南京航空航天大學

Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室

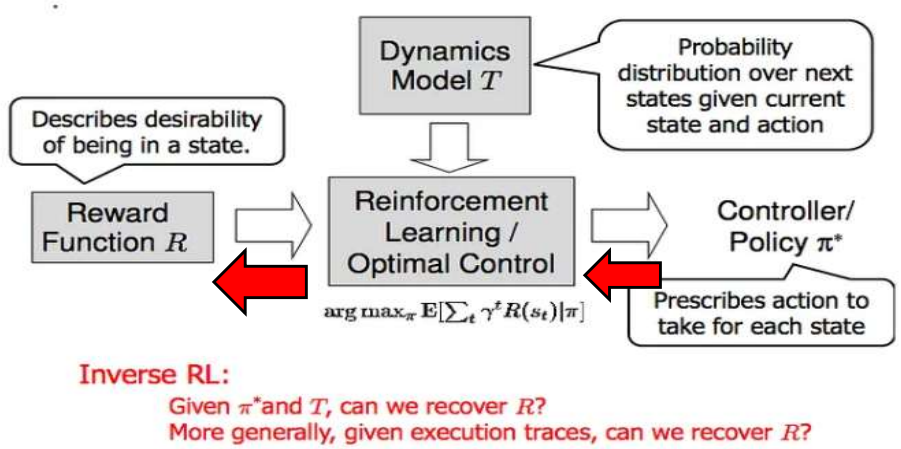
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

Causal Imitation Learning Via Inverse Reinforcement Learning

ICLR | 2023

Inverse Reinforcement learning

- learn a potential reward function under which the experts behavior policy is optimal.



$$maximize_{c \in C}(min_{\pi \in \Pi} - H(\pi) + E_\pi[c(s,a)]) - E_{\pi_E}[c(s,a)]$$

Auto-drive Experiment

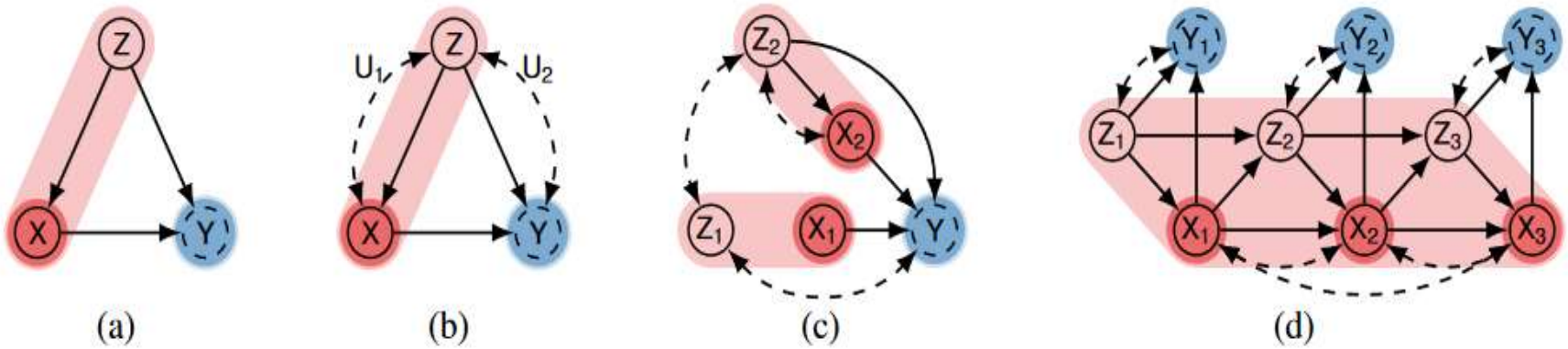
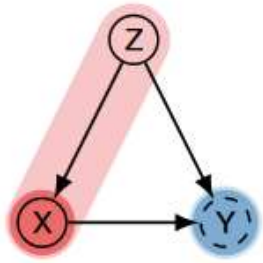


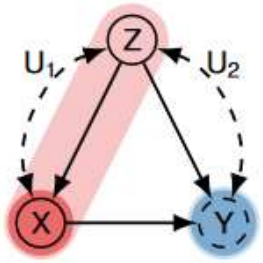
Figure 1: Causal diagrams where X represents an action (shaded red) and Y represents a latent reward (shaded blue). Input covariates of the policy scope $\mathcal{S}$ are shaded in light red.

