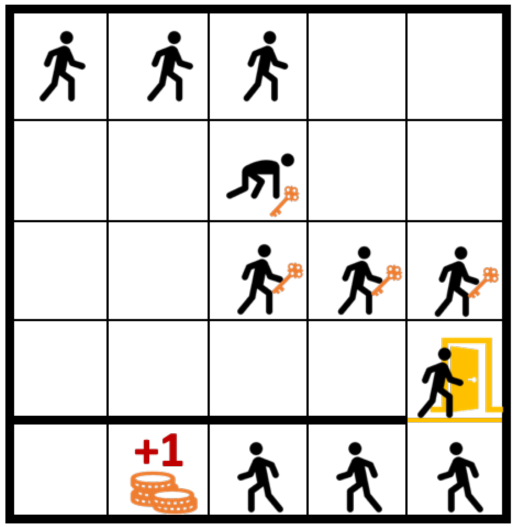
Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach

Yudi Zhang, Yali Du, Biwei Huang, Ziyan Wang, Jun Wang,

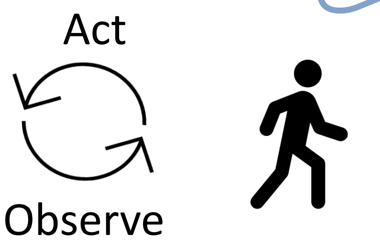
Meng Fang, Mykola Pechenizkiy

[1] Yudi Zhang, Yali Du, Biwei Huang, Ziyan Wang, Jun Wang, Meng Fang, and Mykola Pechenizkiy. "Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach." NeurIPS 2023.

A major challenge in RL: Delayed Reward



Delayed Reward:
Obtaining : positive reward **+1**
Otherwise: reward **0**



Lack of immediate feedback

Unstable policy optimization

Reward Redistribution:

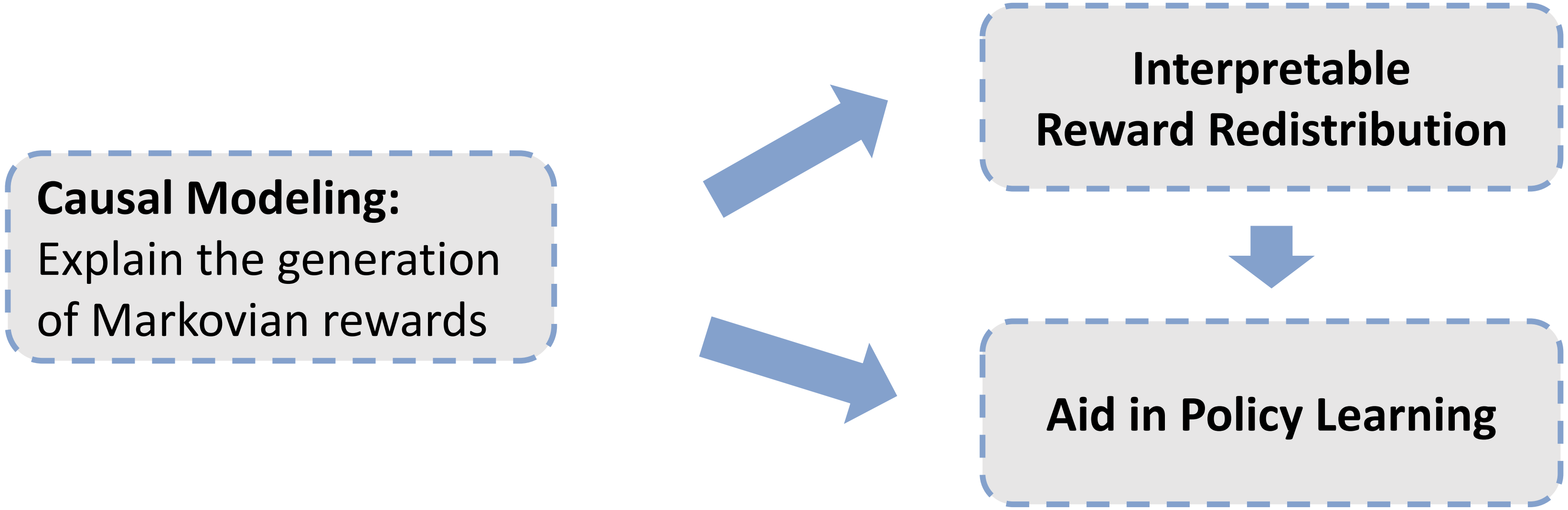
Assign proxy rewards according to the contribution of each state-action pair

[1] Yudi Zhang, Yali Du, Biwei Huang, Ziyan Wang, Jun Wang, Meng Fang, and Mykola Pechenizkiy. "Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach." NeurIPS 2023.

