Agent-Controller Representations:

Principled Offline RL with Rich Exogenous Information

DongHu Kim

Today:

Exogenous vs Endogenous

EX-BMDP

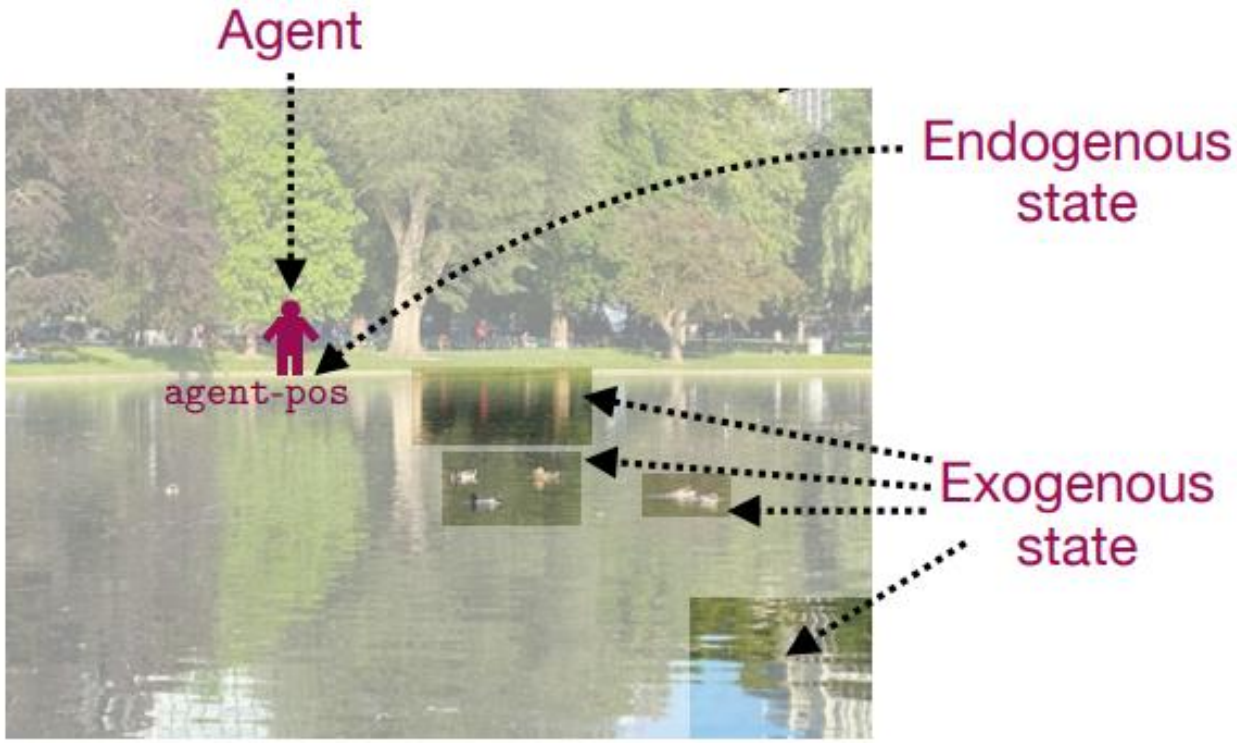
Why other methods fail

Multi-step IDM (ACRO)

Why ACRO works (theoretically)

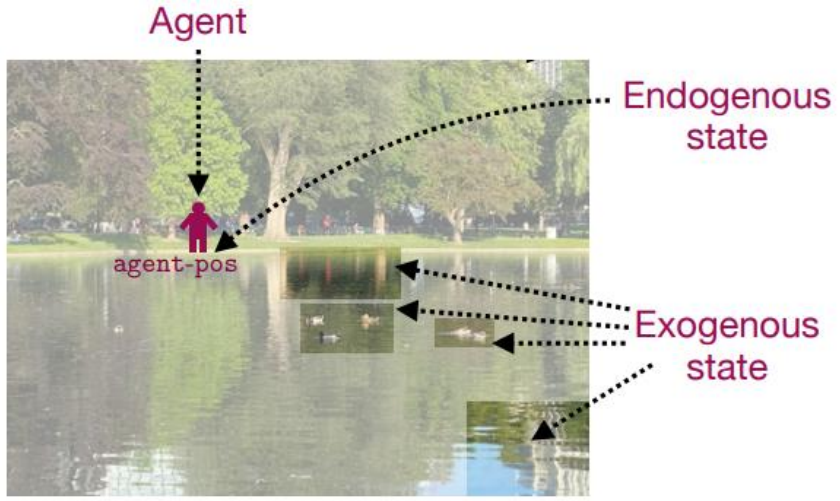
Exogenous vs Endogenous

- In real life, only a part of your visuals are necessary for choosing actions



Exogenous vs Endogenous

- Endogenous [내인성]: Agent-centric information → Necessary!
- Exogenous [외인성]: Agent-unrelated information → Waste of capacity!
- Ambiguity; gray-areas exist (e.g. the bird suddenly flies towards you)