- X: Action(acceleration), [0, 1]
- Z: Location of surrounding cars
- Y: $\alpha x + \beta z - \gamma xz$, $0 < \alpha < \gamma$, Reward function
- U1: car horn of surrounding vehicles
- U2: wind condition

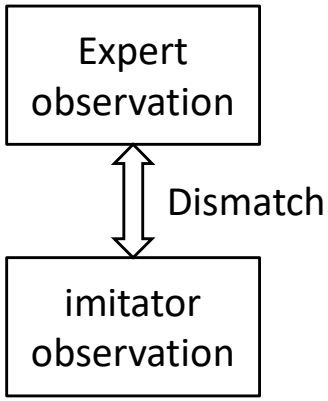
Auto-drive Experiment



(a)



(b)



$Y \leftarrow (1 - X)Z + X(1-Z)$

Behavior policy:

P (X = 1 | Z = 0) = 0.6

P (X = 0 | Z = 1) = 0.4

Z are drawn uniformly over $\{0, 1\}$

E[Y] = 0.5

E[Y |do(π∗)] = 1

outperform

$Z \leftarrow U_1 \oplus U_2$

$Y \leftarrow \neg X \oplus Z \oplus U_2$

U_1 and U_2 are drawn uniformly over $\{0, 1\}$

Expert can hear horn

E[Y] = 1

E[Y |do(π∗)] = 0.5

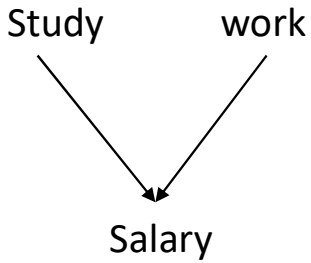
worse

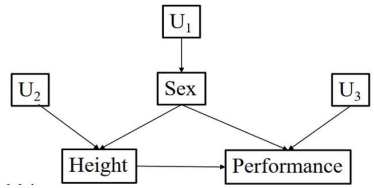
What can we do to prevent?

Preliminary

Definition: SCM is a tuple (X, U, F, P(u)), represent the causal relationships between variables in a system, where the nodes represent the variables in the system and the edges represent the causal relationships between them。

- X : endogenous variables
- U : exogenous variables,
- F : functions determining X,
  for each Xi,
  Xi :=$f_i(X_{pa(i,c)}, U_i)$
- P(u) is a distribution over U





Preliminary

- Policy Scope

  a policy scope S (for short, scope) over actions X is a sequence of tuples $\{\langle Xi, Zi\rangle\}_{i=1}^{n}$ where $Zi \subseteq Pa_i^*$ for every $Xi \in X$, A policy $\pi \sim S$ is a sequence of distributions $\pi = \{\pi1(X1 \mid Z1), \pi2(X2 \mid Z2)\}$

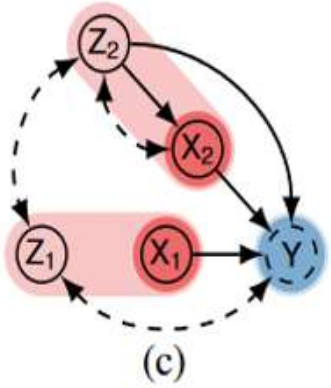
  $S = \{\langle X1, \{Z1\}\rangle, \langle X2, \{Z2\}\rangle\}$

- π-backdoor criterion

  Xi is not an ancestor of Y; $X \notin An(Y)_{\mathcal{G}^{(i)}}$

  or

  Zi blocks all backdoor path from Xi to Y $(Y \perp\!\!\!\perp X_i | \boldsymbol{Z}_i)$ in $\mathcal{G}^{(i)}_{\underline{X_i}}$.



- minimal π-backdoor admissible scope

  there exists no proper subscope S′ ⊂ S satisfying the sequential π-backdoor in G

Preliminary

- MWAL: 基于projection算法的优化方法
  - Projection
    1. 根据给定的奖励函数（由当前w计算出）找出一个最优策略。
    2. 计算给定策略（1中最优策略）的特征期望

⇩ 可以保证学徒的和专家的价值差距有一个上限 可能一样好，也可能一样坏

$$
\begin{aligned}
|V(\overline{\psi}) - V(\pi_E)| &= |\mathbf{w}^* \cdot \boldsymbol{\mu}(\overline{\psi}) - \mathbf{w}^* \cdot \boldsymbol{\mu}_E| \\
&\leq \|\mathbf{w}^*\|_2 \|\boldsymbol{\mu}(\overline{\psi}) - \boldsymbol{\mu}_E\|_2 \\
&\leq \epsilon
\end{aligned}
$$

