Uncertainty-aware Constraint Inference in Inverse Constrained Reinforcement Learning

Sheng Xu, Guiliang Liu

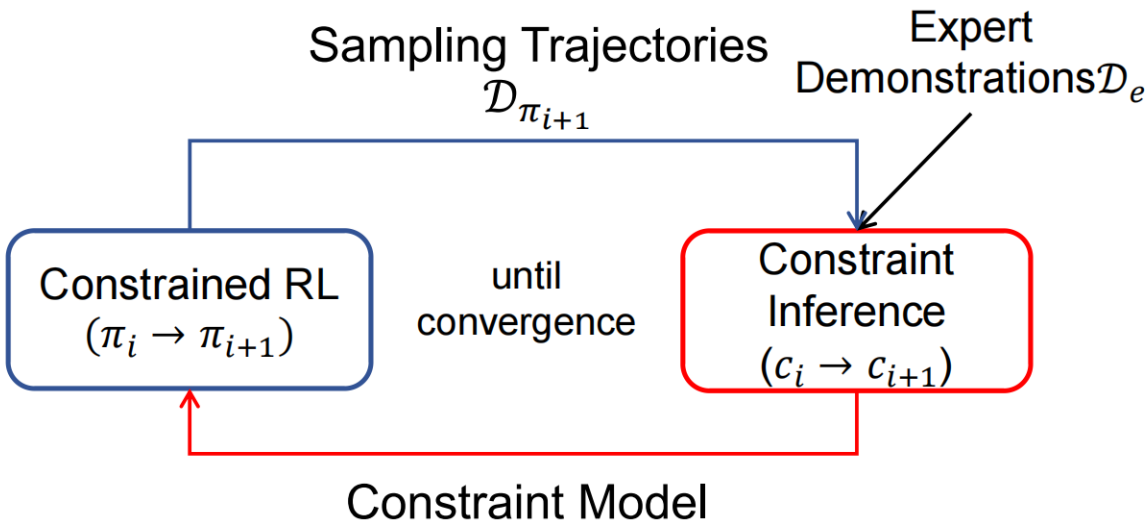
The Chinese University of Hong Kong, Shenzhen

Presented at ICLR 2024

Inverse Constrained Reinforcement Learning (ICRL)

ICRL [4] considers inferring the constraints respected by expert agents from their demonstrations and learning imitation policies that adhere to these constraints, until reproducing the expert demonstrations:

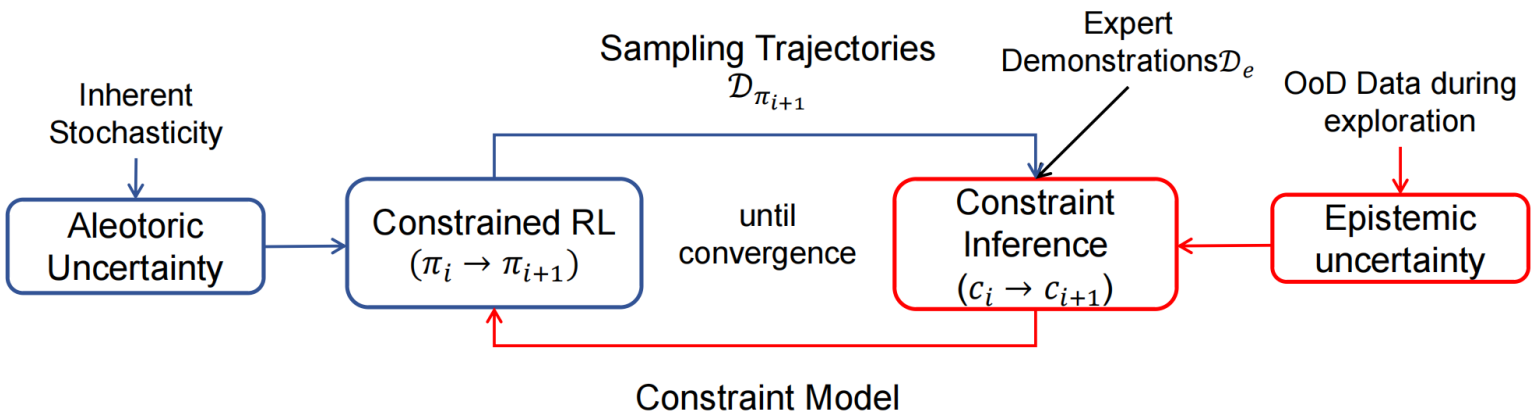
- *(Forward) Constrained Reinforcement Learning:* maximize the cumulative discounted rewards while respecting constraints
- *(Inverse) Constrain Inference:* infer the underlying constraints that best explain the expert behaviors



