig. 1 IRL Framework

This work aims to integrate a control system with IRL to learn a reward function from demonstrations that captures the task objective and guides the manipulator's behavior.

$$x_{k+1} = Ax_k + Bu_k$$

$$J = \sum_{k=0}^{\infty} (x_k^T Q x_k + u_k^T R u_k)$$

$$u_k^* = -K^* x_k$$

$$K^* = \text{DARE}(A, B, Q, R)$$

$$E(Q) = \|K(Q) - K^*\|_F^2$$

IRL for Optimal Control

Given expert demonstrations generated by an unknown quadratic cost Q , infer the equivalent cost such that the resulting control policy matches the expert's control policy.

Model-based IRL by exact gradient

- Discrete Algebraic Riccati Equation (DARE):**

$$P = A^T P A - A^T P B (R + B^T P B)^{-1} B^T P A + Q$$

$$\frac{\partial P}{\partial Q_{ij}} = A^T \frac{\partial P}{\partial Q_{ij}} A - A^T \frac{\partial P}{\partial Q_{ij}} B (R + B^T P B)^{-1} B^T P A - A^T P B (R + B^T P B)^{-1} B^T \frac{\partial P}{\partial Q_{ij}} A + \mathcal{E}_{ij}$$
- Discrete Lyapunov Equation:**

$$(A - BK)^T \frac{\partial P}{\partial Q_{ij}} (A - BK) - \frac{\partial P}{\partial Q_{ij}} + \mathcal{E}_{ij} = 0$$
- Control Gain Sensitivity^[1]:**

$$K = (R + B^T P B)^{-1} B^T P A$$

$$\frac{\partial K}{\partial Q_{ij}} = (R + B^T P B)^{-1} B^T \frac{\partial P}{\partial Q_{ij}} (A - BK)$$

Model-free IRL by inverse Q-learning

Inverse Q-learning

- Data-Driven H Matrix Estimation^[2]**

$$H^{i+1} = \begin{bmatrix} H_{xx}^{i+1} & H_{xu}^{i+1} \\ H_{ux}^{i+1} & H_{uu}^{i+1} \end{bmatrix} = \begin{bmatrix} Q^i + A^T P^{i+1} A & A^T P^{i+1} B \\ B^T P^{i+1} A & R + B^T P^{i+1} B \end{bmatrix}$$

$$Q(x_T(k), u_T(k), P^{i+1}) = \begin{bmatrix} x_T(k) \\ u_T(k) \end{bmatrix}^T H^{i+1} \begin{bmatrix} x_T(k) \\ u_T(k) \end{bmatrix}$$

$$K^{i+1} = (H_{uu}^{i+1})^{-1} H_{ux}^{i+1}$$
- Simultaneous Perturbation Stochastic Approximation (SPSA)^[3]**

$$\theta = [Q_{11}, Q_{12}, \dots, Q_{1n}, Q_{22}, Q_{23}, \dots, Q_{nn}]^T \in \mathbb{R}^p$$

$$\hat{g}^j = \frac{E(\theta^j + \epsilon \Delta^j) - E(\theta^j - \epsilon \Delta^j)}{2\epsilon} \begin{bmatrix} 1/\Delta_1^j \\ 1/\Delta_2^j \\ \vdots \\ 1/\Delta_p^j \end{bmatrix}$$

$$\Delta_i^j \in \{-1, +1\}$$

Algorithm 1 Model-Free IRL via Inverse Q-learning

- Input:** Expert samples $\{(x_k, u_k, x_{k+1}, u_{k+1})\}_{k=1}^N$, expert gain K^* , fixed $R \succ 0$, step size α , tolerance ϵ , finite-difference step ϵ_{fd}
- Output:** Estimated state cost matrix Q
- Construct regression.** For $k = 1, \dots, N$:

$$z_k = \begin{bmatrix} x_k \\ u_k \end{bmatrix}, \quad z_{k+1} = \begin{bmatrix} x_{k+1} \\ u_{k+1} \end{bmatrix},$$

$$\Phi[k, :] = (z_k \otimes z_k - z_{k+1} \otimes z_{k+1})^T, \quad \Theta[k] = x_k^T Q x_k + u_k^T R u_k.$$