⇩

$$
v^* = \max_{\psi \in \Psi} \min_{\mathbf{w} \in \mathbb{S}^k} \left[\mathbf{w} \cdot \boldsymbol{\mu}(\psi) - \mathbf{w} \cdot \boldsymbol{\mu}_E\right]. \tag{3}
$$

$$
\mathbf{G}(i,j) = \boldsymbol{\mu}^j(i) - \boldsymbol{\mu}_E(i)
$$

$$
v^* = \max_{\psi \in \Psi} \min_{\mathbf{w} \in \mathbb{S}^k} \mathbf{w}^T \mathbf{G} \psi = \min_{\mathbf{w} \in \mathbb{S}^k} \max_{\psi \in \Psi} \mathbf{w}^T \mathbf{G} \psi.
$$

Preliminary

- GAIL: 基于GAN生成专家数据直接学习专家策略

1.通过当前policy采样得到的数据与专家数据进行对抗训练来训练Discriminator；
2.利用Discriminator作为surrogate reward function来训练策略Policy（TPRO）

定义 ψ：

$$\psi_{GA}(c) = \begin{cases} \mathbb{E}_{\pi_E}[g(c(s,a))] & \text{if } c < 0 \\ +\infty & \text{otherwise} \end{cases} \quad \text{where}$$

$$g(x) = \begin{cases} -x - \log(1 - e^x) & x < 0 \\ +\infty & \text{otherwise} \end{cases}$$

通过推导，将IL求解cost function的公式变为：

$$\underset{\pi}{\text{minimize}}\ \psi^*_{\text{GA}}(\rho_\pi - \rho_{\pi_E}) - \lambda H(\pi) = D_{\text{JS}}(\rho_\pi, \rho_{\pi_E}) - \lambda H(\pi),$$

$$\Downarrow$$

$$\underset{c \in \mathcal{C}}{\text{maximize}} \left(\min_{\pi \in \Pi} -H(\pi) + \mathbb{E}_\pi[c(s,a)] \right) - \mathbb{E}_{\pi_E}[c(s,a)]$$

Minimal Sequential Backdoor Criterion

Definition 4. Given a causal diagram $\mathcal{G}$, a π-backdoor admissible scope $\mathcal{S}$ is said to be *minimal* if there exists no proper subscope $\mathcal{S}' \subset \mathcal{S}$ satisfying the sequential π-backdoor in $\mathcal{G}$.

Theorem 1. *Given a causal diagram $\mathcal{G}$, if there exists a minimal π-backdoor admissible scope $\mathcal{S} = \{\langle X_i, \boldsymbol{Z}_i \rangle\}_{i=1}^{n}$ in $\mathcal{G}$, consider the following conditions:*

1. *Let effective actions $\boldsymbol{X}^* = \boldsymbol{X} \cap An(Y)_{\mathcal{G}_{\mathcal{S}}}$ and effective covariates $\boldsymbol{Z}^* = \bigcup_{X_i \in \boldsymbol{X}^*} \boldsymbol{Z}_i$;*
2. *For $i = 1, \ldots, n+1$, let $\boldsymbol{X}^*_{<i} = \{\forall X_j \in \boldsymbol{X}^* \mid j < i\}$ and $\boldsymbol{Z}^*_{<i} = \bigcup_{X_j \in \boldsymbol{X}^*_{<i}} \boldsymbol{Z}_j$.*

Then, for any policy $\boldsymbol{\pi} \sim \mathcal{S}$, the expected reward $\mathbb{E}[Y \mid do(\boldsymbol{\pi})]$ is computable from $P(\boldsymbol{O}, Y)$ as:

$$\mathbb{E}[Y \mid do(\boldsymbol{\pi})] = \sum_{\boldsymbol{x}^*, \boldsymbol{z}^*} \mathbb{E}[Y \mid \boldsymbol{x}^*, \boldsymbol{z}^*] \rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*) \tag{2}$$

where the occupancy measure $\rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*) = \prod_{X_i \in \boldsymbol{X}^*} P\left(\boldsymbol{z}_i \mid \boldsymbol{x}^*_{<i}, \boldsymbol{z}^*_{<i}\right) \pi_i(x_i \mid \boldsymbol{z}_i)$.

$$\rho_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R} \text{ as } \rho_\pi(s, a) = \pi(a|s) \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\pi)$$