Inverse Constrained Reinforcement Learning (ICRL)

Existing ICRL works often neglected underlying uncertainties:

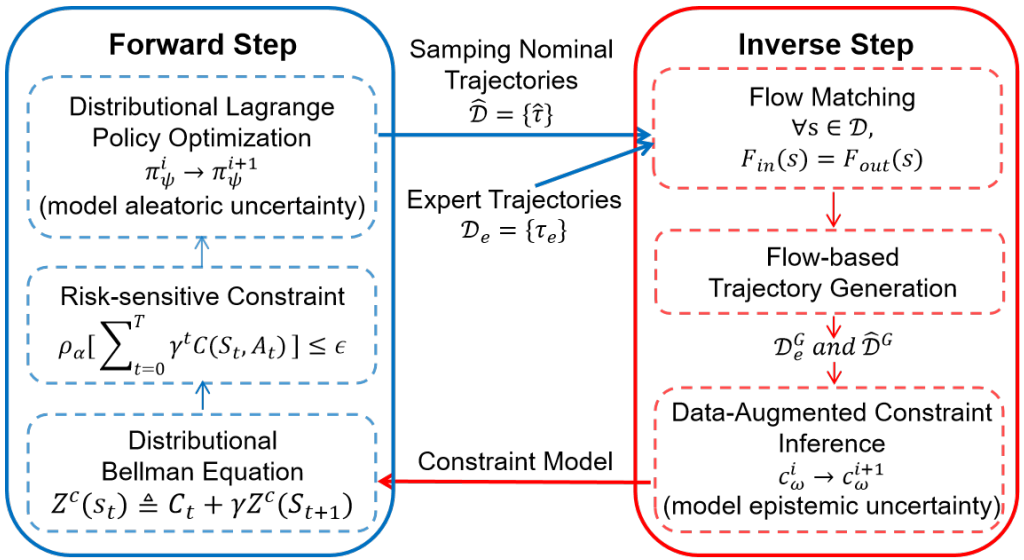
- **Aleatoric uncertainty** arises from the inherent stochasticity of environment dynamics -> leading to constraint-violating behaviors in imitation policies
- **Epistemic uncertainty** results from the model's limited knowledge of Out-of-Distribution (OoD) samples -> affecting the accuracy of step-wise cost predictions



Uncertainty-aware ICRL (UAICRL)

To handle both the aleatoric and epistemic uncertainties, we propose the Uncertainty-aware ICRL, including:

- **Policy Optimization with Risk-sensitive Constraints** by modeling the distribution of the cumulative costs
- **Data-augmented Constraint Inference** with flow-based trajectory generation for constraint inference from limited demonstrations



Forward Step: Policy Optimization with Risk-sensitive Constraints

To drive a risk-sensitive policy, the risk incurred by aleatoric uncertainty can be integrated into the costs distinctly from the reward optimization.

We formulate the trade-off between rewards and costs as a constrained optimization problem:

$$\arg\max_{\pi} \mathbb{E}_{\pi, p_{\mathcal{T}}, p_{\mathcal{R}}, \mu_0}\left[\sum_{t=0}^{T}\left(\gamma^t r(s_t, a_t) + \beta\gamma^t \mathcal{H}[\pi(a_t|s_t)]\right)\right] \text{ s.t. } \rho_\alpha\left[\sum_{t=0}^{T}\gamma^t C(S_t, A_t)\right] \leq \epsilon$$

where $\mathcal{H}[\pi(a_t|s_t)]$ refers to the causal entropy [6] and $C(\cdot)$ denotes the random variable of state-action cost.

