On the Role of Information Structure in Reinforcement Learning

Awni Altabaa

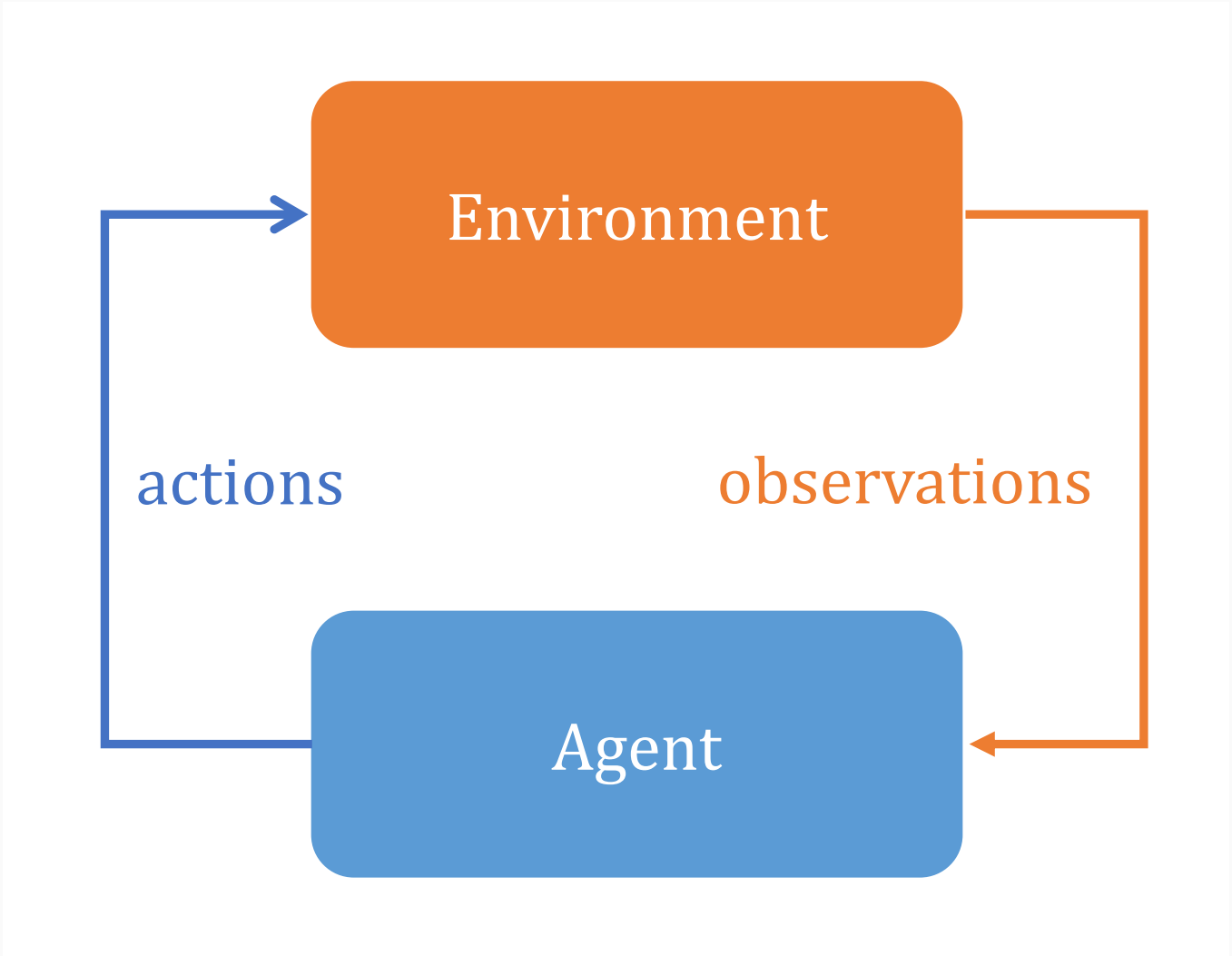
Yale University, Department of Statistics & Data Science

Table of contents

Reinforcement Learning at a High Level

What is Reinforcement Learning, Really?

At the most basic level, reinforcement learning is the problem of learning how to act through interaction with an environment.



Reinforcement Learning is Hard (in general)

Agent aims to learn a policy π which maximizes their objective.

For each choice of policy, there is an expected value for objective, $V(\pi)$

In the worst case, must try every possible policy.

\# of policies is typically $\Omega\left(|\text{action space}|^{|\text{trajectory space}|}\right)$

Entirely impractical, especially as problem scales up.

Many real-world problems have structure that makes them easier to handle.

Making RL Tractable by imposing assumptions on Information Structure

The reinforcement learning literature has identified classes of problems that are tractable.

Most commonly studied is the Markov decision-process (MDP)

Assumes that state of system is "Markovian"—future depends only on present, and not the past.