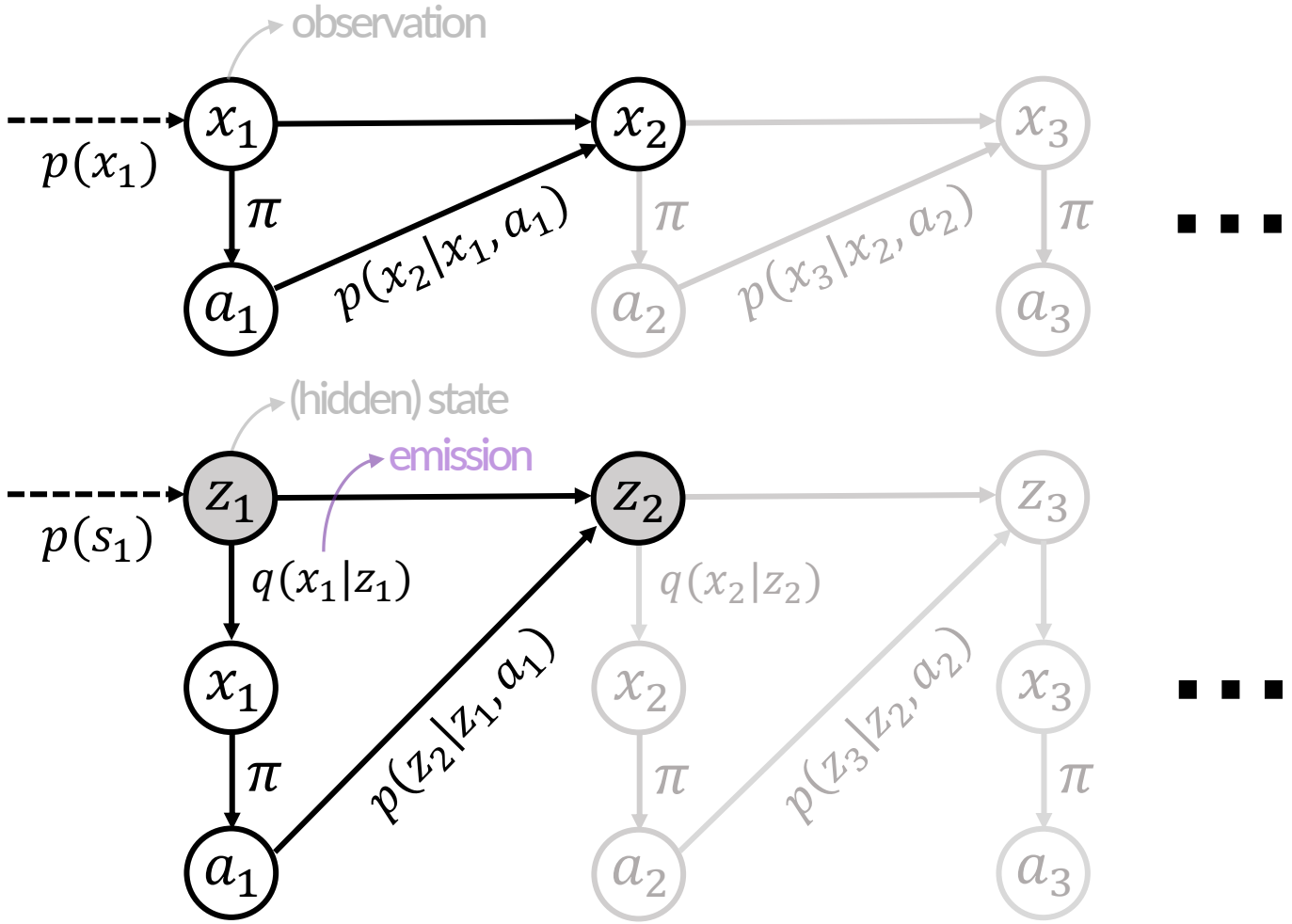


EX-BMDP

MDP

Block-MDP

MDP + Latent state space

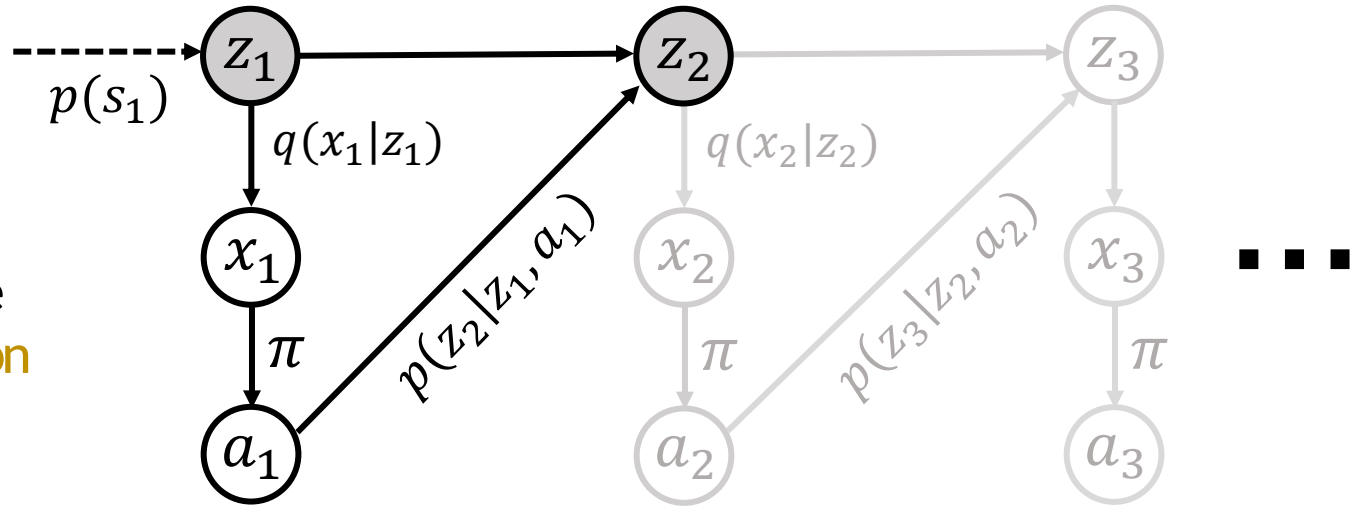


BMDP vs POMDP

Block-MDP