Forward Step: Policy Optimization with Risk-sensitive Constraints

To model the distribution of the cumulative costs, we utilize the distributional Bellman equation [1]:

$$Z_\theta^c(s) = \int_{a\in\mathcal{A}} \pi(a|s) \int_{s'\in\mathcal{S}} p_{\mathcal{T}}(s'|s,a) \int_{c\in\mathcal{C}} p_{\mathcal{C}}(c|s,a)(b_{c,\gamma})_{\#} Z_\theta^c(s')\mathrm{d}a\mathrm{d}s'\mathrm{d}c$$

where $Z^c(s_t)_\theta = \sum_{\iota=0}^{T-t} \gamma^\iota C_\iota | S_0 = s_t$ denotes the variable of discounted cumulative costs parameterized by θ with N supporting quantiles.

We show that the distributional Bellman equation can capture the key components for representing the **aleatoric uncertainty** under the measure of entropy.

Leveraging the aforementioned policy optimization objective and distributional estimator, we design the **Distributional Lagrange Policy Optimization (DLPO)** algorithm to learn the policy under risk-sensitive constraints.

Inverse Step: Constraint Inference with Flow-based Data Augmentation

We utilize the mutual information $I(\omega; y|\boldsymbol{x}, \mathcal{D})$ as a measure of epistemic uncertainty [5]:

- Information gained by model ω when observing true label y for a given input x
- The greater the uncertainty of the model regarding the data, the more additional information it can obtain.

Intuitively, epistemic uncertainty arises when the constraint model ω is required to predict the cost of an OoD trajectory $\bar{\tau}$.
To mitigate the impact of epistemic uncertainty, we need to minimize the mutual information $I(\omega; \Phi|\bar{\tau}, \mathcal{D})$, which can be empirically represented by:

$$\mathcal{H}[p(\Phi|\bar{\tau}, \mathcal{D})] - \frac{1}{M}\sum_{m}\mathcal{H}[p(\Phi|\bar{\tau}; \omega_m)] \textit{where } \omega_m \sim q(\omega)$$

where 1) Φ is a Bernoulli feasibility variable that takes two values $\{\phi, \phi^-\}$ such that $p(\phi|s, \alpha; \omega)$ quantifies to what extent performing action a in the state s is feasible, and 2) $q(\omega)$ denote the dropout distribution.

Inverse Step: Constraint Inference with Flow-based Data Augmentation

To reduce the impact of **epistemic uncertainty** as well as the mutual information $I(\omega; \Phi|\bar{\tau}, \mathcal{D})$, we consider expanding the dataset by generating trajectories, which leads to the data-augmented constraint inference objective:

$$\frac{1}{M}\sum_{m}\mathbb{E}_{\mathcal{D}_e^G}\Big[\sum_{t=0}^{T}\log[p(\phi|s_t^e, a_t^e; \omega_m)]\Big] - \mathbb{E}_{\hat{\mathcal{D}}^G}\Big[\sum_{t=0}^{T}\log[p(\phi|\hat{s}_t, \hat{a}_t; \omega_m)]\Big] + \alpha\mathcal{H}[p(\Phi|\bar{\tau}; \omega_m)]$$

where $\mathcal{D}_e^G$ and $\hat{\mathcal{D}}^G$ are augmented expert and nominal dataset.

We propose a **Flow-based Trajectory Generation (FTG)** algorithm to perform conditional generation by training a Continuous Flow Network (CFlowNet) [2] and then utilize it to generate trajectories based on flows:

- Learning Flow Functions: Train the flow function $F_\xi(\cdot)$, which quantifies the mass of particles passing by, and denser particles indicate a higher probability
- Trajectory Generation: Generate trajectories $\tau^G = (s_0, a_0, \ldots, s_T, a_T)$ by sampling actions based on the scale of $F_\xi(s_t, a_t)$

Experiment Settings

The experiments are based on an ICRL benchmark [3] and extend it to include stochastic dynamics by incorporating noises into transitions.