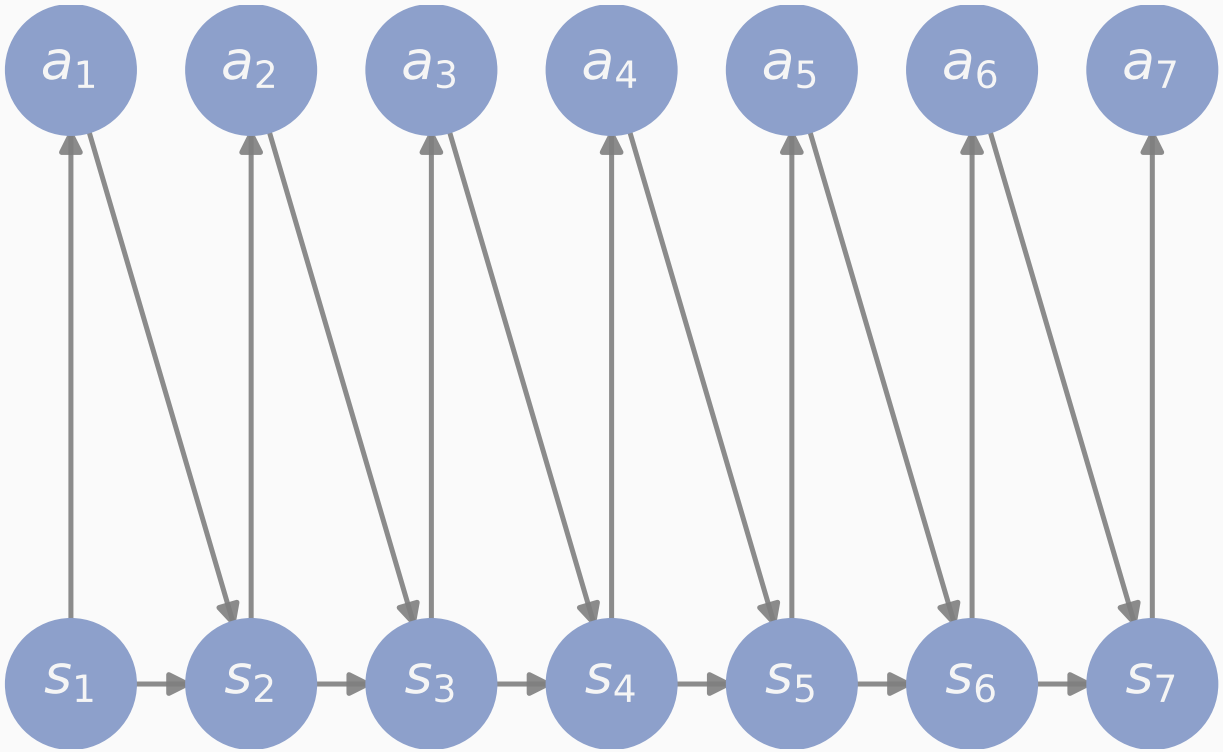
Under such an assumption, ϵ-optimal policy can be learned with

$$\text{poly}\left(S, A, H, \epsilon^{-1}\right)$$

samples, where S is size of state space, A size of action space, H is time horizon.

Compared to $\Omega(((SA)^H)^A)$ in worst case...

When state is Markovian, learning is tractable



The Real-world is not so simple...

This kind of assumption can be captured in the language of **information structure**.

Information Structure = how events in the system occurring at different points in time affect each other

MDPs make a very specific assumption on the system to make the problem tractable.

Real-world sequential decision-making problems involve a complex and time-varying interdependence of system variables

Information Structure

In general, can think of a system as a sequence of variables or events

$$X_1, X_2, X_3, \ldots, X_T$$

Each variable is either a system variable, describing some aspect of the state of the system, or an action by the event.

The *information structure* describes, for each event of the system, what subset of past events it depends on.

Example: MDPs assume the simplest possible information structure—only immediate past.

Importance of "information structure" has long been recognized by control community, but is comparatively unexplored in RL.

Can we characterize the statistical complexity of general reinforcement learning problems via their information structure?

A General Model that Captures Information Structure

A General Model that Captures Information Structure

We need a general model equipped to capture information structure.

Controlled stochastic process: $X_1, \ldots, X_T$

Need several ingredients:

1. **Variable Structure.** Which variables are system variables and which are action variables. Partition of $[T] = \mathcal{S} \cup \mathcal{A}$
2. **Information Structure.** For $t \in [T]$, subset of past variables on which X_t depends. $\mathcal{I}_t \subset [t-1]$. Defines "information space" $\mathbb{I}_t := \prod_{s \in \mathcal{T}_t} \mathbb{X}_s$.
3. **System Kernels.** $\mathcal{T}_t \in \mathcal{P}(\mathbb{X}_t \mid \mathbb{I}_t)$ such that $X_t \sim \mathcal{T}_t(\cdot \mid \{x_s, s \in \mathcal{I}_t\})$.
4. **Observability.** Subset of system variables observable to learning algorithm $\mathcal{O} \subset \mathcal{S}$.
5. **Reward structure.** Reward function(s) for each agent.

A General Model that Captures Information Structure

For system variables $t \in \mathcal{S}$, information set $\mathcal{I}_t$ describes an aspect of system dynamics.

For action variables $t \in \mathcal{A}$, information set $\mathcal{I}_t$ describes information available to agent at time of making decision.

This is an extremely general model.

- Captures events occurring simultaneously or sequentially.
- Captures single-agent or multi-agent problems
- Captures arbitrary system dynamics

E.g., commonly studied models like MDP, POMDP, Dec-POMDP, POMG, are special cases.

Preview: Statistical Complexity Characterized via Information Structure

Theorem (Preview)

For any sequential decision-making problem with information with information structure $\mathcal{I} = \{\mathcal{I}_t, t \in [T]\}$, the sample complexity of learning a near-optimal policy scales as $f(\mathcal{I})$, for some function of the information structure.

For the remainder of the talk, we will build towards this result.

Characterizing the "Complexity" of the Dynamics

A notion of complexity

We are interested in the complexity of the *observable dynamics*.