Motivation

Equally important for Interpretable Reward Redistribution:

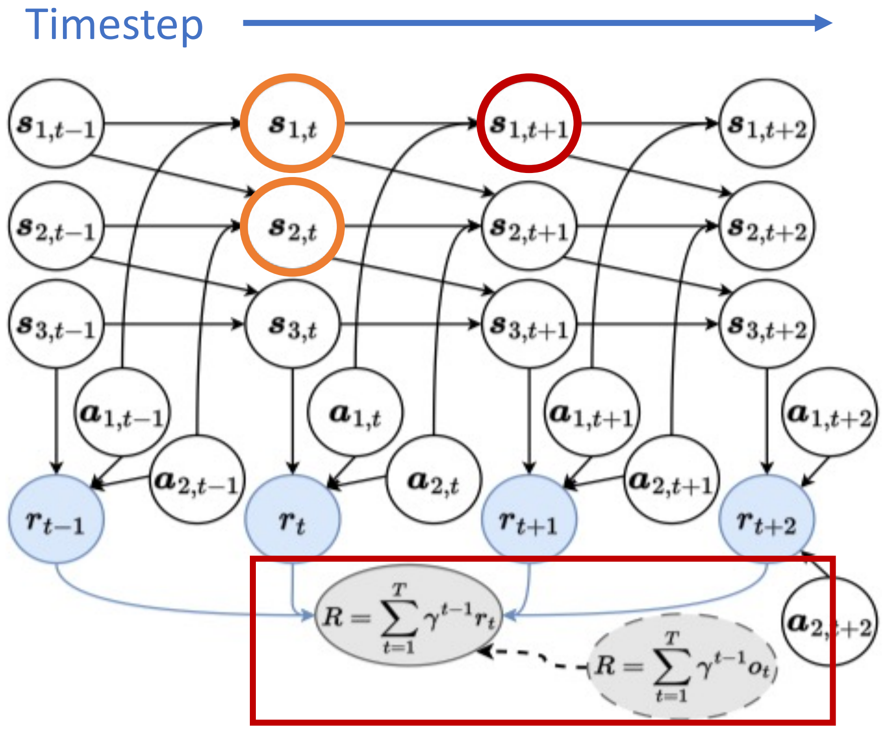
- Computing the contribution of each state-action pair towards delayed rewards?
- Explaining the reasons behind such contribution?



Causal Reformulation of Reward Redistribution

Causality:

causes (which part of the state-action pair) → outcomes (Markovian reward)



- **<u>A graphical example</u>** to model the generation of ,

  Markovian rewards r_t,

  long-term returns R,

  by the causal structure over $\boldsymbol{s}_t$, $\boldsymbol{a}_t$, r_t and R.

Observable

Unobservable &
Goal of Reward Redistribution

[1] Yudi Zhang, Yali Du, Biwei Huang, Ziyan Wang, Jun Wang, Meng Fang, and Mykola Pechenizkiy. "Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach." NeurIPS 2023.

Causal Reformulation of Reward Redistribution

A generative process in MDP:

$$\begin{cases} s_{i,t+1} = f(\boldsymbol{C}^{\boldsymbol{s}\to\boldsymbol{s}}_{\cdot,i} \odot \boldsymbol{s}_t, \boldsymbol{C}^{\boldsymbol{a}\to\boldsymbol{s}}_{\cdot,i} \odot \boldsymbol{a}_t, \epsilon_{s,i,t}) \\ r_t = g(\boldsymbol{C}^{\boldsymbol{s}\to r} \odot \boldsymbol{s}_t, \boldsymbol{C}^{\boldsymbol{a}\to r} \odot \boldsymbol{a}_t, \epsilon_{r,t}) \\ R = \sum_{t=1}^{T} \gamma^{t-1} r_t \end{cases}$$

