Invariant Prediction for Generalization in Reinforcement Learning

Clare Lyle

OATML Group, University of Oxford

Credits

This talk is based off of joint work with some fantastic collaborators and advisors.

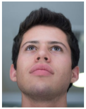
Amy Zhang

Shagun Sodhani

Angelos Filos

Marta Kwiatkowska

Joelle Pineau



Yarin Gal

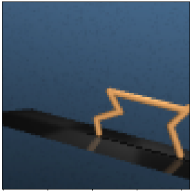
Doina Precup

github.com/facebookreserach/icp-block-mdp

https://arxiv.org/abs/2003.06016

A not-so-simple problem

Train RL agent on environments $\mathcal{E}_1$, $\mathcal{E}_2$

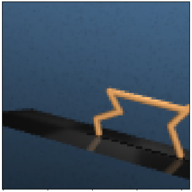


Deploy RL agent on environment $\mathcal{E}_3$

Standard deep RL methods **fail to generalize** to new environments
even when the new environments **share a similar structure** with the training environments

A not-so-simple problem

Train RL agent on environments $\mathcal{E}_1$, $\mathcal{E}_2$



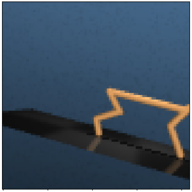
Deploy RL agent on environment $\mathcal{E}_3$

Standard deep RL methods **fail to generalize** to new environments

even when the new environments **share a similar structure** with the training environments

A not-so-simple problem

Train RL agent on environments $\mathcal{E}_1$, $\mathcal{E}_2$



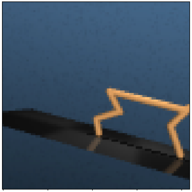
Deploy RL agent on environment $\mathcal{E}_3$

Standard deep RL methods **fail to generalize** to new environments

even when the new environments **share a similar structure** with the training environments

A not-so-simple problem

Train RL agent on environments $\mathcal{E}_1$, $\mathcal{E}_2$



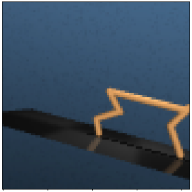
Deploy RL agent on environment $\mathcal{E}_3$

Standard deep RL methods **fail to generalize** to new environments

even when the new environments **share a similar structure** with the training environments

A not-so-simple problem

Train RL agent on environments $\mathcal{E}_1$, $\mathcal{E}_2$



Deploy RL agent on environment $\mathcal{E}_3$

Standard deep RL methods **fail to generalize** to new environments
even when the new environments **share a similar structure** with the training environments

A simple(?) solution

1. Want RL agents that generalize to new environments where **the underlying MDP is the same**.
2. So find a representation that maps equivalent observations from different environments to the same abstract state.
3. Use an idea from the causal inference world to find this representation: **invariant prediction**.

A simple(?) solution

1. Want RL agents that generalize to new environments where **the underlying MDP is the same**.
2. So find a representation that maps equivalent observations from different environments to the same abstract state.
3. Use an idea from the causal inference world to find this representation: **invariant prediction**.

A simple(?) solution

1. Want RL agents that generalize to new environments where **the underlying MDP is the same**.
2. So find a representation that maps equivalent observations from different environments to the same abstract state.
3. Use an idea from the causal inference world to find this representation: **invariant prediction**.

TL;DR

1. Identifying causal variables in the state space $\equiv$ finding *model irrelevance state abstractions* (MISAs)
2. Leveraging the shared structure of the environments leads to more environment-efficient generalization bounds.
3. Even when exact inference is impossible (i.e. deep RL with rich observations), learning an invariant representation leads to improved generalization.

TL;DR

1. Identifying causal variables in the state space $\equiv$ finding *model irrelevance state abstractions* (MISAs)
2. Leveraging the shared structure of the environments leads to more environment-efficient generalization bounds.
3. Even when exact inference is impossible (i.e. deep RL with rich observations), learning an invariant representation leads to improved generalization.