Full description of the system is

$$X_1, X_2, \ldots, X_T$$

But, learning agent can only observe and model

$$(X_h)_{h \in \mathcal{O} \cup \mathcal{A}} \coloneqq \left(X_{t(1)}, \ldots, X_{t(H)}\right)$$

(Here, we index observables by $h \in [H]$, $H \coloneqq |\mathcal{O} \cup \mathcal{A}|$)

To gauge how difficult it is to learn, we need to characterize the "complexity" of the dynamics of the observable system variables

A notion of complexity

A commonly studied notion of the complexity of dynamics is "rank".

For $h \in [H]$, define the dynamics matrix

$$[D_h]_{\text{history},\text{future}} := \mathbb{P}\left[\text{history}, \text{future} \mid \text{actions}\right]$$

Definition

The rank of the dynamics matrices D_h, $r_h := \operatorname{rank}(D_h)$, $r := \max_h r_h$, characterize a notion of complexity of the observable system dynamics.

For our model, these are histories and futures of *observable* variables.

Information-Structural Characterization of Rank of Dynamics

The information structure of a sequential decision-making problem determines an upper bound on its rank.

Information structure can naturally be represented by a DAG $\mathcal{G}$ where edges are determined by the information sets. I.e., $(i,t) \in \mathcal{I} \iff i \in \mathcal{I}_t$.

We will derive a characterization of the rank in terms of this DAG representation.

Definition (Information-Structural State)

Let $G^\dagger$ be the DAG obtained from $\mathcal{G}$ by removing incoming edges towards actions. For each $h \in [H]$, let $\mathcal{I}_h^\dagger \subset [t(h)]$ be the minimal set of past variables (observed or unobserved) which d-separate the past observations $(X_{t(1)}, \dots, X_{t(h)})$ from the future observations $(X_{t(h+1)}, \dots, X_{t(H)})$ in the DAG $\mathcal{G}^\dagger$. Define $\mathbb{I}_h^\dagger := \prod_{s \in \mathcal{I}_h^\dagger} \mathbb{X}_s$.

Intuition: the subset of the past (whether observed or latent) which forms a sufficient statistic for the future.

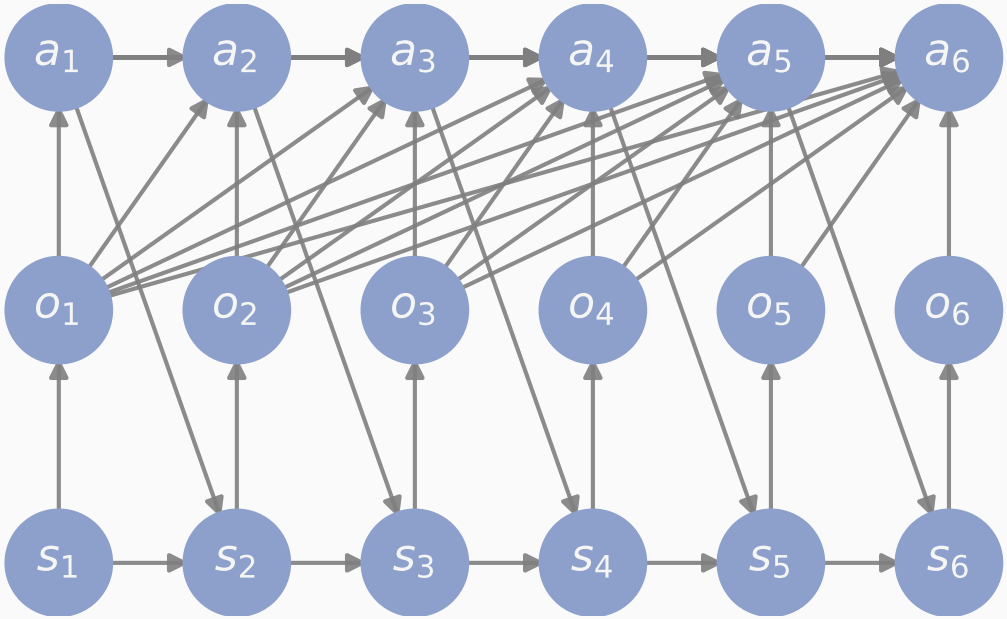
Information-Structural Characterization of Rank of Dynamics

Theorem

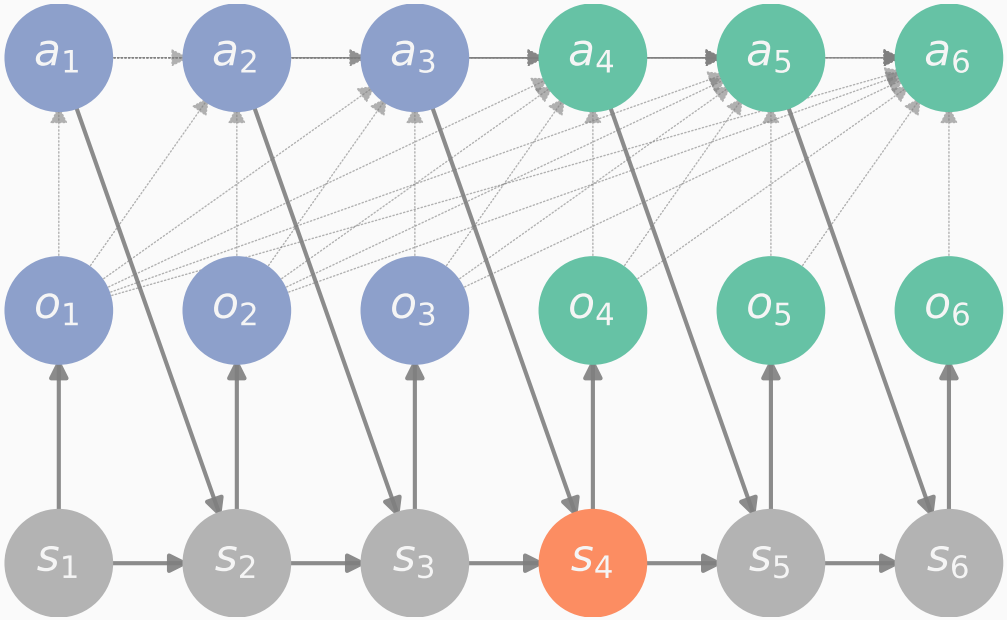
The rank of the observable system dynamics of an arbitrary sequential decision-making problem is bounded by $r_h \leq |\mathbb{I}_h^\dagger|, r \leq \max_{h \in [H]} |\mathbb{I}_h^\dagger|$.