- Initialize.** Choose $Q^0 \succeq 0$ and set $t = 0$.
- while** $\|K^t - K^*\|_F^2 > \epsilon$ **do**
- Least squares.**

$$\text{vec}(H^t) = (\Phi^T \Phi)^{-1} \Phi^T \Theta$$

- Policy evaluation.**

$$K^t = (H_{uu}^t)^{-1} H_{ux}^t$$

- Finite-difference gradient.** For each symmetric basis E_{ij} ($i \leq j$):

$$g_{ij} = \frac{\|K(Q^t + \epsilon_{fd} E_{ij}) - K^*\|_F^2 - \|K^t - K^*\|_F^2}{\epsilon_{fd}}, \quad G^t = \sum_{i \leq j} g_{ij} E_{ij}.$$

- Update.**

$$Q^{t+1} = Q^t - \alpha G^t, \quad t \leftarrow t + 1$$

- end while and return** Q^t

Simulation Results

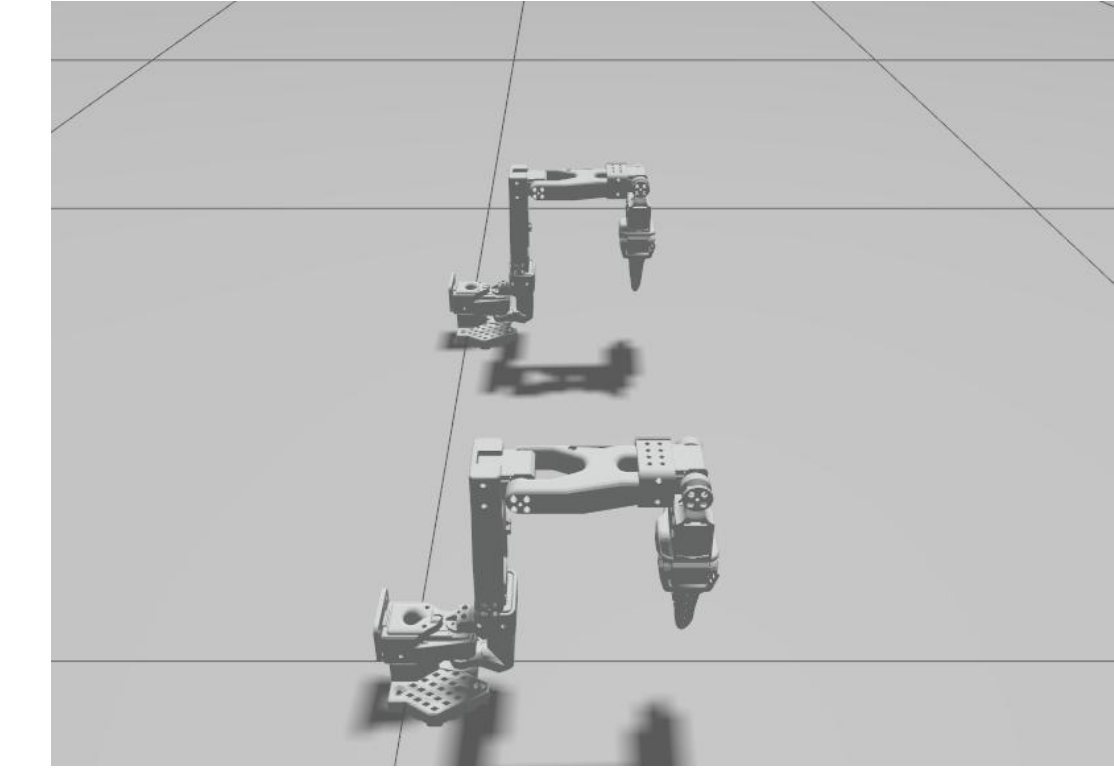


Fig. 2 Gazebo Environment

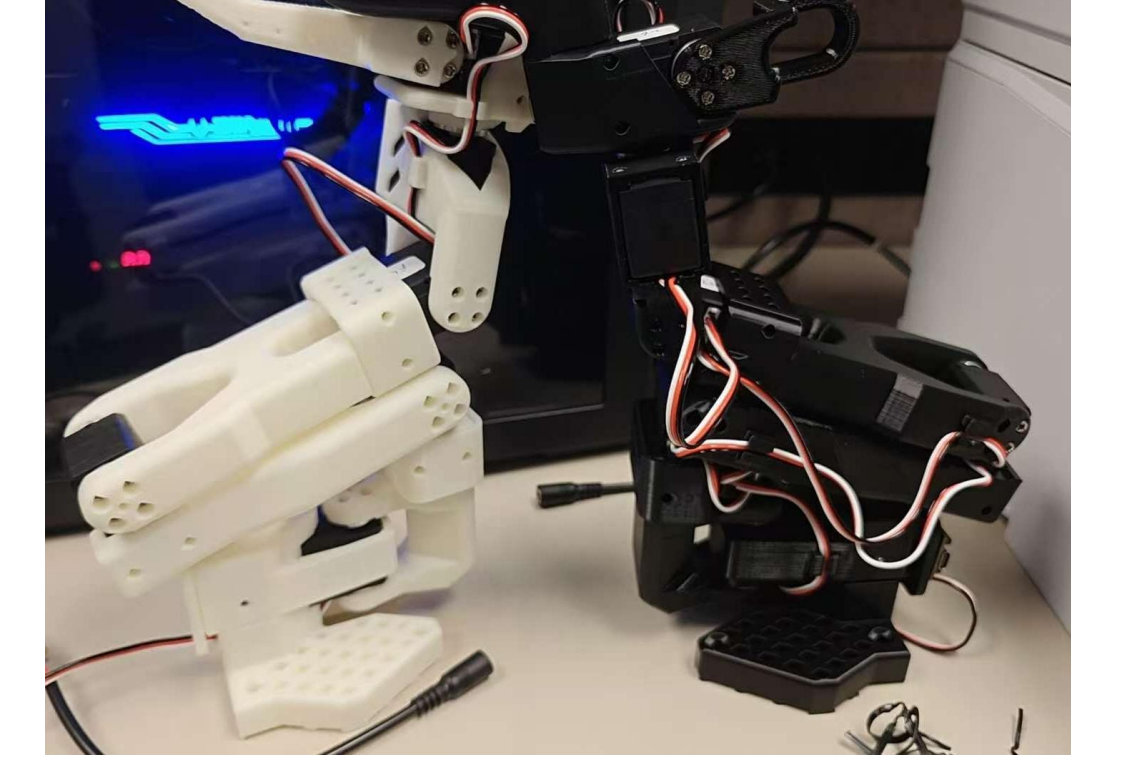
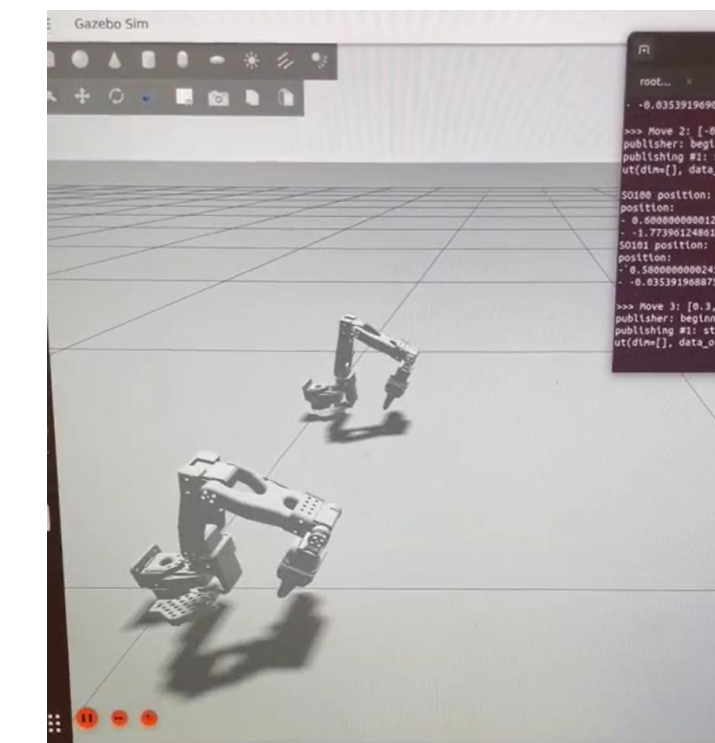
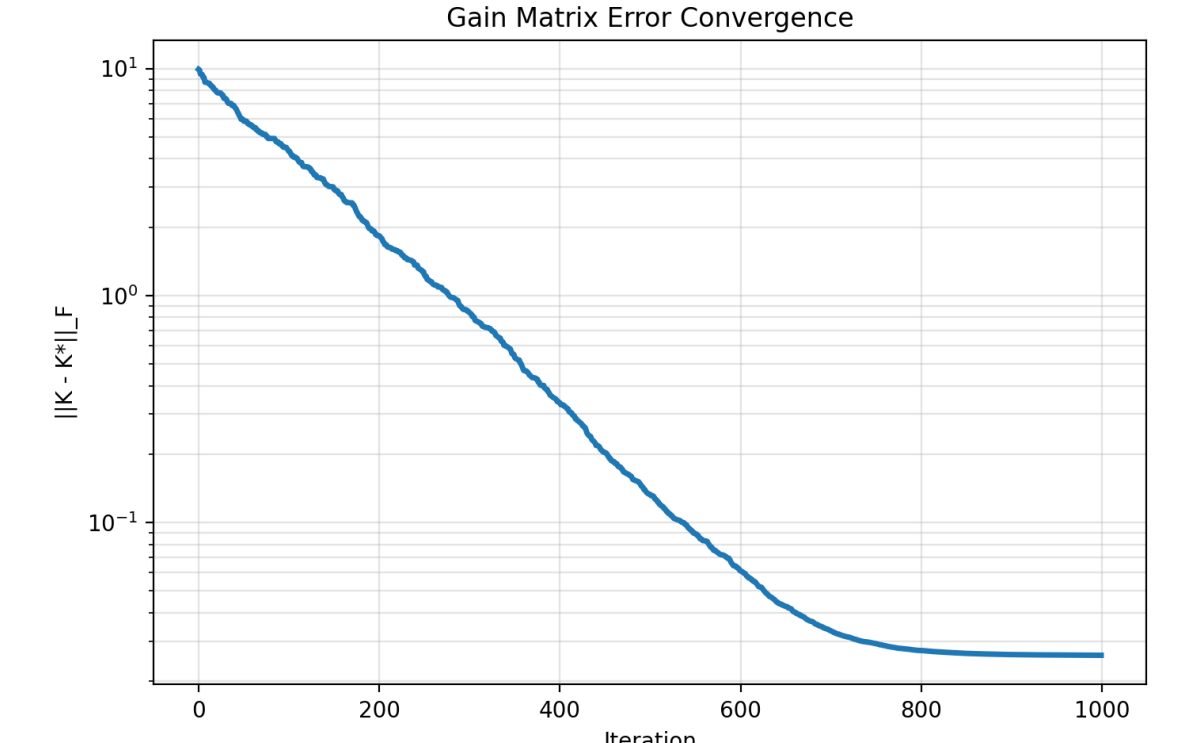