Dynamics function

Markovian reward function

Return Equivalence

Causal structure $\boldsymbol{C}^{\cdot\to\cdot}$:

$$\boldsymbol{C}^{\boldsymbol{s}\to r} \in \{0,1\}^{|s|}, \boldsymbol{C}^{a\to r} \in \{0,1\}^{|a|},$$
$$\boldsymbol{C}^{\boldsymbol{s}\to\boldsymbol{s}} \in \{0,1\}^{|s|\times|s|}, \boldsymbol{C}^{a\to s} \in \{0,1\}^{|a|\times|s|};$$

$\epsilon_{s,i,t}$ and $\epsilon_{r,t}$: i.i.d random noises , $\odot$: element-wise product

Identifiability Result: Given observed state $\boldsymbol{s}_t$, action $\boldsymbol{a}_t$, long-term return R , under the global Markov condition and faithfulness assumption, the causal structure $\boldsymbol{C}^{\cdot\to\cdot}$, unknown functions, f, g and the rewards r_t are identifiable.

[1] Yudi Zhang, Yali Du, Biwei Huang, Ziyan Wang, Jun Wang, Meng Fang, and Mykola Pechenizkiy. "Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach." NeurIPS 2023.

Generative Return Decomposition

> How do we estimate the generative model?

Overall objective to optimize parameterized generative model:

$$L_{\mathrm{m}} = L_{\mathrm{rew}} + L_{\mathrm{cau}} + L_{\mathrm{reg}}$$

Minimize MSE for reward function:

$$L_{\mathrm{rew}}(\phi_{\mathrm{rew}}, \phi_{\mathrm{cau}}^{s\to r}, \phi_{\mathrm{cau}}^{a\to r}) = \mathbb{E}_{\tau\sim\mathcal{D}} \left\| R - \sum_{t=1}^{T} \gamma^{t-1}\hat{r}_t \right\|^2 = \mathbb{E}_{\tau\sim\mathcal{D}} \left\| \sum_{t=1}^{T} \gamma^{t-1} o_t - \sum_{t=1}^{T} \gamma^{t-1}\hat{r}_t \right\|^2$$

Maximize the likelihood for dynamic function:

$$L_{\mathrm{dyn}}\left(\phi_{\mathrm{dyn}}, \phi_{\mathrm{cau}}^{s\to s}, \phi_{\mathrm{cau}}^{a\to s}\right) = \mathbb{E}_{s_t, a_t, s_{t+1}\sim\mathcal{D}}\left[-\sum_{i=1}^{|s|} \log P\left(s_{i,t+1} \middle| s_t, a_t, \boldsymbol{C}^{s\to s}, \boldsymbol{C}^{a\to s}\right)\right]$$

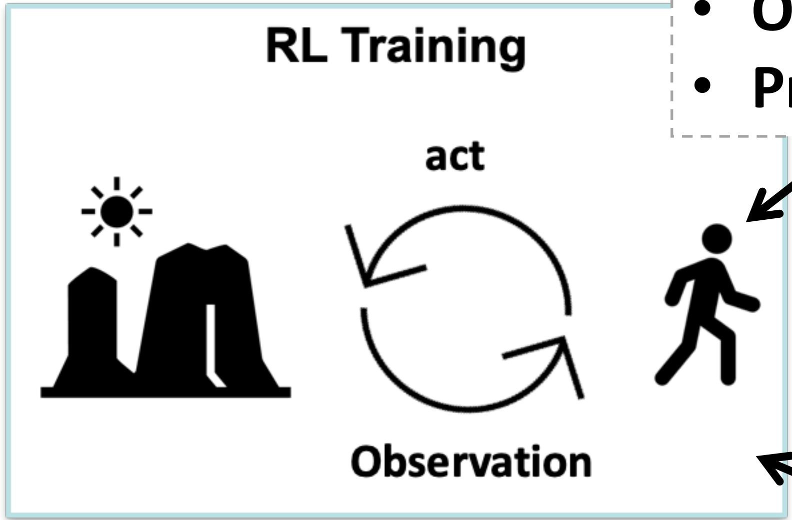
Regularizer:

$$L_{\mathrm{reg}}(\phi_{\mathrm{cau}}) = \lambda_1 \sum_i D_i(C^{s\to r}) + \lambda_2 \sum_i D_i(C^{a\to r}) + \lambda_3 \sum_{j\neq i} D_{i,j}(C^{s\to s})$$
$$+\lambda_4 \sum_{i=j} D_{i,j}(C^{s\to s}) + \lambda_5 \sum_{i=i,j} D_{i,j}(C^{a\to s}), \text{ where } D_i(C) = \log P(C_i = 1)$$

[1] Yudi Zhang, Yali Du, Biwei Huang, Ziyan Wang, Jun Wang, Meng Fang, and Mykola Pechenizkiy. "Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach." NeurIPS 2023.