Intuition: Information must "squeeze" through $\mathbb{I}_h^\dagger$ (a bottleneck). Hence, complexity must be bounded by size of $\mathbb{I}_h^\dagger$.

Example: POMDP [1/2]

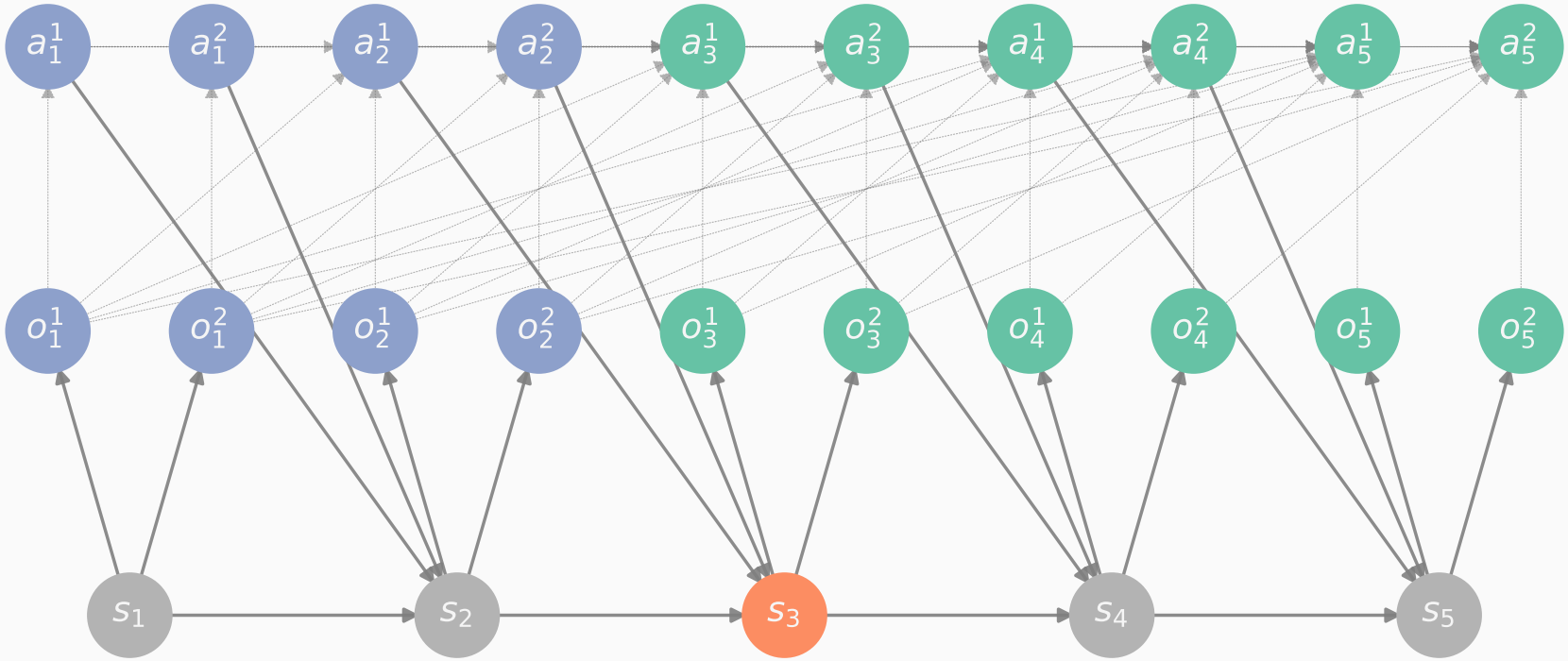


Example: POMDP [2/2]