The evaluation metrics include:

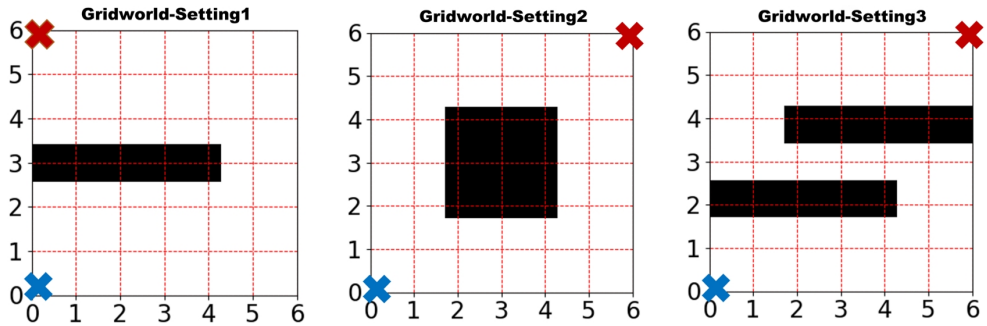
- *Constraint Violation Rate* measures the probability that a policy violates constraint in a trajectory
- *Feasible Cumulative Rewards* calculate the total rewards obtained by the agent before violating any constraints

The comparison methods are shown as follows:

Method	Continues Space	Constraint Optimization	Maximum Entropy	Aleatoric Uncertainty	Epistemic Uncertainty
GACL	✓	×	×	×	×
B2CL	✓	✓	×	×	×
ICRL	✓	✓	✓	×	×
VICRL	✓	✓	✓	×	✓
UAICRL-NRS	✓	✓	✓	×	✓
UAICRL-NDA	✓	✓	✓	✓	×
UAICRL	✓	✓	✓	✓	✓

Discrete Environment: Gridworld

We construct three Gridworld environments with different constraints.