MDP + Latent state space

POMDP + Block assumption



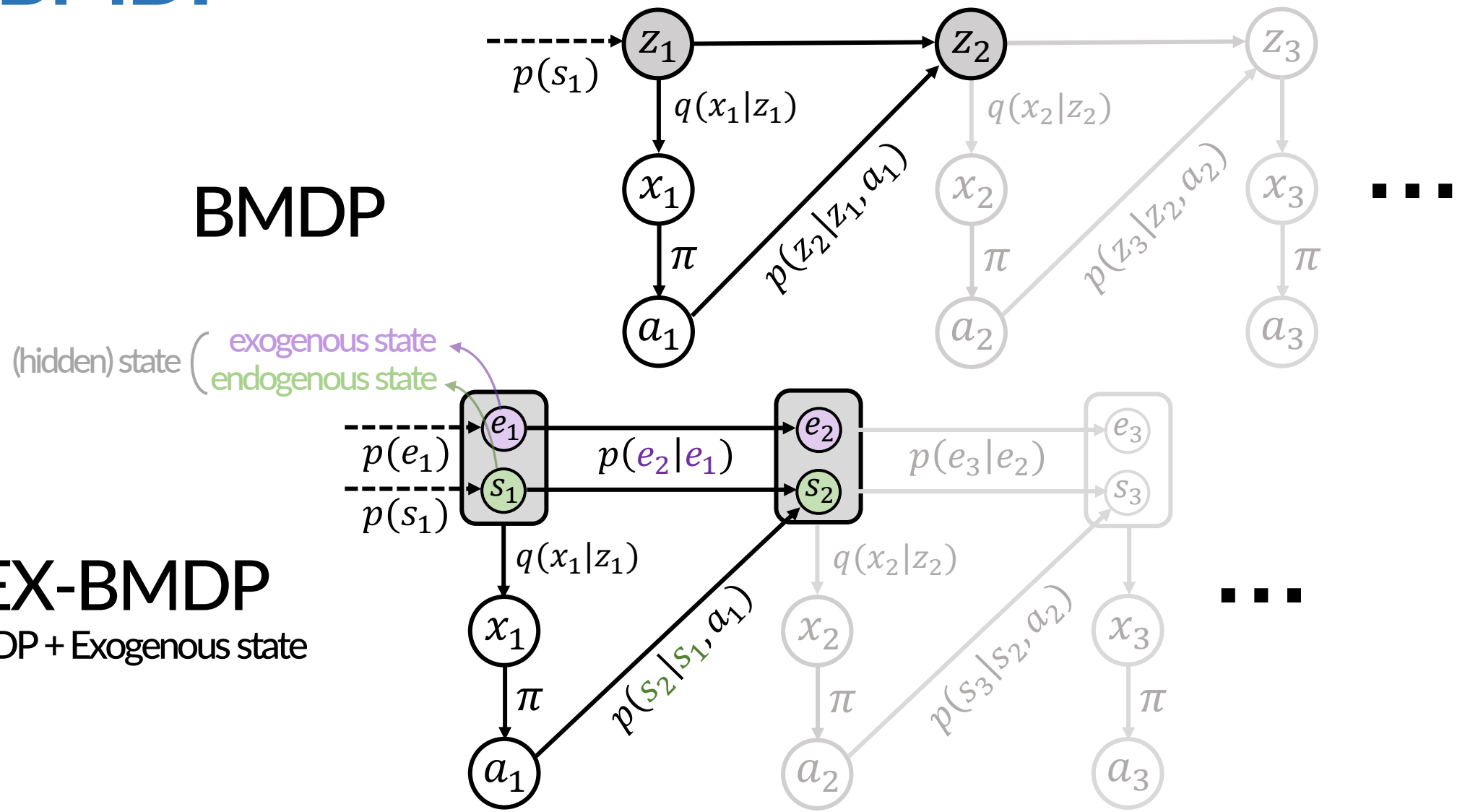
- Block assumption: No two latent states can make the same observation