TL;DR

1. Identifying causal variables in the state space $\equiv$ finding *model irrelevance state abstractions* (MISAs)
2. Leveraging the shared structure of the environments leads to more environment-efficient generalization bounds.
3. Even when exact inference is impossible (i.e. deep RL with rich observations), learning an invariant representation leads to improved generalization.

TL;DR

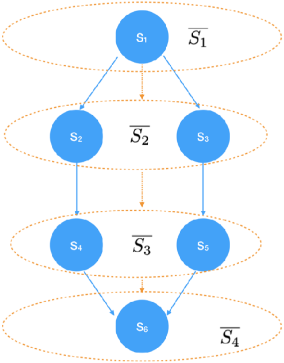
1. Identifying causal variables in the state space $\equiv$ finding *model irrelevance state abstractions* (MISAs)
2. Leveraging the shared structure of the environments leads to more environment-efficient generalization bounds.
3. Even when exact inference is impossible (i.e. deep RL with rich observations), learning an invariant representation leads to improved generalization.

Details

State Abstractions

State Abstractions

A **state abstraction** is a function $\phi : \mathcal{S} \rightarrow \bar{\mathcal{S}}$ which maps states $s \in \mathcal{S}$ to simpler abstract state space $\bar{\mathcal{S}}$. This can make it easier for an agent to learn and plan.



MISAs

A **model-irrelevance state abstraction** (MISA) is a state abstraction that preserves the reward function and transition dynamics of the MDP (Li et al.). i.e.

$$\phi(s) = \phi(s') \implies R(s) = R(s')$$

and $\sum p(s''|s) = \sum p(s''|s')$

State Abstractions

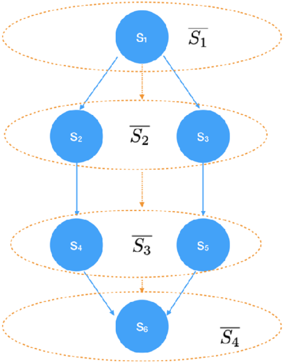
A **state abstraction** is a function $\phi : \mathcal{S} \to \bar{\mathcal{S}}$ which maps states $s \in \mathcal{S}$ to simpler abstract state space $\bar{\mathcal{S}}$. This can make it easier for an agent to learn and plan.

MISAs

A **model-irrelevance state abstraction** (MISA) is a state abstraction that preserves the reward function and transition dynamics of the MDP (Li et al.). i.e.

$$\phi(s) = \phi(s') \implies R(s) = R(s')$$

and $\sum \quad p(s''|s) = \sum \quad p(s''|s')$



State Abstractions

State Abstractions

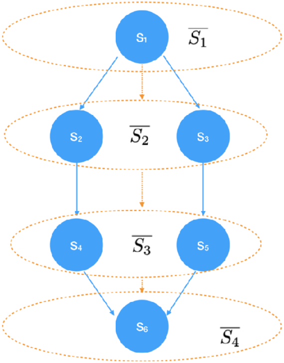
A **state abstraction** is a function $\phi : \mathcal{S} \to \bar{\mathcal{S}}$ which maps states $s \in \mathcal{S}$ to simpler abstract state space $\bar{\mathcal{S}}$. This can make it easier for an agent to learn and plan.

MISAs

A **model-irrelevance state abstraction** (MISA) is a state abstraction that preserves the reward function and transition dynamics of the MDP (Li et al.). i.e.

$$\phi(s) = \phi(s') \implies R(s) = R(s')$$

$$\text{and} \quad \sum p(s''|s) = \sum p(s''|s')$$



State Abstractions in Deep RL

1. Canonical state abstraction model

$$\mathcal{X} \to \Phi = \phi(\mathcal{X}) \to \hat{V} = f(\Phi)$$

2. Deep RL model

$$\underbrace{\overbrace{\mathcal{X} \to \Phi_1 = \phi_1(\mathcal{X})}^{\Phi(X)} \to \Phi_2 \cdots \to \Phi_N = \phi_N(\Phi_{N-1})}_{\Phi(X)} \underbrace{\to \hat{V} = f(\Phi_N}_{f(\Phi)}$$

