On the Role of Information Structure in Reinforcement Learning for Partially-Observable Sequential Teams and Games

Awni Altabaa, Zhuoran Yang

Department of Statistics & Data Science, Yale University

Introduction

- In a sequential decision-making problem, the ***information structure*** is the description of how events in the system occurring at different points in time affect each other.
- Classical models of reinforcement learning (e.g., MDPs, POMDPs, Dec-POMDPs/POMGs) assume a very simple and highly regular information structure, while more general models like predictive state representations do not explicitly model the information structure.
- Real-world sequential decision-making problems typically involve a complex and time-varying interdependence of system variables, requiring a rich and flexible representation of information structure.
- The control community has long recognized the importance of information structure, leading to the development of the celebrated Witsenhausen intrinsic model (Witsenhausen, 1975), and extensive study since the 1970s. This includes characterizing the tractability of planning (i.e., computing the optimal policy given a model) as a function of the information structure.
- A general theory of information structures in reinforcement learning is missing.

Key take-aways

- An explicit representation of information structure enables a richer analysis of reinforcement learning problems and more tailor-designed algorithms.
- The information structure of a reinforcement learning problem determines (in part) its statistical tractability. In particular, when learning general sequential decision-making problems with arbitrary information structures, the information structure determines key quantities in the sample complexity.
- We identify a quantity, which we call the ***information-structural state*** due to its role as an "effective state", that we show is central to constructing compact representations that enable efficient reinforcement learning. This quantity is derived in terms of the DAG representation of the information structure.

Generic Sequential Decision-Making Problems

Consider a controlled stochastic process $(X_1, \dots, X_H)$, where X_h is a random variable corresponding to the variable at time h. X_h may be either an 'observation' or an 'action'. A choice of policy $\pi = \{\pi_h\}_{h\in\mathcal{A}}$ induces a probability distribution on $\mathbb{X}_1 \times \cdots \times \mathbb{X}_H$ as follows

$$\mathbb{P}^{\pi}(x_1, \dots, x_H) = \prod_{h\in\mathcal{O}} \mathbb{P}_h(x_h \mid x_1, \dots, x_{h-1}) \cdot \prod_{h\in\mathcal{A}} \pi_h(x_h \mid x_1, \dots, x_{h-1}), \tag{1}$$

We define the *system dynamics matrix* $\boldsymbol{D}_h \in \mathbb{R}^{|\mathbb{H}_h|\times|\mathbb{F}_h|}$ as the matrix giving the probability of each possible pair of history and future at time h given the execution of the actions,

$$[\boldsymbol{D}_h]_{\tau_h,\omega_h} = \overline{\mathbb{P}}[\tau_h, \omega_h] = \mathbb{P}\left[\tau_h^o, \omega_h^o \;\middle|\; \mathrm{do}(\tau_h^a, \omega_h^a)\right], \quad \tau_h \in \mathbb{H}_h, \omega_h \in \mathbb{F}_h, \tag{2}$$

where $\omega_h^o = \texttt{obs}(\omega_h)$ are is the observation component of the future ω_h, $\omega_h^a = \texttt{act}(\omega_h)$ is the action component, and similarly for τ_h^o, τ_h^a.

The rank of such a controlled stochastic process is the maximal rank of its dynamics matrices. This is a measure of the complexity of the dynamics. **Definition (Rank of dynamics).** The rank of the dynamics $\{\boldsymbol{D}_h\}_{h\in[H]}$ is $r = \max_{h\in[H]} \mathrm{rank}(\boldsymbol{D}_h)$.

Explicitly representing information structures in RL: Partially-Observable Sequential Teams (and Games)

A partially-observable sequential team (POST) is a controlled stochastic process that specifies the joint distribution of T variables $(X_t)_{t\in[T]}$, and is specified by the following components.

1. **Variable Structures.** The variables $\{X_t\}_{t\in[T]}$ are partitioned into two disjoint subsets — $\mathcal{S} \subset [T]$ indexes system variables and $\mathcal{A} \subset [T]$ indexes action variables.
2. **Information Structure.** For $t \in [T]$, the "information set" $\mathcal{I}_t \subset [t-1]$ of the variable X_t is the set of past variables that are coupled to X_t in the dynamics. That is, the value of $I_t := (X_s : s \in \mathcal{I}_t)$ directly determines the distribution of X_t. We call I_t the "information variable" at time t, and call $\mathbb{I}_t = \prod_{s\in\mathcal{I}_t} \mathbb{X}_s$ the "information space".
3. **System Kernels.** For any $t \in \mathcal{S}$, $\mathcal{T}_t$ is a mapping from $\mathbb{I}_t$ to $\mathcal{P}(\mathbb{X}_t)$ that specifies the conditional distribution of a system variable X_t given I_t.
4. **Decision Kernels.** Each agent chooses a decision kernel (i.e., policy) $\pi_t : \mathbb{I}_t \to \mathcal{P}(\mathbb{X}_t)$, specifying the distribution over actions at time $t \in \mathcal{A}$.
5. **Observability.** We denote the observable system variables by $\mathcal{O} \subset \mathcal{S}$. We require that the information sets of the action variables are observable, $\mathcal{O} \supset \cup_{t\in\mathcal{A}}(\mathcal{I}_t \cap \mathcal{S})$. We define $\mathcal{U} := \mathcal{O} \cup \mathcal{A}$, and let $H := |\mathcal{U}|$ be the time-horizon of the *observable* variables.
6. **Reward Function.** At the end of an episode, the team receives the reward $R(x_s,\ s \in \mathcal{U})$, where $R : \prod_{s\in\mathcal{U}} \mathbb{X}_s \to [0, 1]$ is the "reward function."