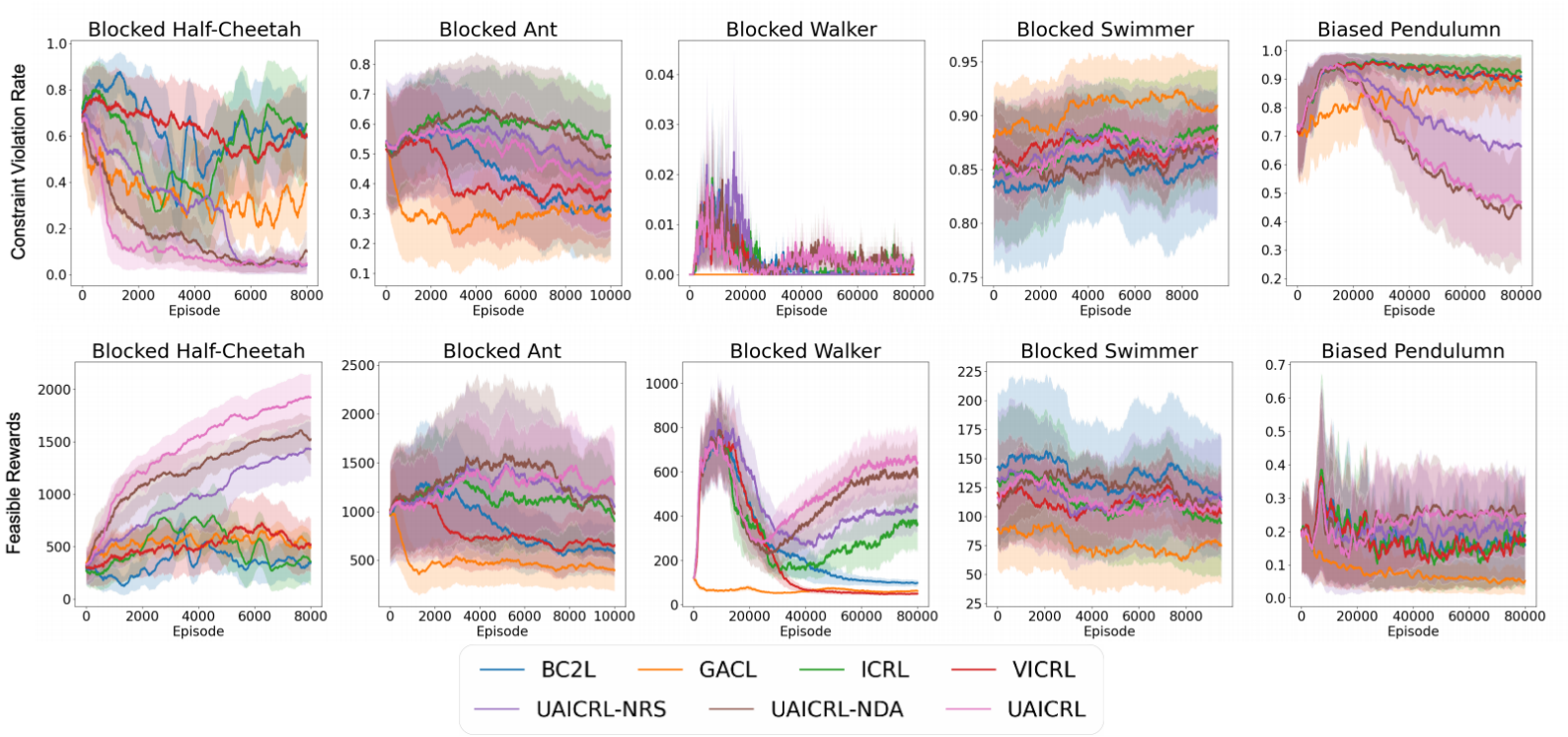
The evaluation results of different methods in three Gridworlds with the random rate p_s=0.01 and 0.001 are as follows:

		$p_s = 0.01$			$p_s = 0.001$		
		Gridworld Setting 1	Gridworld Setting 2	Gridworld Setting 3	Gridworld Setting 1	Gridworld Setting 2	Gridworld Setting 3
Feasible Rewards	BC2L	0.451	**0.716**	0.125	0.647	0.602	0.192
	GACL	0.032	0.109	0.000	0.011	0.070	0.000
	ICRL	0.244	0.532	0.033	0.356	0.368	0.089
	VICRL	0.537	0.310	0.051	0.778	0.610	0.070
	UAICRL	**0.650**	0.683	**0.359**	**0.797**	**0.739**	**0.401**
Constraint Violation Rate	BC2L	0.33	0.19	0.58	0.29	0.27	0.52
	GACL	0.43	0.29	0.78	0.67	0.11	0.84
	ICRL	0.53	0.33	0.63	0.36	0.27	0.73
	VICRL	0.35	0.33	0.45	0.19	0.28	0.53
	UAICRL	**0.13**	**0.09**	**0.34**	**0.09**	**0.07**	**0.38**