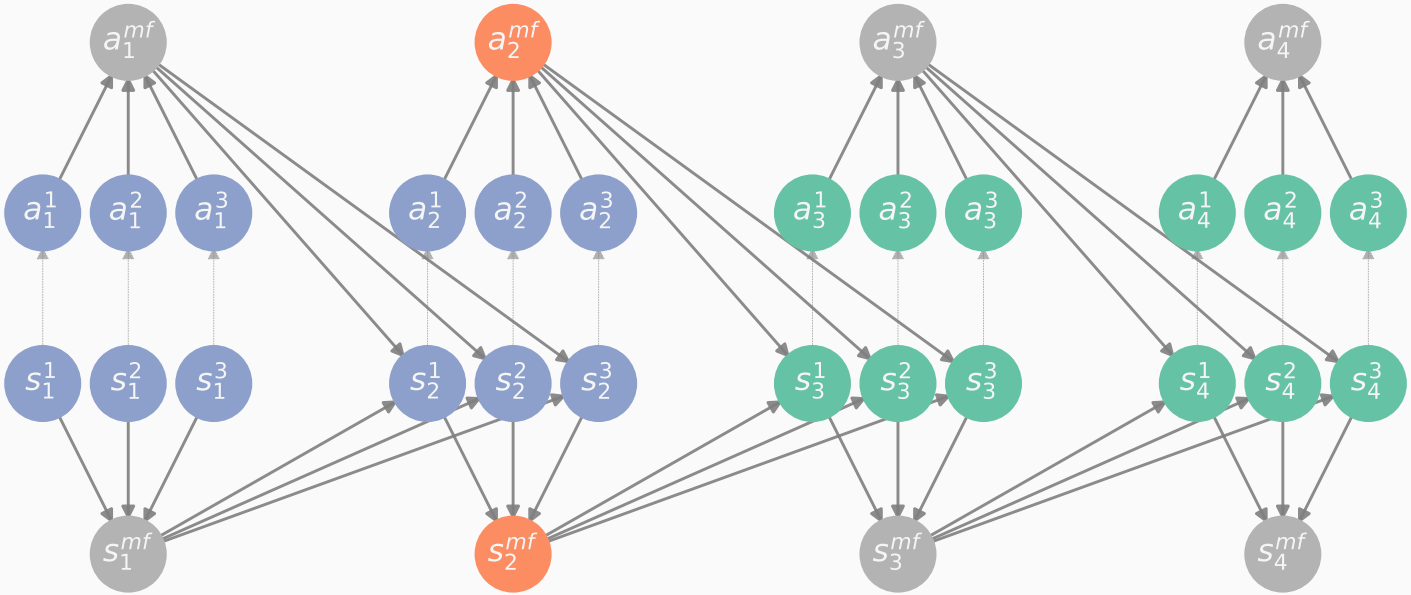
The information structural state coincides with the Markovian state.

Example: Dec-POMDP / POMG



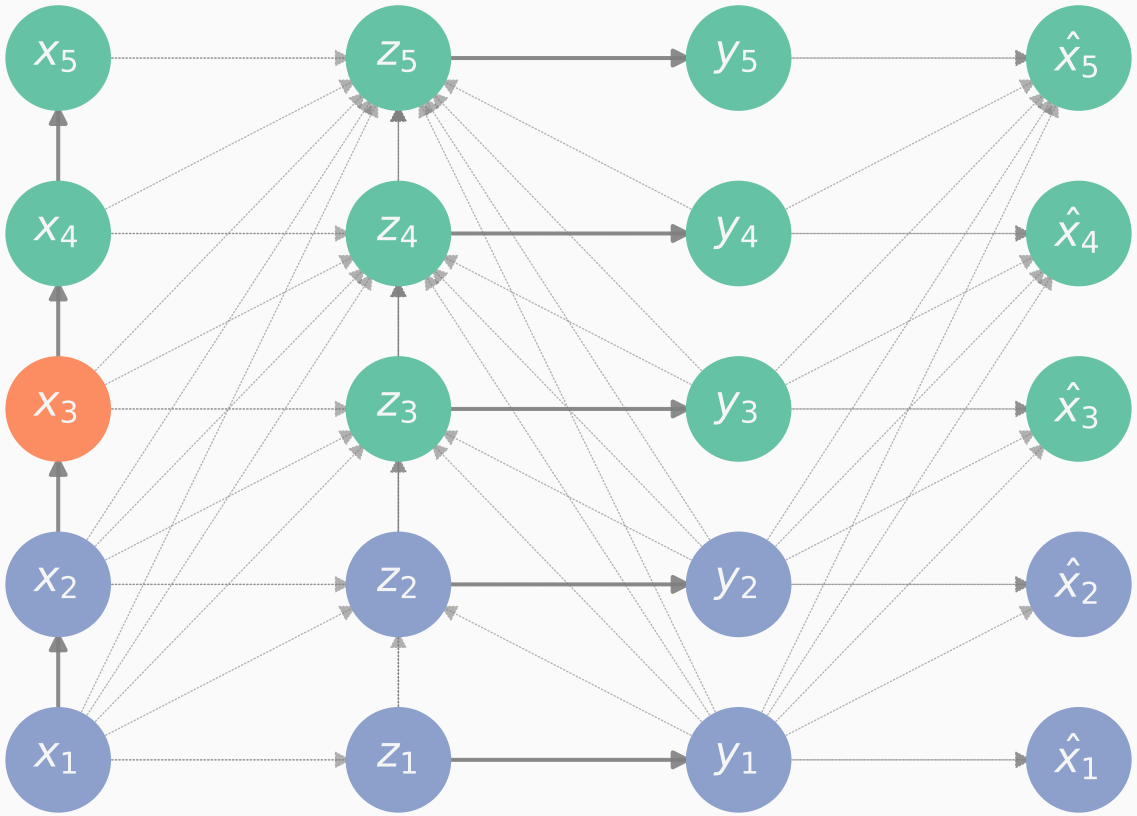
The information structural state coincides with the Markovian state.

Example: "Mean-field"



A mean-field-like information structure. $\mathbb{I}_h^\dagger$ is mean-field state/action.