3. Can think of each layer in DNN as *both* representation and value function.
4. When auxiliary loss used to train representation, will say representation layer is the one where this loss is applied.

$$\begin{tikzcd} & T \arrow[d, bend left, "\ell_{\text{rep}}"] & \\ \mathcal{X} \arrow[r, "\phi"] & \Phi \arrow[u, bend left, "g"] \arrow[r, bend left, "f"] & \hat{V} = f(\Phi) \end{tikzcd}$$

Block MDPs: Intuition

- We're interested in families of MDPs $\mathcal{M}_1, \ldots, \mathcal{M}_k$ that are 'behaviourally equiivalent'.
- I.e. want $\mathcal{M}_1, \ldots, \mathcal{M}_k$ with state spaces $\mathcal{X}_1, \ldots, \mathcal{X}_k$
- for which $\exists \phi$ s.t. $\phi(\mathcal{X}_i) = \phi(\mathcal{X}_j)$ forall i, j and ϕ is a MISA for the union $\cup_{i \in I} \mathcal{X}_i$.
- **Question:** how can we learn ϕ from a subset of the environments $\{\mathcal{M}_i\}$?
- **Answer:** use the causal structure of the problem.

Block MDPs: Intuition

- We're interested in families of MDPs $\mathcal{M}_1, \ldots, \mathcal{M}_k$ that are 'behaviourally equiivalent'.
- I.e. want $\mathcal{M}_1, \ldots, \mathcal{M}_k$ with state spaces $\mathcal{X}_1, \ldots, \mathcal{X}_k$
- for which $\exists \phi$ s.t. $\phi(\mathcal{X}_i) = \phi(\mathcal{X}_j)$ forall i, j and ϕ is a MISA for the union $\cup_{i \in I} \mathcal{X}_i$.
- **Question:** how can we learn ϕ from a subset of the environments $\{\mathcal{M}_i\}$?
- **Answer:** use the causal structure of the problem.

Block MDPs: Intuition

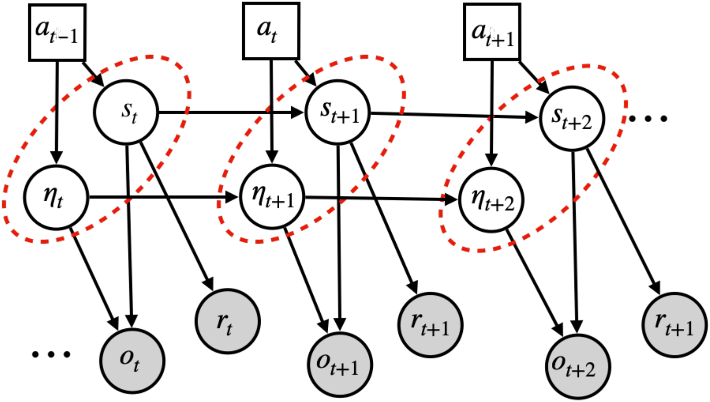
- We're interested in families of MDPs $\mathcal{M}_1, \dots, \mathcal{M}_k$ that are 'behaviourally equiivalent'.
- I.e. want $\mathcal{M}_1, \dots, \mathcal{M}_k$ with state spaces $\mathcal{X}_1, \dots, \mathcal{X}_k$
- for which $\exists\phi$ s.t. $\phi(\mathcal{X}_i) = \phi(\mathcal{X}_j)$ forall i, j and ϕ is a MISA for the union $\cup_{i \in I}\mathcal{X}_i$.
- **Question:** how can we learn ϕ from a subset of the environments $\{\mathcal{M}_i\}$?
- **Answer:** use the causal structure of the problem.

Block MDPs: Intuition

- We're interested in families of MDPs $\mathcal{M}_1, \ldots, \mathcal{M}_k$ that are 'behaviourally equiivalent'.
- I.e. want $\mathcal{M}_1, \ldots, \mathcal{M}_k$ with state spaces $\mathcal{X}_1, \ldots, \mathcal{X}_k$
- for which $\exists \phi$ s.t. $\phi(\mathcal{X}_i) = \phi(\mathcal{X}_j)$ forall i, j and ϕ is a MISA for the union $\cup_{i \in I} \mathcal{X}_i$.
- **Question:** how can we learn ϕ from a subset of the environments $\{\mathcal{M}_i\}$?
- **Answer:** use the causal structure of the problem.