Continuous Environment: MuJoCo

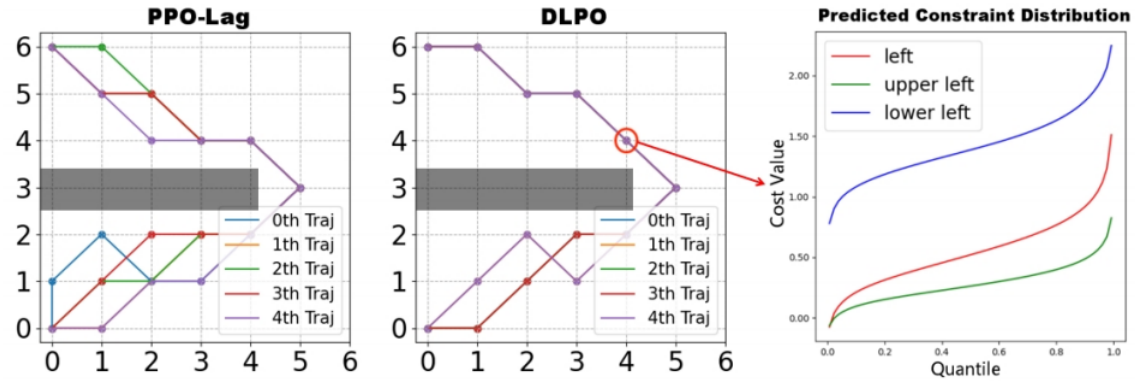
We utilize five MuJoCo environments and additionally incorporate Gaussian noise into the transition function as $p_{\mathcal{T}}(s_{t+1}|s_t, a_t) = f(s_t, a_t) + \mathcal{N}(\mu, \sigma)$

The constraint violation rate (top) and feasible rewards (bottom) in five MuJoCo environments during training with stochasticity of $\mathcal{N}(0, 0.1)$ are as follows:

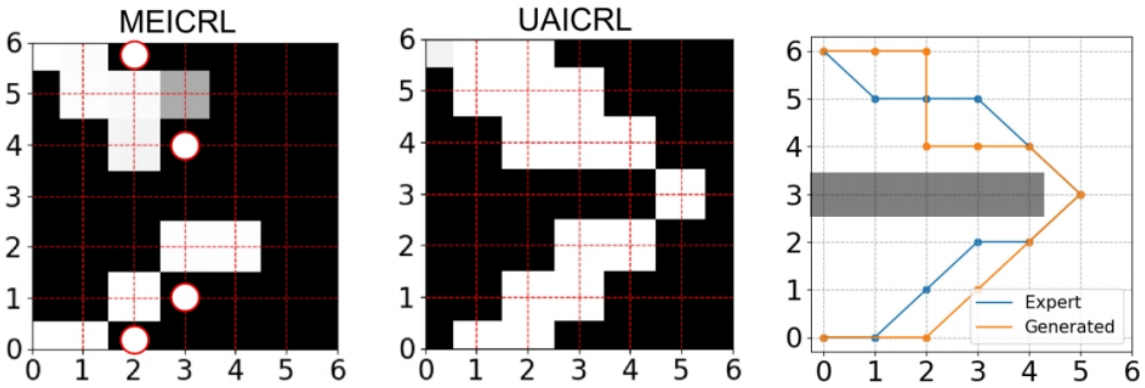


Visualization Results

The trajectories generated by PPO-Lag and DLPO, with the predicted cost distributions at the red circle:



The constraint map recovered by MEICRL and UAICRL, along with the trajectories generated by our FTG:



[1] Gerstenberg, J., Neininger, R., and Spiegel, D.
On solutions of the distributional bellman equation.
Electronic Research Archive 31, 8 (2023), 4459–4483.

[2] Li, Y., Luo, S., Wang, H., and Hao, J.
Cflownets: Continuous control with generative flow networks.
In *International Conference on Learning Representations (ICLR)* (2023).

[3] Liu, G., Luo, Y., Gaurav, A., Rezaee, K., and Poupart, P.
Benchmarking constraint inference in inverse reinforcement learning.
In *International Conference on Learning Representations (ICLR)* (2023).

[4] Malik, S., Anwar, U., Aghasi, A., and Ahmed, A.
Inverse constrained reinforcement learning.
In *International Conference on Machine Learning (ICML)* (2021), pp. 7390–7399.

[5] Smith, L., and Gal, Y.
Understanding measures of uncertainty for adversarial example detection.
In *The Conference on Uncertainty in Artificial Intelligence (UAI)* (2018), pp. 560–569.

[6] Ziebart, B. D., Bagnell, J. A., and Dey, A. K.
Modeling interaction via the principle of maximum causal entropy.
In *International Conference on Machine Learning (ICML)* (2010), pp. 1255–1262.

Thanks for your listening!