Generative Return Decomposition

> How do we exploit the estimated generative model?



Reward Signal for Oprimizing Φ_π:
- Observed delayed rewards ✗
- Predicted Markovian Rewards ✓

Input of Policy Φ_π:
- State ✗
- Compact Representation (CR) ✓

Learned Generative Model

[1] Yudi Zhang, Yali Du, Biwei Huang, Ziyan Wang, Jun Wang, Meng Fang, and Mykola Pechenizkiy. “Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach.” NeurIPS 2023.

Compact Representation (CR)

> a minimally sufficient subset of all state components for policy learning

CR: All the state components influence ***rewards***.

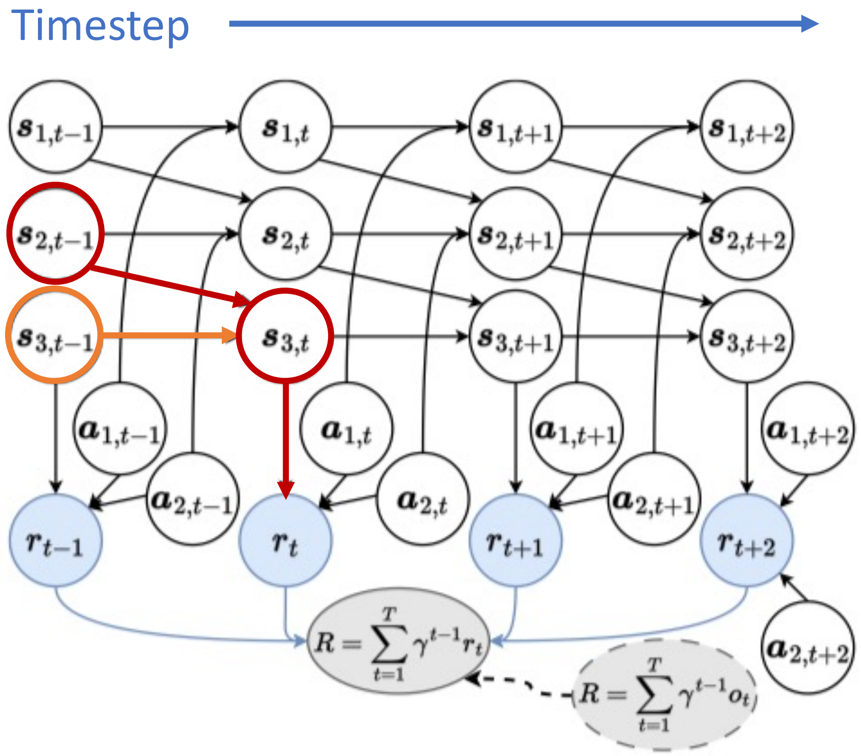
$\boldsymbol{s}_{i,t}$ is selected into **CR** if,

1) $C_i^{s\rightarrow r} = 1$:

$\boldsymbol{s}_{i,t}$ directly impacts r_t ($s_{2,t}$)

2) $C_{i,j}^{s\rightarrow s} = C_j^{s\rightarrow r} = 1$:

$\boldsymbol{s}_{i,t}$ indirectly impacts r_{t+1}, ($s_{3,t}$)



[1] Yudi Zhang, Yali Du, Biwei Huang, Ziyan Wang, Jun Wang, Meng Fang, and Mykola Pechenizkiy. "Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach." NeurIPS 2023.