$$supp(q(\cdot|s_1)) \cap supp(q(\cdot|s_2)) = \emptyset, s_1 \neq s_2$$

i.e. There exists a unique mapping from observations to states $\phi_*: \mathcal{X} \longrightarrow \mathcal{S}$

Often assumed in theoretical proofs!

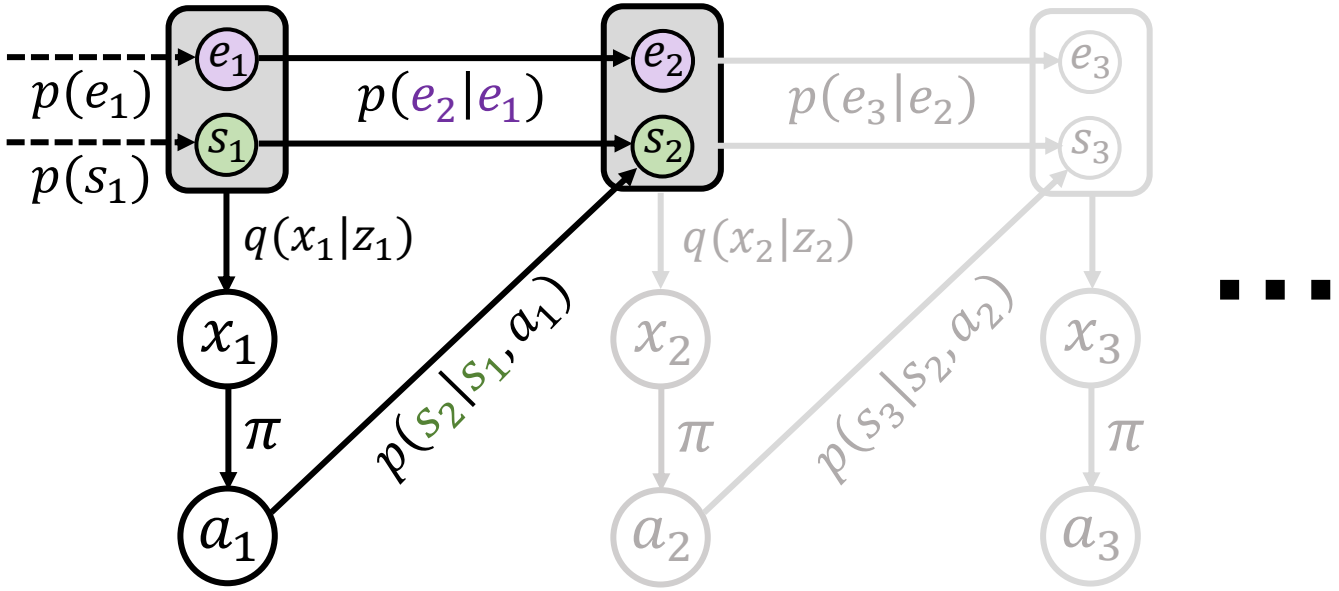
EX-BMDP


BMDP
z_1
$p(s_1)$
$q(x_1|z_1)$
x_1
π
a_1
$p(z_2|z_1, a_1)$
z_2
$q(x_2|z_2)$
x_2
π
a_2
$p(z_3|z_2, a_2)$
z_3
x_3
π
a_3
(hidden) state
exogenous state
endogenous state
e_1
s_1
$p(e_1)$
$p(s_1)$
$p(e_2|e_1)$
e_2
s_2
$p(e_3|e_2)$
e_3
s_3
EX-BMDP
BMDP + Exogenous state
$q(x_1|z_1)$
x_1
π
a_1
$p(s_2|s_1, a_1)$
$q(x_2|z_2)$
x_2
π
a_2
$p(s_3|s_2, a_2)$
x_3
π
a_3