Causal Structure

Decompose agent's observation o_t into 'causal' and 'spurious' components s_t and η_t.



Block MDP

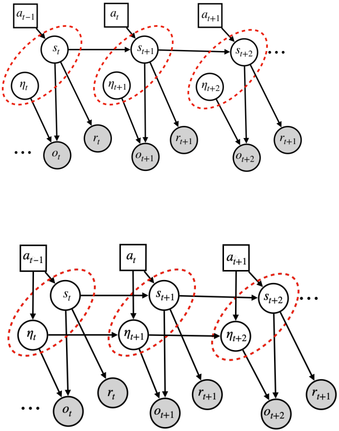


Figure 2: Top: Graphical model for a Block MDP. The observation o_t is modelled here as a function of the state s_t and a noise variable η_t. In a Block MDP, there is

Definition

A Block MDP is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{X}, p, q, R \rangle$

- unobservable state space $\mathcal{S}$
- finite action space $\mathcal{A}$
- observation space $\mathcal{X}$
- transition distribution p
- reward function R
- (injective) emission function $q : \mathcal{S} \to \mathcal{X}$

Invariant Causal Prediction: Intuition

- We want a way of leveraging data collected from many environments to find a representation that captures the causal structure of the underlying dynamics model.
- ICP Hypothesis:

$$\textbf{Causality} \iff \textbf{Invariance} \tag{1}$$

- i.e. given a set of environments corresponding to *interventions* on variables in the causal graph, a predictor that depends on variables that are causal parents of the target will be invariant across the environments.
- Can use this invariance criterion as a means of selecting (or even learning) a representation.

Invariant Causal Prediction

<1>

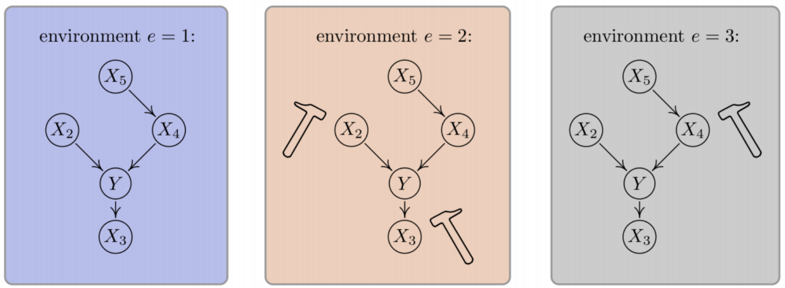


Figure 1: An example including three environments. The invariance (1) and (2) holds if we consider $S^* = \{X_2, X_4\}$. Considering indirect causes instead of direct ones (e.g. $\{X_2, X_5\}$) or an incomplete set of direct causes (e.g. $\{X_4\}$) may not be sufficient to guarantee invariant prediction.

Figure 3: Invariant Causal Prediction (Peters et al., 2016)

Assumptions

- **Assumption 1**: The observation space of a Block MDP is fully observable, and therefore exhibits the Markov property.
- **Assumption 2**: The components of the current observation are independent conditioned on the previous observation, i.e.

$$p(X^1_{t+1}|X_t, X^2_{t+1}) = P(X^1_{t+1}|X_t) \tag{2}$$

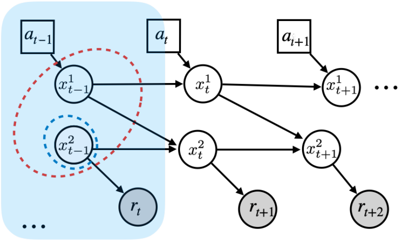


Figure 4: Graphical model demonstrating assumption 2.