Imitation Via Inverse Reinforcement Learning

$$\nu^* = \min_{\pi \sim \mathcal{S}} \max_{M \in \mathcal{M}} \mathbb{E}_M[Y] - \mathbb{E}_M[Y \mid \mathrm{do}(\pi)]. \tag{1}$$

$$\mathbb{E}[Y \mid do(\pi)] = \sum_{x^*, z^*} \mathbb{E}[Y \mid x^*, z^*] \rho_\pi(x^*, z^*) \tag{2}$$

- Object Function:

$$\nu^* = \min_{\pi \sim \mathcal{S}} \max_{r \in \mathcal{R}} \sum_{x^*, z^*} r(x^*, z^*) \left(\rho(x^*, z^*) - \rho_\pi(x^*, z^*)\right) \tag{3}$$

Imitation Via Inverse Reinforcement Learning

- Causal MWAL

  Proposition 1. For a hypothesis class R = {r = w · ϕ | w ∈ Sk}, the solution v∗ of the canonical equation in Eq. (3) is obtainable by solving the following minimax problem:

$$\nu^* = \min_{\pi \sim \mathcal{S}} \max_{w \in \mathbb{S}^k} w^\top G \pi, \tag{4}$$

  where G is a k × n matrix given by $G(i, j) = \sum_{x^*, z^*} \phi^{(i)}(x^*, z^*)\,(\rho(x^*, z^*) - \rho_{\pi^{(j)}}(x^*, z^*))$.

- Causal GAIL

  Proposition 2. *For a hypothesis class* $\mathscr{R} = \{r : \mathscr{D}_{X^*} \times \mathscr{D}_{Z^*} \mapsto \mathbb{R}\}$ *regularized by* ψ, *the solution* ν^* *of the penalized canonical equation in Eq. (5) is obtainable by solving the following problem:*

$$\nu^* = \min_{\pi \sim \mathcal{S}} \psi^* (\rho - \rho_\pi) \tag{6}$$

  where ψ^* *be a conjugate function of* ψ *and is given by* $\psi^* = \max_{r \in \mathbb{R}^{X \times Z}} a^\top r - \psi(r)$.

Imitation Via Inverse Reinforcement Learning

Causal MWAL

Algorithm 3: Causal MWAL

Input: $\mathcal{G}$, Expert demonstrations τ_E

1: Apply Thm. 1 (or Thm. 2) to obtain formulas for the expert's $\rho(\boldsymbol{x}^*, \boldsymbol{z}^*)$ and imitator's occupancy measure $\rho_{\pi_j}(\boldsymbol{x}^*, \boldsymbol{z}^*)$

2: Let $\boldsymbol{\mu}(i) = \sum_{\boldsymbol{x}^*, \boldsymbol{z}^*} \phi^{(i)}(\boldsymbol{x}^*, \boldsymbol{z}^*) \rho(\boldsymbol{x}^*, \boldsymbol{z}^*)$ and let $\boldsymbol{\mu}_{\boldsymbol{\pi}}(i) = \sum_{\boldsymbol{x}^*, \boldsymbol{z}^*} \phi^{(i)}(\boldsymbol{x}^*, \boldsymbol{z}^*) \rho_{\boldsymbol{\pi}}(\boldsymbol{x}^*, \boldsymbol{z}^*)$

3: Let $\tilde{\boldsymbol{G}}(i, \boldsymbol{\mu}_{\boldsymbol{\pi}}) = \left((\hat{\boldsymbol{\mu}}(i) - \boldsymbol{\mu}_{\boldsymbol{\pi}}(i)) - 2\right)/4$

4: Let $\beta = \left(1 + \sqrt{\frac{2\ln(k)}{J}}\right)^{-1}$

5: Initialize $\boldsymbol{w}^1(i) = 1$ for $i = 1, ..., k$

6: **for** iteration $j = 0, 1, 2, \ldots, J$ **do**

7: Set $\boldsymbol{w}^j(i) = \frac{\boldsymbol{w}^j(i)}{\sum_i \boldsymbol{w}^j(i)}$ for $i = 1, ..., k$

8: Compute the policy $\hat{\boldsymbol{\pi}}_j$ by $\arg\min_{\boldsymbol{\pi}} \boldsymbol{w}^\top \tilde{\boldsymbol{G}} \boldsymbol{\pi}$, where $\boldsymbol{w} := \boldsymbol{w}^j$

9: Compute $\hat{\boldsymbol{\mu}}_j = \boldsymbol{\mu}_{\hat{\boldsymbol{\pi}}_j}$

10: $\boldsymbol{w}^{j+1}(i) = \boldsymbol{w}^j(i) \cdot \exp(\ln(\beta) \cdot \tilde{\boldsymbol{G}}(i, \hat{\boldsymbol{\mu}}_j))$ for $i = 1, ..., k$

11: **end for**

12: Output: The mixed policy that has a probability $\frac{1}{J}$ of choosing $\hat{\boldsymbol{\pi}}_j$, for all $t \in \{1, \ldots, J\}$

Algorithm 1 The MWAL algorithm

1: **Given:** An MDP\R M and an estimate of the expert's feature expectations $\hat{\mu}_E$.