EX-BMDP

EX-BMDP

BMDP + Exogenous state



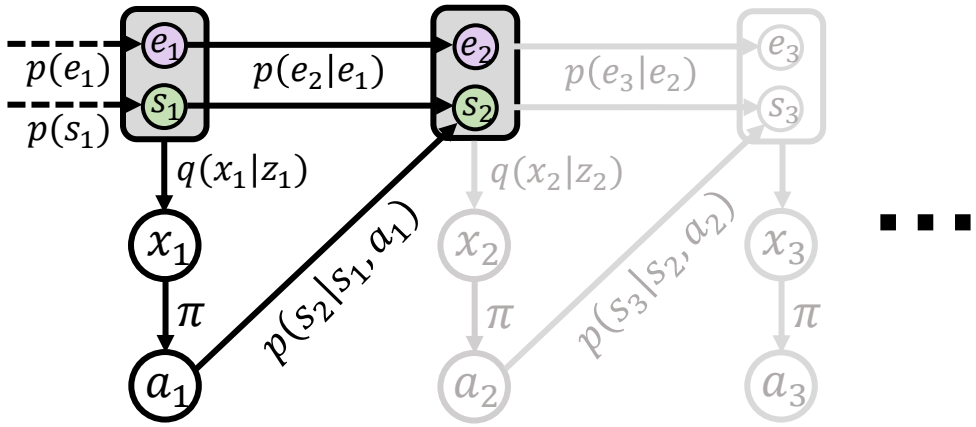
- Exogenous state also has temporally correlated dynamics, but independent from actions.

  State transition can be decoupled into: $p(z_{t+1}|z_t, a_t) = p(s_{t+1}|s_t, a_t) \cdot p(e_{t+1}|e_t)$
- Doesn't affect the reward function: $r(x_t, a_t) = r(z_t, a_t) = r(s_t, a_t)$
- Block assumption still holds(even for exogenous state): $\phi_*: \mathcal{X} \longrightarrow \mathcal{S}$, $\phi_{*,e}: \mathcal{X} \longrightarrow \mathcal{E}$

Offline-RL on EX-BMDP

EX-BMDP

BMDP + Exogenous state



GOAL: Pretrain an encoder that models endogenous info and discards exogenous info!

1. Can the algorithm fully recover endogenous information? (Full Rep.)
2. Can the algorithm discard time-independent exogenous information? (Time-Ind. Exo. Inv.)
3. Can the algorithm discard temporally correlated exogenous information? (Exogenous Invariant)
4. Can the algorithm learn without reward signal? (Reward Free)
5. Can the algorithm learn from non-expert policy? (Non-Expert Policy)

Previous Methods

Algorithms	TD3 (DrQ)	CURL	DRIML	DBC	AE	1-Step Inverse	Behavior Cloning	BYOL Explore	**ACRO** (Ours)
Time-Ind. Exo. Inv.	✓	✗	✓	✓	✗	✓	✓	✓	✓
Reward Free	✗	✓	✓	✗	✓	✓	✓	✓	✓
Exogenous Invariant	✗	✗	✗	✓	✗	✓	✓	?	✓
Non-Expert Policy	✓	✓	✓	✓	✓	✓	✗	✓	✓
Full Rep.	✓	✗	✓	✗	✓	✗	✓	✗	✓

GOAL: Pretrain an encoder that models endogenous info and discards exogenous info!

1. Can the algorithm fully recover endogenous information? (Full Rep.)
2. Can the algorithm discard time-independent exogenous information? (Time-Ind. Exo. Inv.)
3. Can the algorithm discard temporally correlated exogenous information? (Exogenous Invariant)
4. Can the algorithm learn without reward signal? (Reward Free)
5. Can the algorithm learn from non-expert policy? (Non-Expert Policy)

Previous Methods

Algorithms	TD3 (DrQ)	CURL	DRIML	DBC	AE	1-Step Inverse	Behavior Cloning	BYOL Explore	**ACRO** (Ours)
Time-Ind. Exo. Inv.	✓	✗	✓	✓	✗	✓	✓	✓	✓
Reward Free	✗	✓	✓	✗	✓	✓	✓	✓	✓
Exogenous Invariant	✗	✗	✗	✓	✗	✓	✓	?	✓
Non-Expert Policy	✓	✓	✓	✓	✓	✓	✗	✓	✓
Full Rep.	✓	✗	✓	✗	✓	✗	✓	✗	✓