- **Assumption 3**: The training

Assumptions

- **Assumption 1**: The observation space of a Block MDP is fully observable, and therefore exhibits the Markov property.
- **Assumption 2**: The components of the current observation are independent conditioned on the previous observation, i.e.

$$p(X^1_{t+1}|X_t, X^2_{t+1}) = P(X^1_{t+1}|X_t) \tag{2}$$

- **Assumption 3**: The training

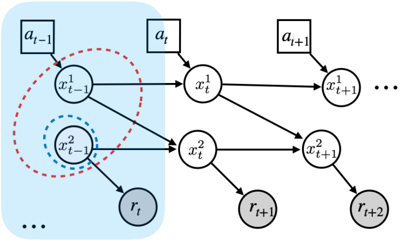


Figure 4: Graphical model demonstrating assumption 2.

Assumptions

- **Assumption 1**: The observation space of a Block MDP is fully observable, and therefore exhibits the Markov property.
- **Assumption 2**: The components of the current observation are independent conditioned on the previous observation, i.e.

$$p(X^1_{t+1}|X_t, X^2_{t+1}) = P(X^1_{t+1}|X_t) \quad (2)$$

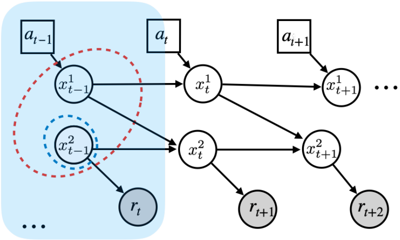


Figure 4: Graphical model demonstrating assumption 2.

- **Assumption 3**: The training

Assumptions

- **Assumption 1**: The observation space of a Block MDP is fully observable, and therefore exhibits the Markov property.
- **Assumption 2**: The components of the current observation are independent conditioned on the previous observation, i.e.

$$p(X^1_{t+1} | X_t, X^2_{t+1}) = P(X^1_{t+1} | X_t) \quad (2)$$

- **Assumption 3**: The training

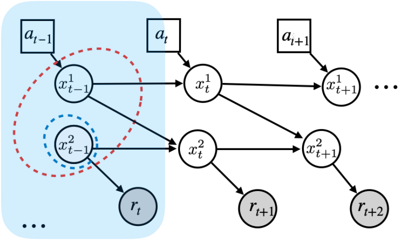


Figure 4: Graphical model demonstrating assumption 2.

Results

Causality and MISAs

Causal Variables $\iff$ State Abstractions

- Consider the setting where variables are observable: state $s = (x_1, \ldots, x_n)$.
- Take the variables which are **causal ancestors** of the return, $\bar{s} = (x_{i_1}, \ldots, x_{i_k})$
- Then the mapping $\phi : (x_1, \ldots x_n) \mapsto (x_{i_1}, \ldots, x_{i_k})$ ... is a **model irrelevance state abstraction**

Theorem 1

Let $S_R \subseteq \{1, \ldots, k\}$ be the set of variables such that the reward $R(x, a)$ is a function only of $[x]_{S_R}$ (x restricted to the indices in S_R). Then let $S = \mathbf{AN}(R)$ denote the ancestors of S_R in the (fully observable) causal graph corresponding to the transition dynamics of $M_{\mathcal{E}}$. Then the state abstraction $\phi_S(x) = [x]_S$ is a *model-irrelevance* abstraction for every $e \in \mathcal{E}$

Causality and MISAs

Causal Variables $\iff$ State Abstractions

- Consider the setting where variables are observable: state $s = (x_1, \dots, x_n)$.
- Take the variables which are **causal ancestors** of the return, $\bar{s} = (x_{i_1}, \dots, x_{i_k})$
- Then the mapping $\phi : (x_1, \dots x_n) \mapsto (x_{i_1}, \dots, x_{i_k})$ ... is a **model irrelevance state abstraction**

Theorem 1

Let $S_R \subseteq \{1, \dots, k\}$ be the set of variables such that the reward $R(x, a)$ is a function only of $[x]_{S_R}$ (x restricted to the indices in S_R). Then let $S = \mathbf{AN}(R)$ denote the ancestors of S_R in the (fully observable) causal graph corresponding to the transition dynamics of $M_{\mathcal{E}}$. Then the state abstraction $\phi_S(x) = [x]_S$ is a *model-irrelevance* abstraction for every $e \in \mathcal{E}$

Bounds on Generalization Error

Good state abstractions

MISAs generalize well to new environments because the agent can immediately apply its knowledge from previous environments.