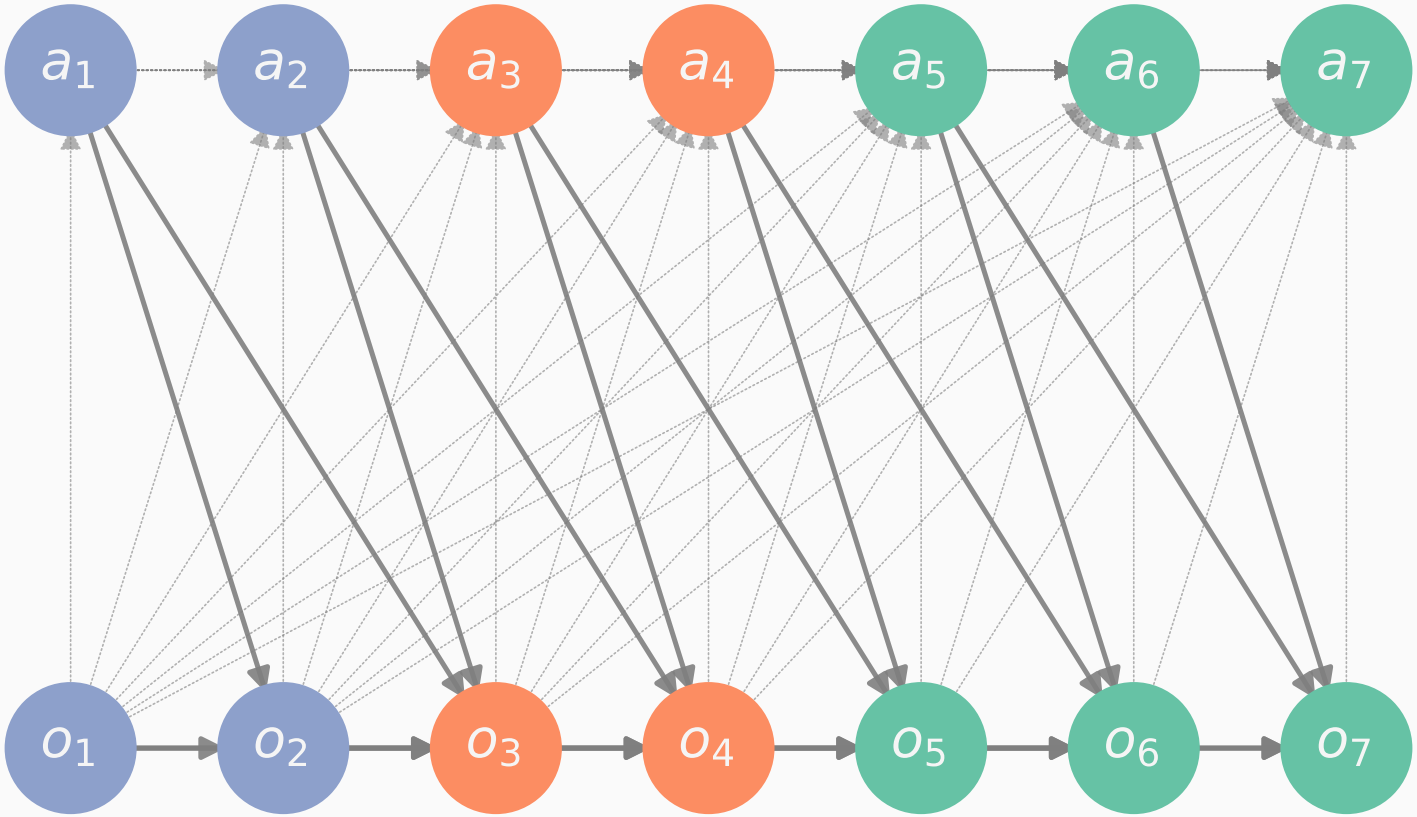
Example: Point-to-Point Real-time Communication



X is source, z is encoding, y is output of noisy channel, $\hat{x}$ is decoding.

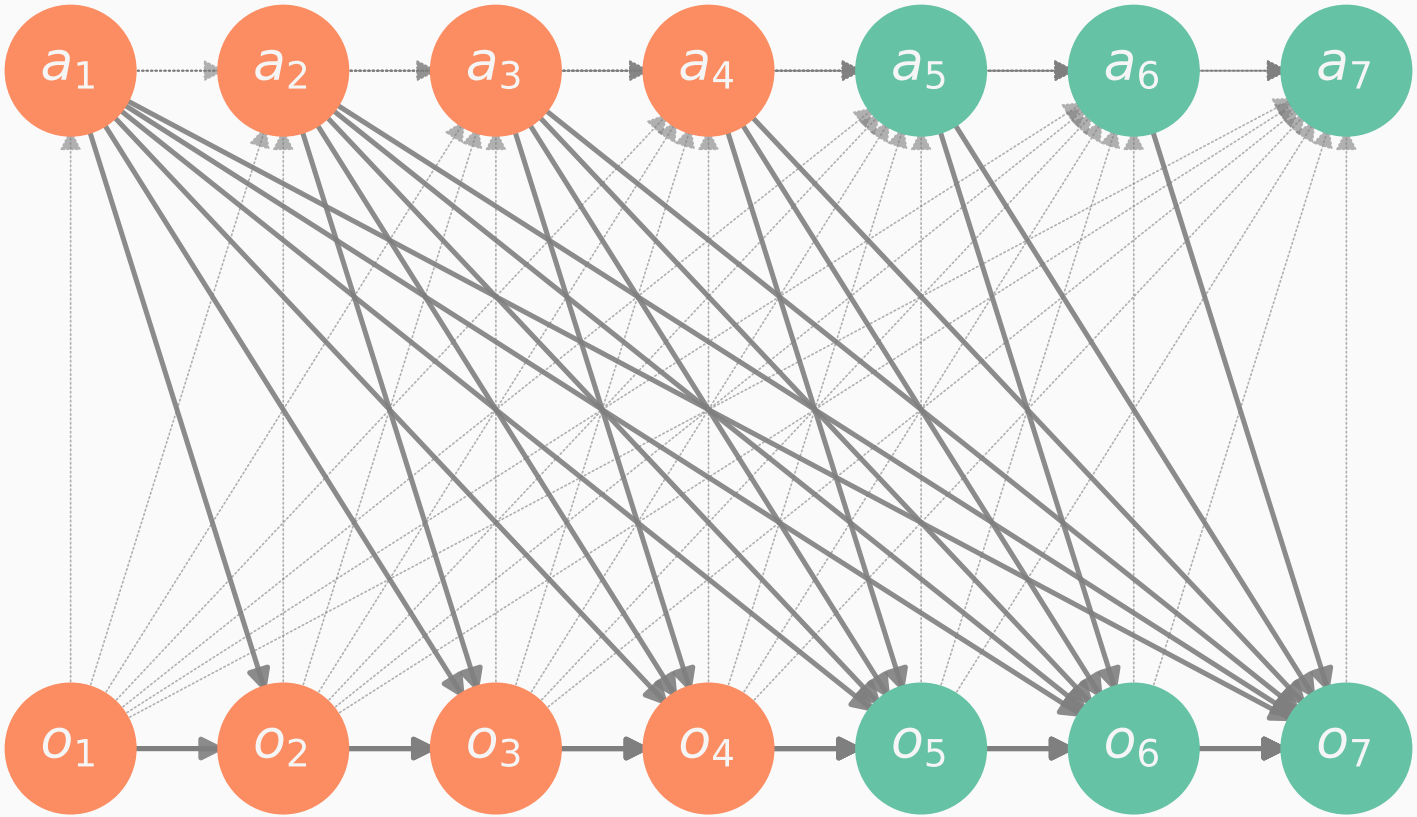
Information-structural state varies at different points in time. Either source, or source + encoding.