Experimental Results – Episodic MuJoCo

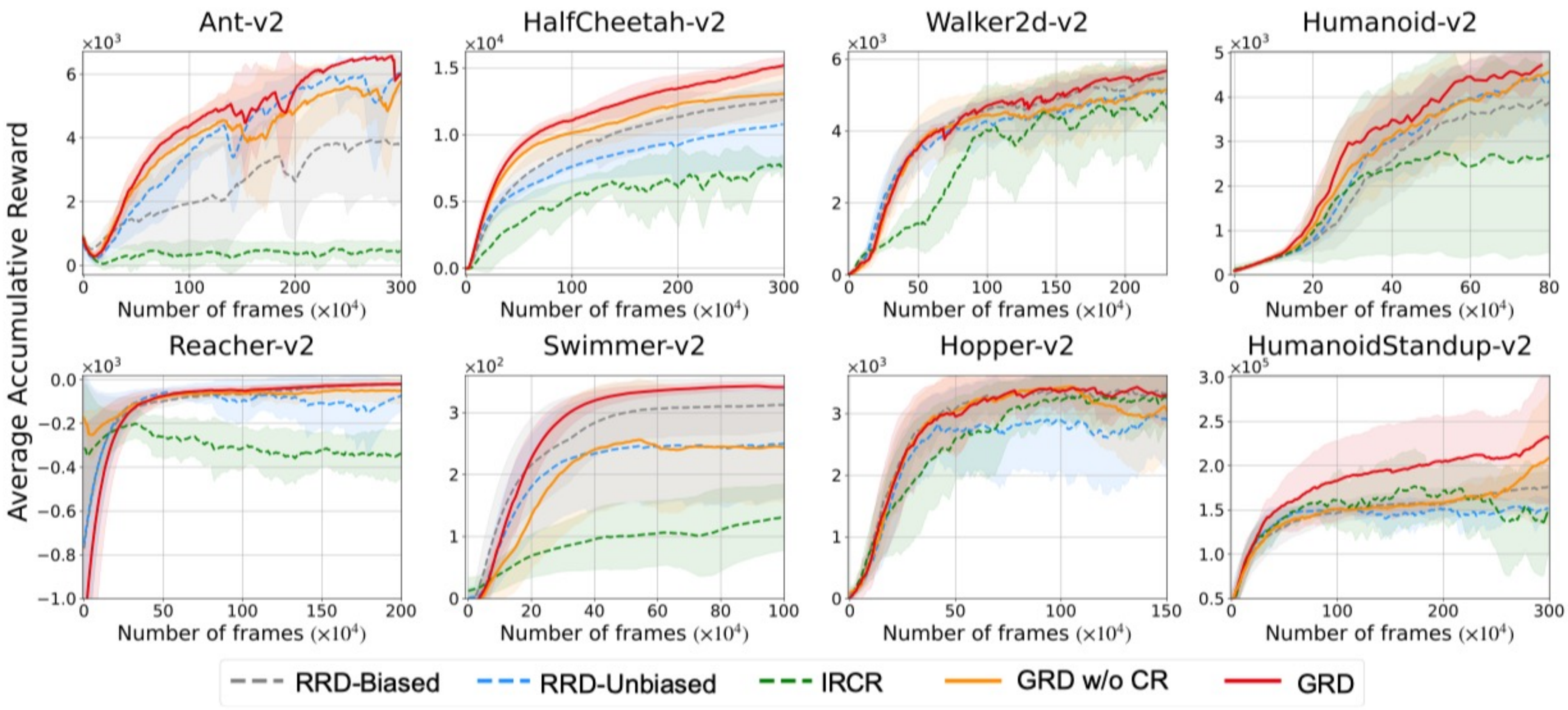
At time step t, the agent is marked as r_t.

The observed sparse and delayed rewards are,

$$o_t = \begin{cases} 0, & \text{if } t \neq T \\ \gamma^{\mathrm{t-1}} r_t, & \text{if } t = T \end{cases}$$