2: Let $\beta = \left(1 + \sqrt{\frac{2\ln k}{T}}\right)^{-1}$.

3: Define $\tilde{\mathbf{G}}(i, \boldsymbol{\mu}) \triangleq \left((1-\gamma)(\boldsymbol{\mu}(i) - \hat{\mu}_E(i)) + 2\right)/4$, where $\boldsymbol{\mu} \in \mathbb{R}^k$.

4: Initialize $\mathbf{W}^{(1)}(i) = 1$ for $i = 1, \ldots, k$.

5: **for** $t = 1, \ldots, T$ **do**

6: Set $\mathbf{w}^{(t)}(i) = \frac{\mathbf{W}^{(t)}(i)}{\sum_i \mathbf{W}^{(t)}(i)}$ for $i = 1, \ldots, k$.

7: Compute an ϵ_P-optimal policy $\hat{\pi}^{(t)}$ for M with respect to reward function $R(s) = \mathbf{w}^{(t)} \cdot \boldsymbol{\phi}(s)$.

8: Compute an ϵ_F-good estimate $\hat{\boldsymbol{\mu}}^{(t)}$ of $\boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}(\hat{\pi}^{(t)})$.

9: $\mathbf{W}^{(t+1)}(i) = \mathbf{W}^{(t)}(i) \cdot \exp(\ln(\beta) \cdot \tilde{\mathbf{G}}(i, \hat{\boldsymbol{\mu}}^{(t)}))$ for $i = 1, \ldots, k$.

10: **end for**

11: Post-processing: Return the mixed policy $\overline{\psi}$ that assigns probability $\frac{1}{T}$ to $\hat{\pi}^{(t)}$, for all $t \in \{1, \ldots, T\}$.

Imitation Via Inverse Reinforcement Learning

Causal GAIL

Algorithm 4: Causal GAIL

Input: $\mathcal{G}$, Expert demonstrations τ_E

1: Apply Thm. 1 (or Thm. 2) to obtain formulas for the expert's $\rho(x^*, z^*)$ and imitator's occupancy measure $\rho_{\pi_j}(x^*, z^*)$

2: **for** iteration $j = 0, 1, 2, \ldots$ **do**

3: Collect trajectories from distributions $\rho(x^*, z^*)$ and $\rho_{\pi_j}(x^*, z^*)$

4: Update the parameters w of discriminator D_j with gradient

$$\hat{\mathbb{E}}[\nabla_w \log(D_j(x^*, z^*))] + \hat{\mathbb{E}}_{\pi_j}[\nabla_w \log(1 - D_j(x^*, z^*))]$$

5: Update the policy $\pi_j = \arg\min_\pi \mathbb{E}_\pi[\log(1 - D(x^*, z^*))]$ with policy optimization for DTR

6: **end for**

Algorithm 1 Generative adversarial imitation learning

1: **Input:** Expert trajectories $\tau_E \sim \pi_E$, initial policy and discriminator parameters θ_0, w_0