Example: m-step Limited Memory



Information-structural state is m-step past.

Example: Fully-Connected Information Structure



Information-structural state is full history—motivating example: intractable.

A general theory towards understanding role of information structure in reinforcement learning

In the worst case (fully-connected information structure), RL is intractable.

So many possible information structures. Not just very simple ones like MDPs/POMDPs.

RL is missing a general theory on the role of information-structure in learning general sequential decision-making problems.

Proof Sketch

By d-separation,

$$
\begin{aligned}
[D_h]_{\text{history},\text{future}} &= \mathbb{P}\left[\text{history}, \text{future} \,|\, \text{actions}\right] \\
&= \mathbb{P}\left[\text{history} \mid \text{actions}\right] \sum_{i_h^\dagger} \mathbb{P}\left[i_h^\dagger \,\middle|\, \text{history}\right] \mathbb{P}\left[\text{future} \,\middle|\, i_h^\dagger\right]
\end{aligned}
$$

Define $D_{h,1} \in \mathbb{R}^{|\text{history}| \times |\mathbb{I}_h^\dagger|}$ and $D_{h,2} \in \mathbb{R}^{|\mathbb{I}_h^\dagger| \times |\text{future}|}$ such that $D_h = D_{h,1} D_{h,2}$.

$\therefore \operatorname{rank}(D_h) \leq \left|\mathbb{I}_h^\dagger\right|$.

A Robust Parameterization Amenable to Reinforcement Learning

Reinforcement Learning Requires Robust Representations

Key challenge in RL: Constructing *representations* which enable robustly and efficiently modeling probabilities of system trajectories. I.e., probabilities of the form $\mathbb{P}\left[\text{future} \mid \text{history}\right]$.

We can use information-structure to do this in a general and systematic way for *arbitrary* sequential decision-making problems.

An identifiability condition

We introduce an "identifiability" condition that enables the construction of such a parameterization.

Assumption ($\mathcal{I}^\dagger$-weakly revealing)

At each point in time $h \in [H]$, the m-step futures are informative about the information structural state. I.e., for two mixtures of info-struct state, the distribution of m-step futures are distinct.

$i_h^\dagger$ is the sufficient statistic of the system at time h. In general, not observable. This assumption basically says, the observations are coupled to $i_h^\dagger$ such that different $i_h^\dagger$ result in different distributions of observations.