Model error bound

Consider an MDP M, with M' denoting a coarser bisimulation of M. Let ϕ denote the mapping from states of M to states of M'. Suppose that the dynamics of M are L-Lipschitz w.r.t. $\phi(X)$ and that T is some approximate transition model satisfying $\max_s \mathbb{E}\| T(\phi(s)) - \phi(T_M(s))\| < \delta$, for some $\delta > 0$. Let $W_1(\pi_1, \pi_2)$ denote the 1-Wasserstein distance. Then

$$\mathbb{E}_{x\sim M'}[\| T(\phi(x)) - \phi(T_{M'}(x))\|] \leq \delta + 2LW_1(\pi_{\phi(M)}, \pi_{\phi(M')}). \quad (3)$$

Bounds on Generalization Error

Good state abstractions

MISAs generalize well to new environments because the agent can immediately apply its knowledge from previous environments.

Model error bound

Consider an MDP M, with M' denoting a coarser bisimulation of M. Let ϕ denote the mapping from states of M to states of M'. Suppose that the dynamics of M are L-Lipschitz w.r.t. $\phi(X)$ and that T is some approximate transition model satisfying $\max_s \mathbb{E}\| T(\phi(s)) - \phi(T_M(s))\| < \delta$, for some $\delta > 0$. Let $W_1(\pi_1, \pi_2)$ denote the 1-Wasserstein distance. Then

$$\mathbb{E}_{x\sim M'}[\| T(\phi(x)) - \phi(T_{M'}(x))\|] \leq \delta + 2LW_1(\pi_{\phi(M)}, \pi_{\phi(M')}). \quad (3)$$

Observable Variables Setting

When state is equal to the variables in the causal graph, it's straightforward to apply known causal prediction methods to find the causal ancestors of the reward.

Algorithm: ICP for Model Irrelevance State Abstractions

Result: $S \subset \{1, \ldots, k\}$, the causal state variables

Input: α, a confidence parameter, $\mathcal{D}$, an replay buffer with observations $\mathcal{X}$ (partitioned into environments $e_1, \ldots, e_k$).

$S \leftarrow \emptyset$;

stack $\leftarrow$ r ;

while *stack is not empty* **do**

  $v =$ stack.pop() ;

  if $v \notin S$ **then**

    $S' \leftarrow$ `ICP`(v, $\mathcal{D}$, $\frac{\alpha}{\dim(\mathcal{X})}$) ;

    $S \leftarrow S \cup S'$;

    stack.push(S')

Observable Variables Setting

When state is equal to the variables in the causal graph, it's straightforward to apply known causal prediction methods to find the causal ancestors of the reward.

Algorithm: ICP for Model Irrelevance State Abstractions

Result: $S \subset \{1, \ldots, k\}$, the causal state variables

Input: α, a confidence parameter, $\mathcal{D}$, an replay buffer with observations $\mathcal{X}$ (partitioned into environments $e_1, \ldots, e_k$).

$S \leftarrow \emptyset$;

stack $\leftarrow$ r ;

while *stack is not empty* **do**

  $v =$ stack.pop() ;

  if $v \notin S$ **then**

    $S' \leftarrow$ `ICP`(v, $\mathcal{D}$, $\frac{\alpha}{\dim(\mathcal{X})}$) ;

    $S \leftarrow S \cup S'$;

    stack.push(S')

Rich Observation Setting

In the rich observation setting, we can't obtain guarantees.

However, we propose a method for learning approximate MISAs.