2: **for** $i = 0, 1, 2, \ldots$ **do**

3: Sample trajectories $\tau_i \sim \pi_{\theta_i}$

4: Update the discriminator parameters from w_i to w_{i+1} with the gradient

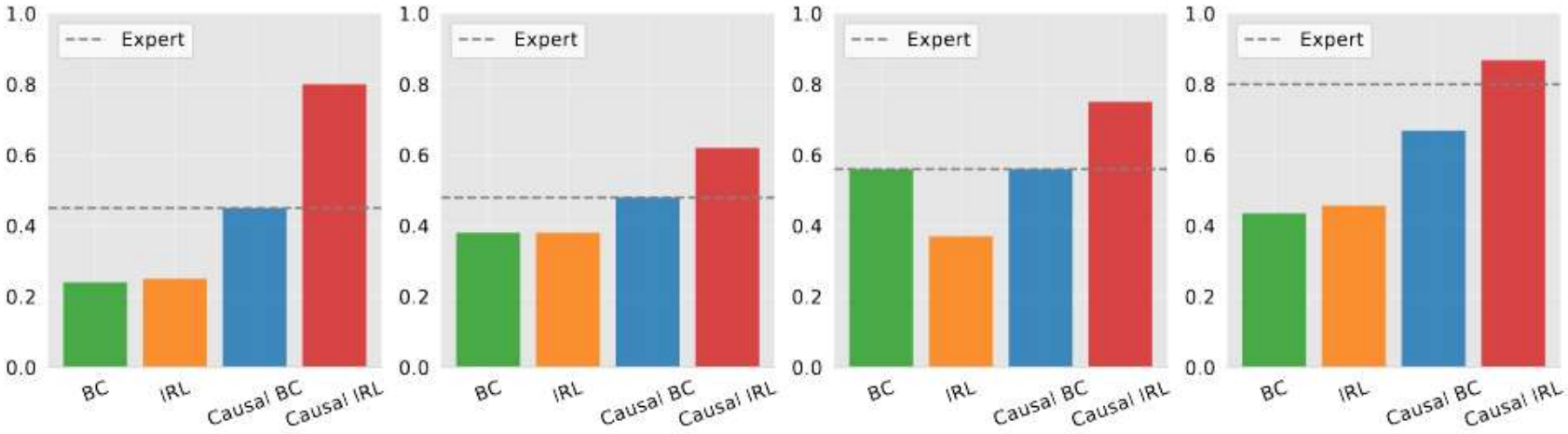
$$\hat{\mathbb{E}}_{\tau_i}[\nabla_w \log(D_w(s,a))] + \hat{\mathbb{E}}_{\tau_E}[\nabla_w \log(1 - D_w(s,a))] \tag{17}$$

5: Take a policy step from θ_i to θ_{i+1}, using the TRPO rule with cost function $\log(D_{w_{i+1}}(s,a))$. Specifically, take a KL-constrained natural gradient step with

$$\hat{\mathbb{E}}_{\tau_i}[\nabla_\theta \log \pi_\theta(a|s) Q(s,a)] - \lambda \nabla_\theta H(\pi_\theta), \quad \text{where } Q(\bar{s},\bar{a}) = \hat{\mathbb{E}}_{\tau_i}[\log(D_{w_{i+1}}(s,a)) \mid s_0 = \bar{s}, a_0 = \bar{a}] \tag{18}$$

6: **end for**

Experiment



(a) Backdoor (b) Highway Driving (c) MNIST Digits (d) Infinite MDPUC

Figure 3: Simulation results (a, b, c, d) for our experiments, where y-axis represents the expected reward of learned policies in the actual causal model; the grey dashed line denotes the expert's reward.

Experiment

Table 1: Expected rewards $\mathbb{E}[Y \mid \text{do}(\pi)]$ for all imitation strategies. $\mathbb{E}[Y]$ measures the expert's performance. For each experiment, we highlight policies with the optimal performance.

	Experiment	BC	IRL	$\pi_{c\text{-}bc}$	$\pi_{c\text{-}irl}$	$\mathbb{E}[Y]$
1	Backdoor	0.24 ± 0.0042	0.25 ± 0.017	0.45 ± 0.0057	$\mathbf{0.80} \pm 0.0048$	0.45
2	HighD + RA	0.38 ± 0.0073	0.38 ± 0.0815	0.49 ± 0.0066	$\mathbf{0.62} \pm 0.0026$	0.48
3	MNIST	0.56 ± 0.0046	0.37 ± 0.0035	0.56 ± 0.0048	$\mathbf{0.75} \pm 0.0027$	0.56
4	MDPUC	0.435 ± 7.160	0.457 ± 3.718	0.669 ± 1.789	$\mathbf{0.868} \pm 0.327$	0.8
5	Frontdoor	0.52 ± 0.0046	0.51 ± 0.077	0.52 ± 0.0054	$\mathbf{0.63} \pm 0.0045$	0.60
6	Backdoor (Linear)	0.70 ± 0.0044	0.72 ± 0.013	0.75 ± 0.0036	$\mathbf{0.98} \pm 0.0003$	0.75
7	Frontdoor (Linear)	0.62 ± 0.0037	0.50 ± 0.0040	0.62 ± 0.0036	$\mathbf{0.75} \pm 0.0030$	0.62
8	HighD	0.38 ± 0.0073	0.38 ± 0.082	0.49 ± 0.0066	$\mathbf{0.49} \pm 0.051$	0.48