(Generalized) Predictive State Representations

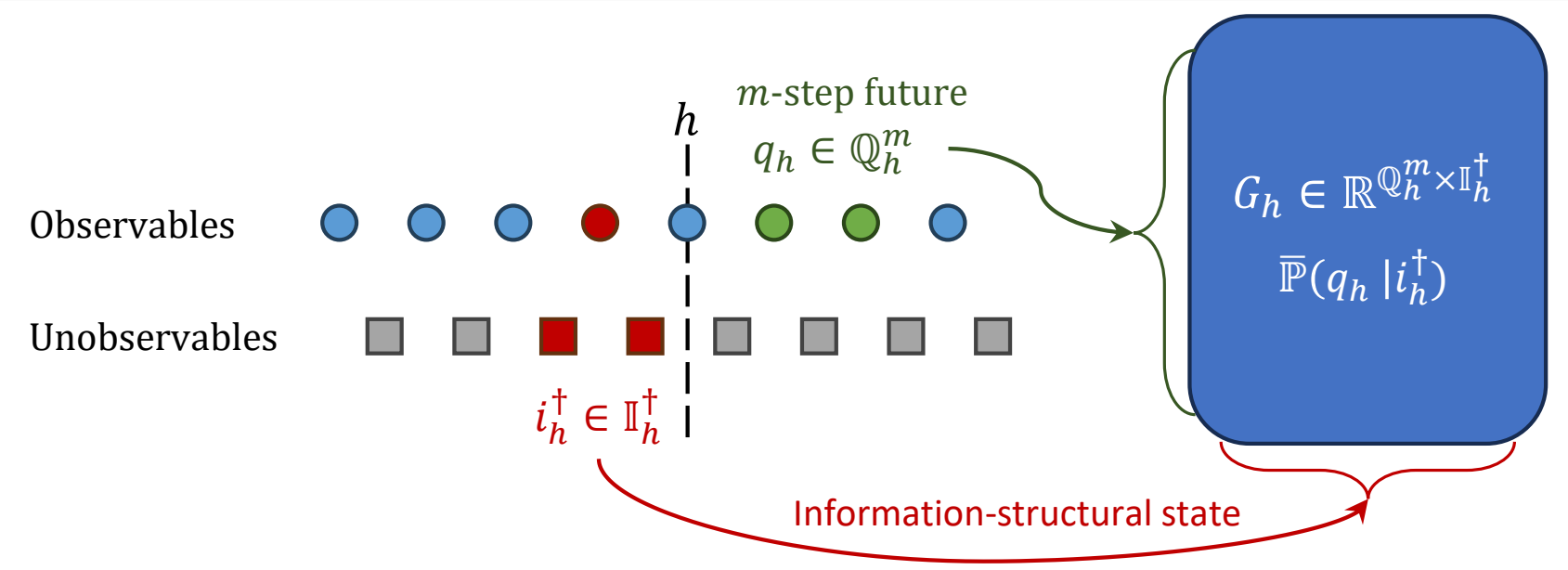
With this, we can construct a "Observable Operator Model" representation of the dynamics of this system, through its information structure.

PSRs are a type of representation where a certain "prediction vector" $\psi_h(\text{history}) \in \mathbb{R}^d$ summarizes all the needed information about the history.

GenPSR Parameterization: $\{M_h : \mathbb{X}_{t(h)} \to \mathbb{R}^{d\times d}\}_{h\in[H]}, \psi_0, \phi_H$ such that

$$\mathbb{P}\left[x_{t(1)}, \ldots, x_{t(H)}\right] = \phi_H(x_{t(h)})^\intercal M_{H-1}(x_{t(H-1)}) \cdots M_1(x_{t(1)})\psi_0$$
$$\psi_h(x_{t(1)}, \ldots, x_{t(h)}) = M_h(x_{t(h)})\psi_{h-1}(x_{t(h-1)}).$$

Constructing a Robust Parameterization



This kind of representation can be constructed by exploiting the information structure.

$$m_h(\text{future}) := (G_h^\dagger)^\intercal \left[\mathbb{P}\left[\text{future} \;\middle|\; i_h^\dagger\right]\right]_{i_h^\dagger},$$

$$\left[M_h(x_{t(h)})\right]_{\text{future},\cdot} = m_{h-1}(x_{t(h)}, \text{future})^\intercal.$$

This Parameterization is Robust

If the identifiability condition is robust (in a minimum singular-value sense), then this parameterization is robust.

I.e., when the error in estimating M_h is small, the error in the estimated probabilities of trajectories is also small.

$$\mathtt{TV}(\mathbb{P}_{\hat{\theta}}, \mathbb{P}_{\theta}) \lesssim \alpha^{-1}\, \mathtt{D}(\hat{\theta}, \theta),$$

where α describes robustness of identifiability condition.

This makes it a good parameterization for reinforcement learning.

Payoff: Characterizing the Sample Complexity of General Reinforcement Learning Problems via Information Structure

Information-structural characterization of statistical hardness

These tools that we developed enable us to characterize an upper bound on the statistical hardness of a general reinforcement learning problem in terms of its information structure.

We have a result that roughly says "any sequential decision-making problem with an information structure $\mathcal{I}$ an be learned with a sample complexity at most $f(\mathcal{I})$"

Information-structural characterization of statistical hardness

Theorem

Suppose a sequential decision-making problem is $\mathcal{I}^\dagger$-weakly revealing. There exists an algorithm which learns an ϵ-optimal policy with a sample complexity

$$\frac{1}{\epsilon^2} \times \mathrm{poly}\left(\frac{1}{\alpha}, \max_h \left|\mathbb{I}_h^\dagger\right|, Q_m, A, H\right)$$

In the game setting, the same assumption imply the existence of a self-play algorithm that learns an ϵ-equilibrium (NE or CCE) with the same sample complexity.

- α: robustness parameter assoc. w/ $\mathcal{I}^\dagger$-weakly revealing condition
- $\mathbb{I}_h^\dagger$: information-structural state
- Q_m: size of m-step trajectories.
- A: max size of actions space.
- H: time horizon.

Proof idea

We prove this result by exhibiting an algorithm achieving this sample complexity.

We crucially make use of the robust generalized PSR parameterization, that we constructed by exploiting information structure.

Online algorithm. Basic outline:

1. Estimate confidence set for θ^* according to genPSR param
2. Choose policy which explores optimally (i.e., visits trajectories for which model gives high uncertainty)

Repeat until confidence set is small enough, then compute optimal policy using estimated model.

Discussion

Where does this leave us?

- Insights: information-structural perspective.
- Fundamental understanding of RL as a problem: what is and isn't possible?
- Practical implications? Information structure as an inductive bias.

Thank you. Questions?

Thank you.

Joint work with: Zhuoran Yang (Faculty @ Yale S&DS)

Paper: `https://arxiv.org/abs/2403.00993`