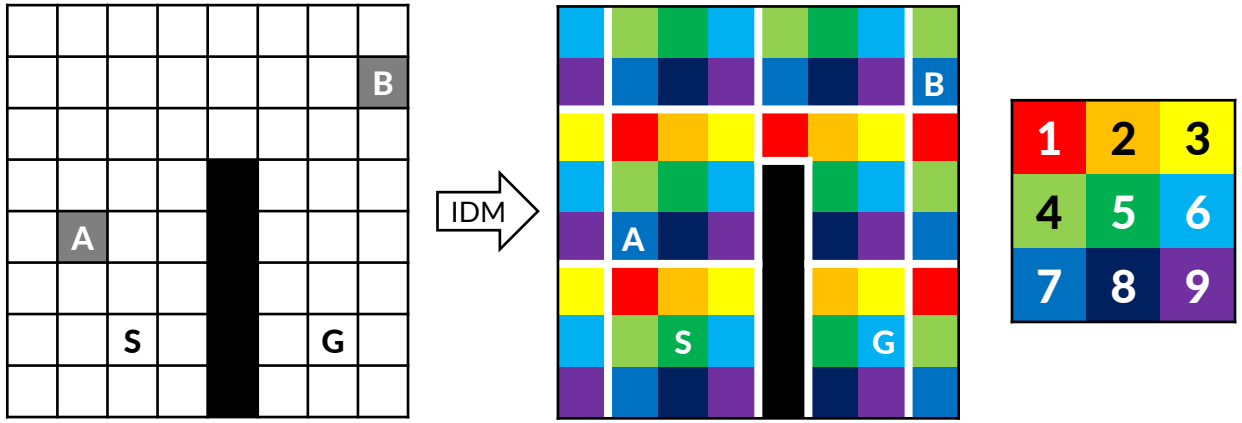
- Augmentation Contrastive Methods (e.g. CURL)
  - Augmentation invariant representation doesn't guarantee exo-invariance.
  - Augmentation function may remove some endogenous information.
- Temporal Contrastive Methods (e.g. DRIML, HOMER, ATC, …)
  - Counter-example: If the agent is staying still, observations should be encoded into identical representation.

    That's possible only if the model cannot tell them apart, but they can (and will) use exogenous information!

Previous Methods

Algorithms	TD3 (DrQ)	CURL	DRIML	DBC	AE	1-Step Inverse	Behavior Cloning	BYOL Explore	**ACRO** (Ours)
Time-Ind. Exo. Inv.	✓	✗	✓	✓	✗	✓	✓	✓	✓
Reward Free	✗	✓	✓	✗	✓	✓	✓	✓	✓
Exogenous Invariant	✗	✗	✗	✓	✗	✓	✓	?	✓
Non-Expert Policy	✓	✓	✓	✓	✓	✓	✗	✓	✓
Full Rep.	✓	✗	✓	✗	✓	✗	✓	✗	✓

- Inverse-Dynamics Model (1-step Inverse)
  - Predict what action was taken between given two consecutive observations.
  - Counter-example: Grid World
  - IDM can map the whole grid into 9 states and still get 100% accuarcy. Agent being in A or B are obviously different states, but IDM fails.



Previous Methods

Algorithms	TD3 (DrQ)	CURL	DRIML	DBC	AE	1-Step Inverse	Behavior Cloning	BYOL Explore	**ACRO** (Ours)
Time-Ind.	✓	✗	✓	✓	✗	✓	✓	✓	✓
Exo. Inv.									
Reward Free	✗	✓	✓	✗	✓	✓	✓	✓	✓
Exogenous Invariant	✗	✗	✗	✓	✗	✓	✓	?	✓
Non-Expert Policy	✓	✓	✓	✓	✓	✓	✗	✓	✓
Full Rep.	✓	✗	✓	✗	✓	✗	✓	✗	✓