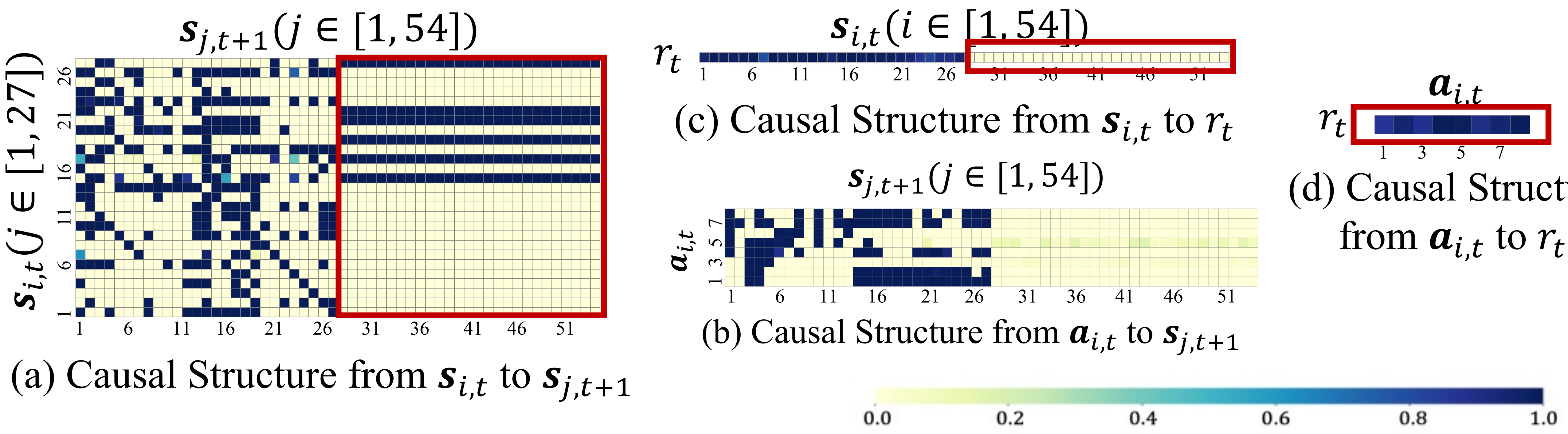
[1] Yudi Zhang, Yali Du, Biwei Huang, Ziyan Wang, Jun Wang, Meng Fang, and Mykola Pechenizkiy. "Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach." NeurIPS 2023.

Experimental Results – Episodic MuJoCo



[1] Yudi Zhang, Yali Du, Biwei Huang, Ziyan Wang, Jun Wang, Meng Fang, and Mykola Pechenizkiy. "Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach." NeurIPS 2023.

Visualization of Learned Causal Structure



(a) Causal Structure from $s_{i,t}$ to $s_{j,t+1}$

(c) Causal Structure from $s_{i,t}$ to r_t

(b) Causal Structure from $a_{i,t}$ to $s_{j,t+1}$

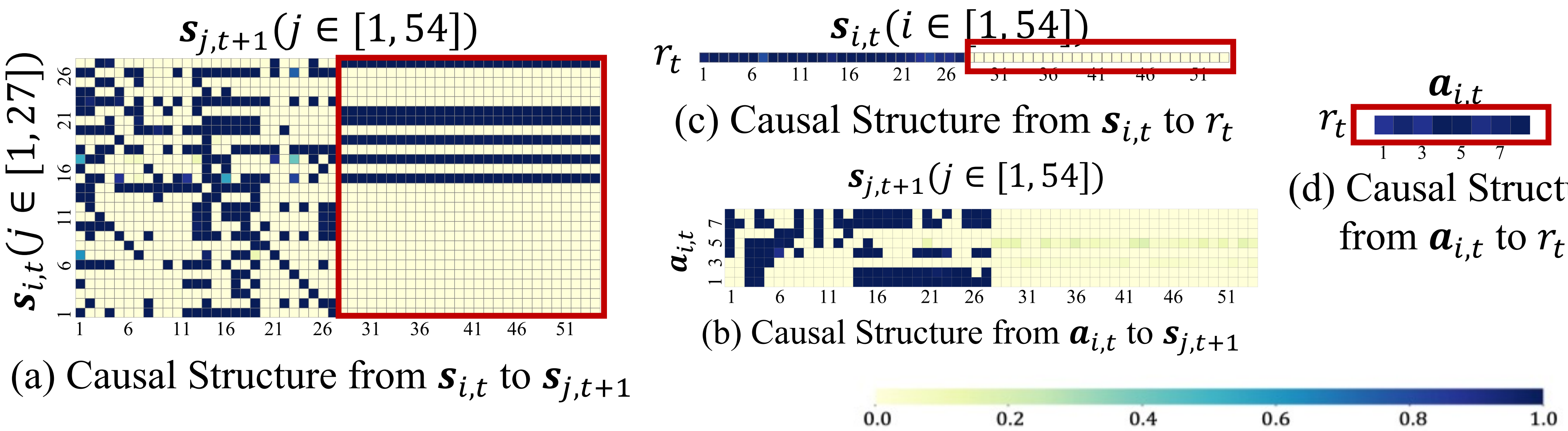
(d) Causal Structure from $a_{i,t}$ to r_t

Two characters of Ant:

- 28~111 dimensions in the state are not used.
- Low-cost control: all dimensions of action cause rewards.