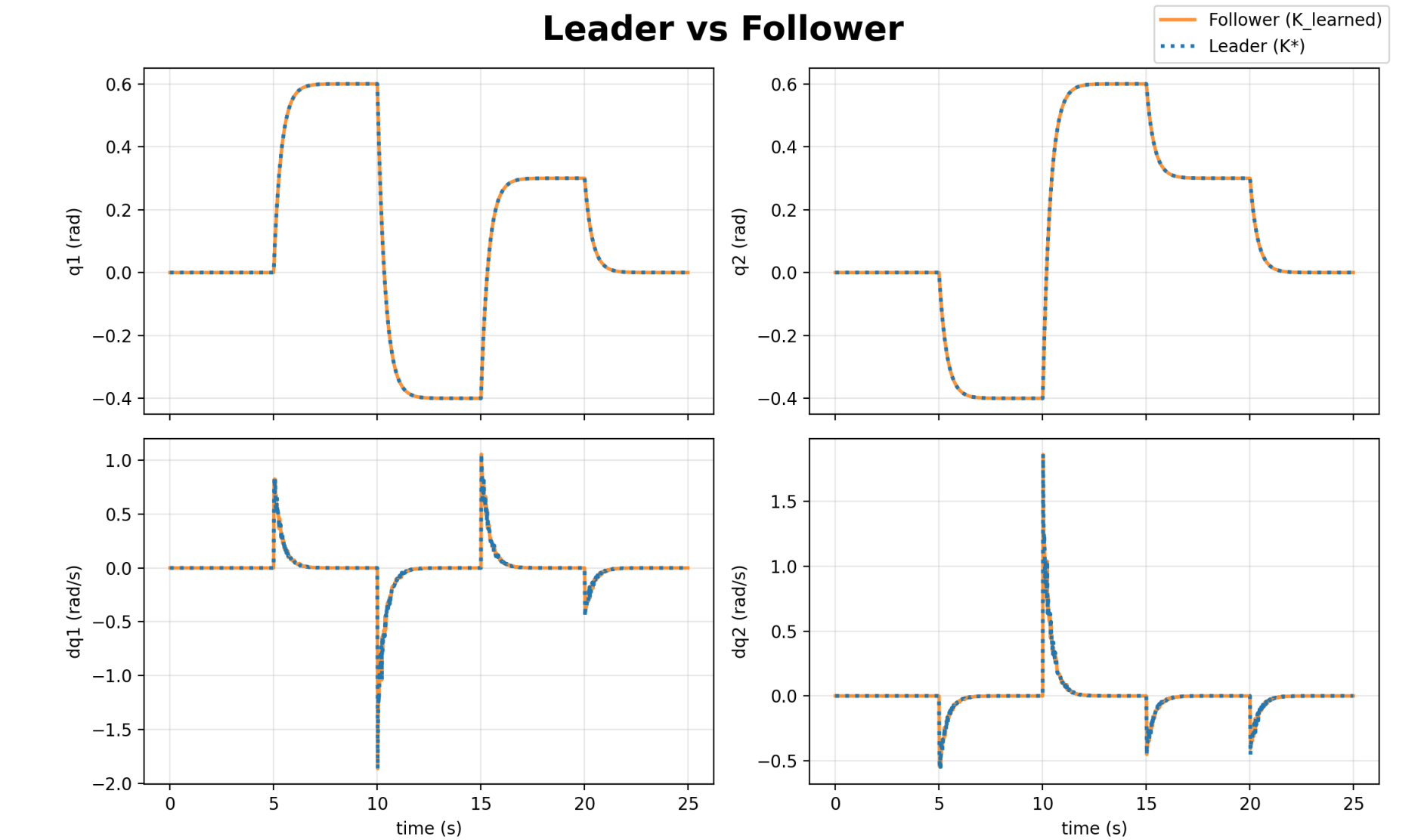
Fig. 3 Leader–Follower Robotic Arms



a) Joint Control Task in Gazebo



b) Convergence of the Control Gain of Follower



c) Joint Angle and Velocity of Shoulder and Elbow

Fig. 4 Simulation Results

Conclusion

Model-based and model-free inverse reinforcement learning methods are developed for discrete-time linear systems to recover quadratic control objectives from demonstrations. Results on leader–follower robotic arms demonstrate effective objective recovery and interpretable control behavior.

References

- [1] M. Cody Priess, Conway, R., Choi, J., Popovich, J.M. and Radcliffe, C.J. (2015). Solutions to the Inverse LQR Problem With Application to Biological Systems Analysis. IEEE Transactions on Control Systems and Technology, 23(2), pp.770–777
- [2] Lian, B., Xue, W., Xie, Y., Lewis, F.L. and Davoudi, A. (2023). Off-policy inverse Q-learning for discrete-time antagonistic unknown systems. Automatica, 155, p.111171.
- [3] Spall, J.C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE Transactions on Automatic Control, 37(3), pp.332–341.