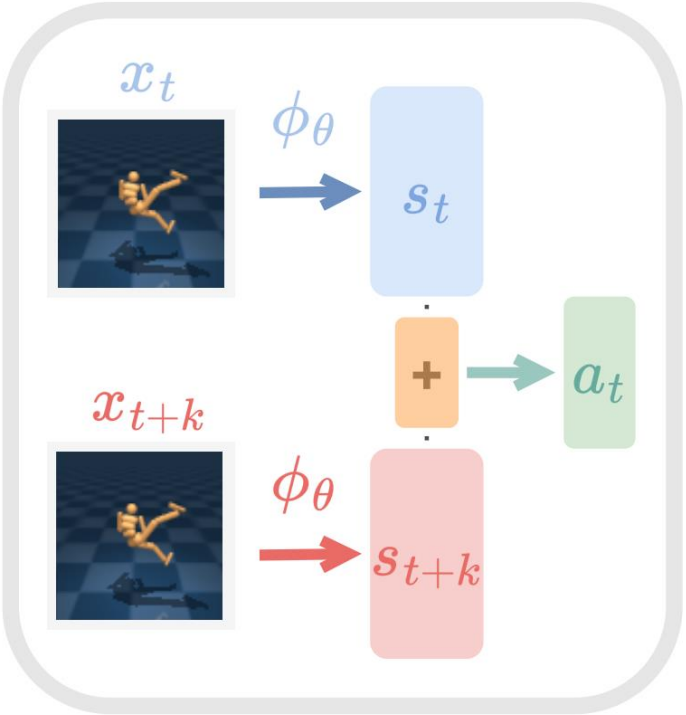
*inconsistency between table and text

- Behavior Cloning (Imitation Learning)
  - Requires data collection policy to be exogenous invariant (expert policy).
- Predictive Method: Autoencoder ~~DrQ~~
  - Learns the whole latent space, without distinguishing endogenous and exogenous.
- Reward-dependent Method: Bisimulation(DBC) ~~BYOL Explore~~
  - Counter-example: If the reward is constant, the model cannot distinguish anything.

Multi-step Inverse Modeling (ACRO)

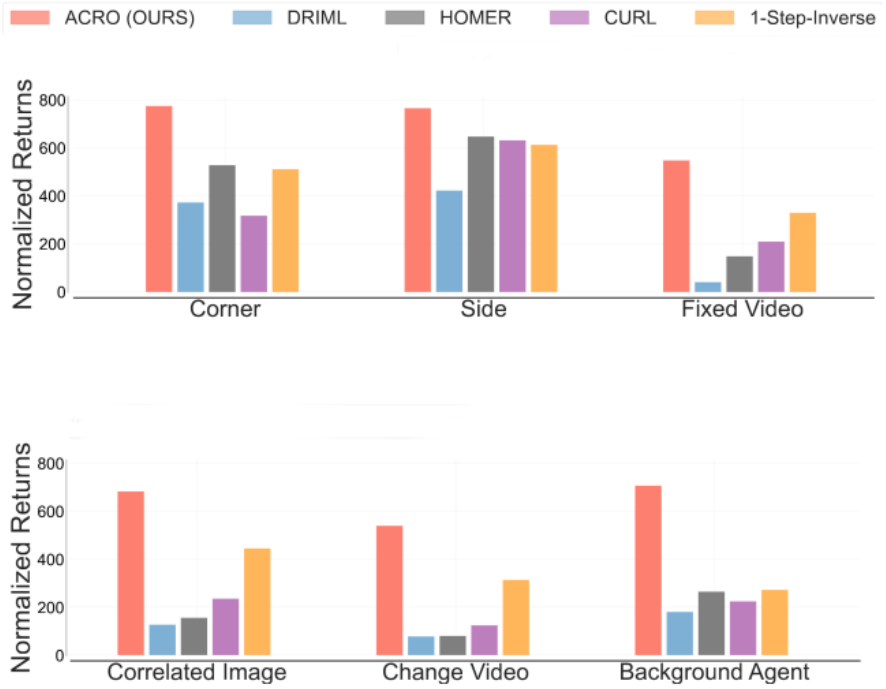
Inverse Dynamics Modeling but with k-step gap:

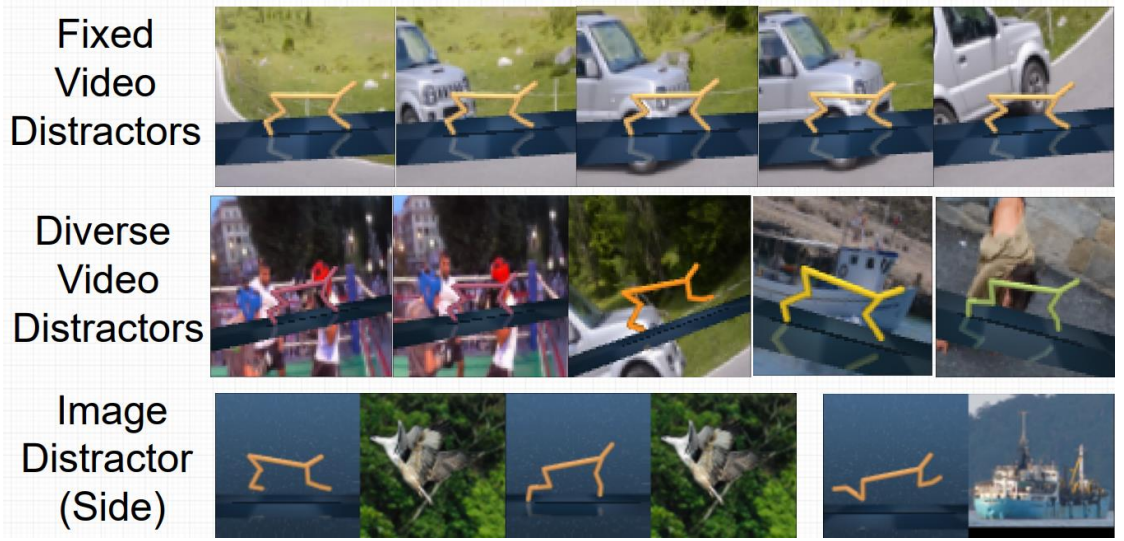
$$\phi_\star \in \arg\max_{\phi \in \Phi} \mathop{\mathbb{E}}_{\substack{t \sim U(0,N), \\ k \sim U(0,K)}} \log\left(\mathbb{P}(a_t \mid \phi(x_t), \phi(x_{t+k}))\right)$$



Multi-step Inverse Modeling (ACRO)

Hey it works!





…but how?

But how?

Three main questions:

1. How can ACRO ignore exogenous information?

2. Is endogenous-only representation really good for Offline RL?

3. **How can ACRO fully recover endogenous information**(when one-step IDM cannot)**?**

How can ACRO ignore exogenous information?

Invariance Lemma (Efroni et al.)

- ACRO objective can be optimized without relying on exogenous information.

$$\hat{\phi} = argmax_{\phi} \log(\ \mathbb{P}(a_t | x_t, x_{t+k}; \phi)\ )$$

- …as long as the dataset collection policy was exogenous-invariant.

  Given a policy π that doesn't depend on exogenous information(exo-free),

$$\mathbb{P}_{\pi}(a_t | x_t, x_{t+k}) = \mathbb{P}_{\pi}(a_t | \phi_*(x_t), \phi_*(x_{t+k}))$$

Why is 'endogenous-only' so important?

Main assumptions used in provable Offline-RL (Foster et al.)

1. The function class $\mathcal{F}$ contains the optimal Q-function. (Realizability)
2. The data distribution is sufficiently diverse. (Concentrability)
3. The function class $\mathcal{F}$ is closed under the Bellman operator. (Bellman completeness)

   Any $f \in \mathcal{F}$ still stays $\mathcal{F}$ after updating by Bellman operator: $\mathcal{T}f \in \mathcal{F}$

Proposition 2.2, 2.3

- A function class built upon an endogenous encoder $\mathcal{F}(\phi_*)$ is Bellman complete, and adding some exogenous information may break that.

How come multi-step IDM works but not single-step?

In theoretical vein, IDM is used for provable exploration in BMDP. (Efroni et al., Du et al., …)

- Intuition: In a deterministic setting, two observations from the same latent state must have identical backward dynamics!

- This idea is (likely) only possible in recursive/exhaustive exploration algorithms, and there's no explanation to how the same logic can apply to representation learning.

- Therotical paper of ACRO (Efroni et al.) claim their algorithm learns inverse dynamics modeling, but the action information only exist for recreating the path and not prediction.

How come multi-step IDM works but not single-step?

In theoretical vein, IDM is used for provable exploration in BMDP. (Efroni et al., Du et al., …)

- In the grid world counter-example, multi-step IDM may also suffer the same failure in a much larger grid world (e.g. 10000x10000).

- For multi-step IDM to succeed, we can:
  1. Predict every single action between current & initial observation. (intractable, unstable)
  2. Use an expert policy for data-collection (why not just use BC then?)

Citation/Further readings

Agent Controller Representations: Principled Offline RL with Rich Exogenous Information (Islam et al.)

Provably filtering exogenous distractors using multistep inverse dynamics (Efroni et al.)

Provably efficient RL with Rich Observations via Latent State Decoding (Du et al.)

On oracle-efficient PAC reinforcement learning with rich observations (Dann et al.)

PAC Reinforcement Learning with Rich Observations (Krishnamurthy et al.)

Kinematic State Abstraction and Provably Efficient Rich-Observation Reinforcement Learning (Misra et al.)

Guaranteed Discovery of Control-Endogenous Latent States with Multi-Step Inverse Models (Lamb et al.)

Sample-efficient reinforcement learning in the presence of exogenous information. (Efroni et al.)