Any choice of decision kernels (joint policy) $\boldsymbol{\pi}$ induces a unique probability measure over $\mathbb{X}_1 \times \cdots \times \mathbb{X}_T$, which is given by

$$\mathbb{P}^{\boldsymbol{\pi}}[X_1 = x_1, \dots X_T = x_t] = \prod_{t\in\mathcal{S}} \mathcal{T}_t(x_t \mid \{x_s : s \in \mathcal{I}_t\}) \prod_{t\in\mathcal{A}} \pi_t(x_t \mid \{x_s : s \in \mathcal{I}_t\}). \tag{3}$$

From the perspective of a learning agent, we are interested in modeling the observable variables, which we can index by $h \in [H]$ as follows,

$$\left(X_{t(h)}\right)_{h\in[H]} = \left(X_{t(1)}, \dots, X_{t(H)}\right) = (X_t)_{t\in\mathcal{U}}. \tag{4}$$

Characterizing the rank of dynamics via information structural state

The information structure of a POST can be naturally represented as a (labeled) directed acyclic graph (DAG). Given the variable structure and information structure of a POST, $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \{\mathcal{I}_t\}_t)$, its DAG representation is given by $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{L})$. The nodes of the graph are the set of variables, $\mathcal{V} = [T] = \mathcal{S} \cup \mathcal{A}$. The edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ of the DAG are given by

$$\mathcal{E} = \{(i, t) : t \in [T], i \in \mathcal{I}_t\}.$$

Definition (Information-structural state). 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-separates 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 $\mathbb{I}_h^\dagger := \prod_{s\in\mathcal{I}_h^\dagger} \mathbb{X}_s$ as the joint space of those variables.

Theorem (complexity of dynamics of POST/POSG). The rank of the observable system dynamics of a POST or POSG is bounded by

$$r \le \max_{h\in[H]} \left|\mathbb{I}_h^\dagger\right|.$$

Signifcance & Implications: Generalized Predictive State Representations and Sample-Efficent Reinforcement Learning

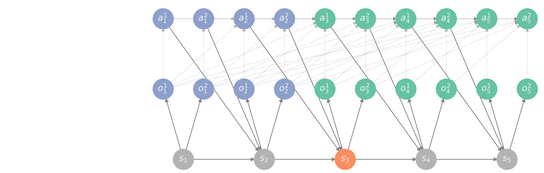
Theorem (Generalized PSR) A POST/POSG (under certain conditions) can be represented compactly by a set of operators $M_h : \mathbb{X}_h \to \mathbb{R}^{d\times d}, h \in \{1, \dots, H-1\}$ which capture the probability of any trajectory,

$$\overline{\mathbb{P}}[x_{t(1)}, \dots, x_{t(H)}] = \phi_H(x_{t(H)})^\top M_{H-1}(x_{t(H-1)}) \cdots M_1(x_{t(1)})\psi_0. \tag{5}$$

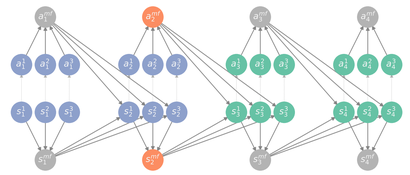
Theorem (Sample complexity of RL) There exists a reinforcement learning algorithm which learns an ϵ-optimal policy/an ϵ-equilibrium for POSTs/POSGs (under certain conditions) with a polynomial sample complexity depending on the information-structural state,

$$\frac{1}{\epsilon^2} \times \mathrm{poly}\left(\frac{1}{\alpha}, \max_h \left|\mathbb{I}_h^\dagger\right|, \max_h |\mathbb{Q}_h^m|, \max_{s\in\mathcal{U}} |\mathbb{X}_s|, Q_A, H\right) \tag{6}$$

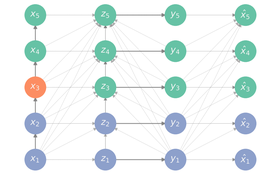
Examples of Information Structures & their Information-Structral State



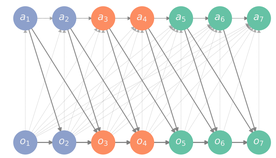
(a) Decentralized POMDP/POMG information-structure.



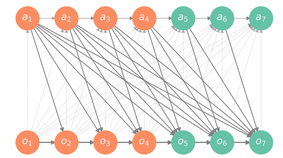
(b) "Mean-field" information structure.



(c) Point-to-point real-time communication with feedback information structure.



(d) Limited-memory ($m = 2$) information structures.



(e) Fully connected information structure.

Figure 1. DAG representation of various information structures. Grey nodes represent unobservable variables, blue nodes represent past observable variables, green nodes represent future observable variables, and red nodes represent the information structural state $\mathcal{I}_h^\dagger$.