[1] Yudi Zhang, Yali Du, Biwei Huang, Ziyan Wang, Jun Wang, Meng Fang, and Mykola Pechenizkiy. "Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach." NeurIPS 2023.

Visualization of Learned Causal Structure



(a) Causal Structure from $\boldsymbol{s}_{i,t}$ to $\boldsymbol{s}_{j,t+1}$

(c) Causal Structure from $\boldsymbol{s}_{i,t}$ to r_t

(b) Causal Structure from $\boldsymbol{a}_{i,t}$ to $\boldsymbol{s}_{j,t+1}$

(d) Causal Structure from $\boldsymbol{a}_{i,t}$ to r_t

- The edges related to no-used states do not exist (in yellow);
- Although learned redundant edges, they are not in CR;
- The edges from all dimensions of action to reward exist (in blue).

[1] Yudi Zhang, Yali Du, Biwei Huang, Ziyan Wang, Jun Wang, Meng Fang, and Mykola Pechenizkiy. "Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach." NeurIPS 2023.

Visualization of Redistributed Rewards

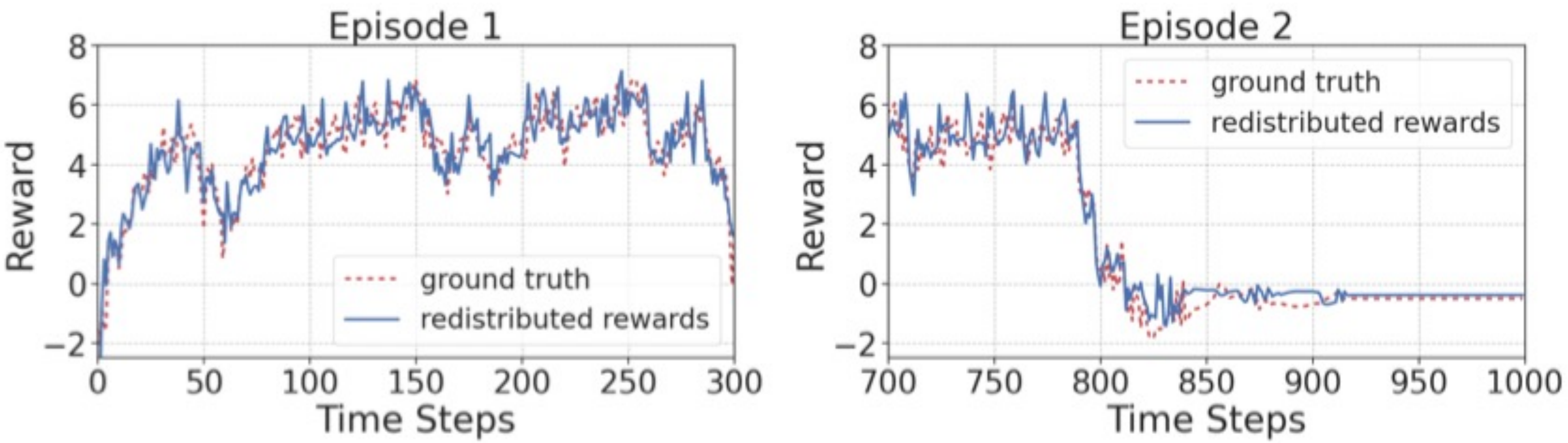


Figure 5: The visualization of redistributted rewards (blue solid lines) and the grounded rewards (red dotted lines).

[1] Yudi Zhang, Yali Du, Biwei Huang, Ziyan Wang, Jun Wang, Meng Fang, and Mykola Pechenizkiy. "Interpretable Reward Redistribution in Reinforcement Learning: A Causal Approach." NeurIPS 2023.

Thanks!

Paper : https://arxiv.org/abs/2305.18427
Project Page: https://reedzyd.github.io/GenerativeReturnDecomposition/

Project

Find Me