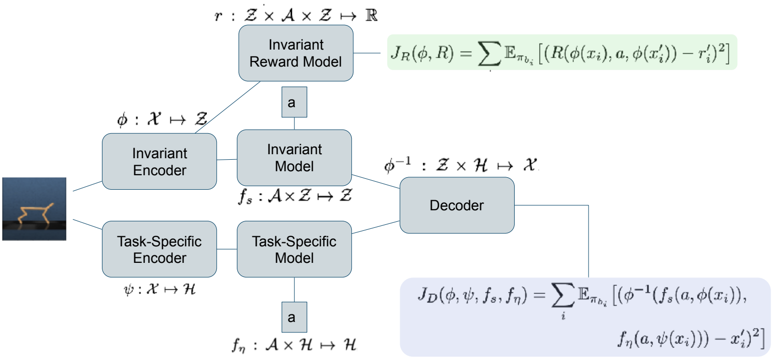
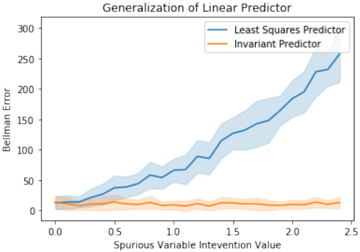
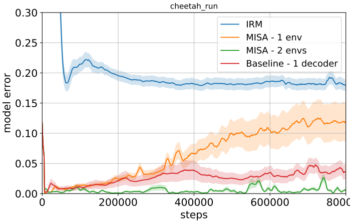


Figure 5: Architecture for Rich Observation Setting

Empirical Results

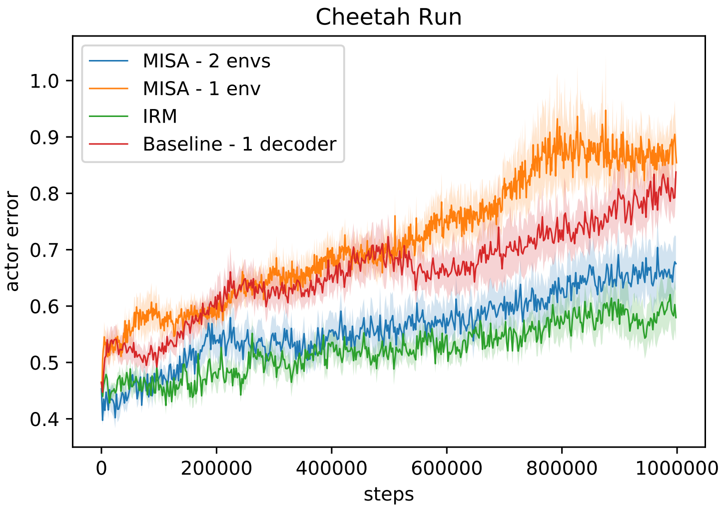
Model Learning





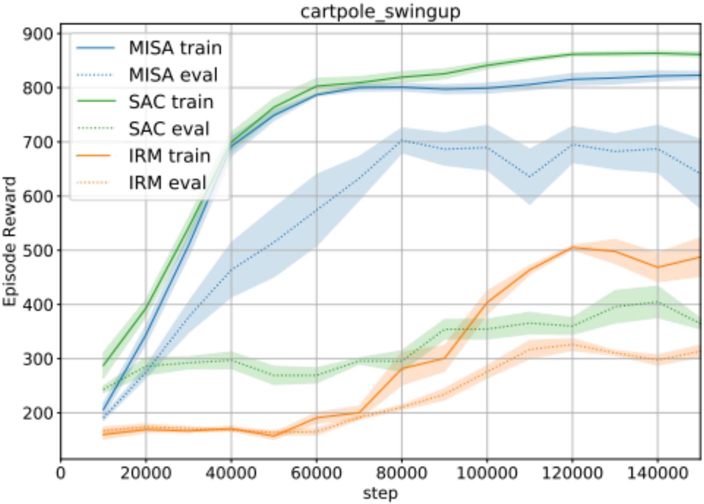
Empirical Results

Imitation Learning



Empirical Results

Reinforcement Learning



Conclusions

- We show that invariant prediction can be used to find good state abstractions that pick up on the shared causal structure between environments.
- We prove some results on how to find these state abstractions and how well they'll generalize.
- We present an approach that leverages invariant prediction to obtain improved generalization to new environments on a variety of tasks.

Thanks!