Provably Efficient Offline Reinforcement Learning for Partially Observable Markov Decision Processes

Hongyi Guo[1], Qi Cai[1], Yufeng Zhang[1], Zhuoran Yang[2], Zhaoran Wang[1]

[1]Northwestern University

[2]Yale University

July 21, 2022

Background

Offline reinforcement learning (RL) for partially observable Markov decision processes (POMDPs).

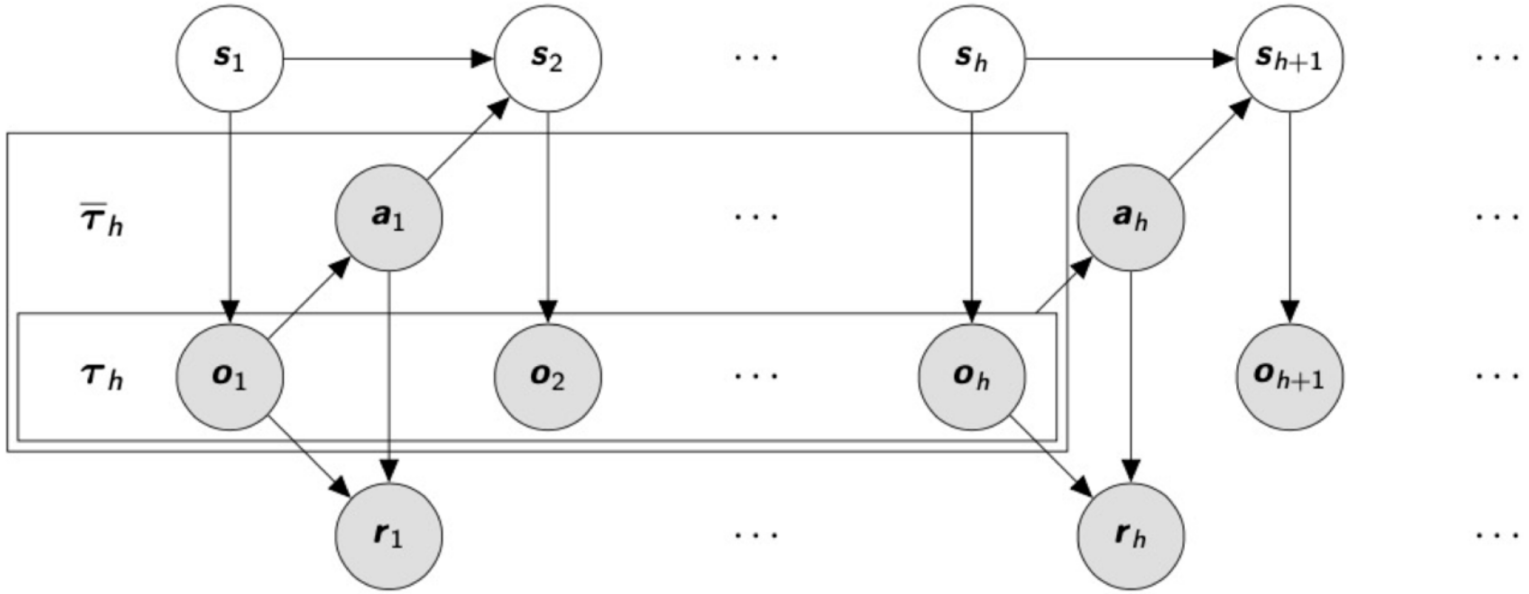

Figure: POMDP

Overview

- Linear POMDP (emission kernel and transition kernel are linear in known feature mappings)
- Undercompleteness assumption (observation dist. $\Rightarrow$ state dist.)

Overview

- Linear POMDP (emission kernel and transition kernel are linear in known feature mappings)
- Undercompleteness assumption (observation dist. $\Rightarrow$ state dist.)
- A pessimistic offine RL algorithm
- Finite sample guarantee $\widetilde{\mathcal{O}}(1/\epsilon^2)$

Intuition

Define random functions

$$X_{h,a}(o) = \mathbb{1}^{\bar{\pi}}_{\mathrm{do}(A_{h-1}=a)}\{O_h = o\},$$
$$Y_{h,a,a'}(o,o') = \mathbb{1}^{\bar{\pi}}_{\mathrm{do}(A_{h-1:h}=(a,a'))}\{O_{h:h+1} = (o,o')\}.$$

Intuition

Define random functions

$$X_{h,a}(o) = \mathbb{1}^{\bar{\pi}}_{\mathrm{do}(A_{h-1}=a)}\{O_h = o\},$$
$$Y_{h,a,a'}(o,o') = \mathbb{1}^{\bar{\pi}}_{\mathrm{do}(A_{h-1:h}=(a,a'))}\{O_{h:h+1} = (o,o')\}.$$

Regress Y on X and in the mean time we can solve the model parameter θ:

$$Y_{h,a,a'} = \mathbb{F}^{\theta}_{h,a'} X_{h,a} + U_{h,a,a'}, \tag{2.1}$$

where $\mathbb{F}$ is some linear operator and $U_{h,a,a'}$ is the zero-mean perturbation term.

Intuition

Define random functions

$$X_{h,a}(o) = \mathbb{1}^{\bar{\pi}}_{\mathrm{do}(A_{h-1}=a)}\{O_h = o\},$$
$$Y_{h,a,a'}(o,o') = \mathbb{1}^{\bar{\pi}}_{\mathrm{do}(A_{h-1:h}=(a,a'))}\{O_{h:h+1} = (o,o')\}.$$

Regress Y on X and in the mean time we can solve the model parameter θ:

$$Y_{h,a,a'} = \mathbb{F}^{\theta}_{h,a'} X_{h,a} + U_{h,a,a'}, \tag{2.1}$$

where $\mathbb{F}$ is some linear operator and $U_{h,a,a'}$ is the zero-mean perturbation term.

- U is correlated with X! $\Rightarrow$ IV regression.

IV Regression

An instrumental variable (IV) is correlated with X but uncorrelated with the perturbation U. We choose $Z_h = O_{h-1}$ as the IV.

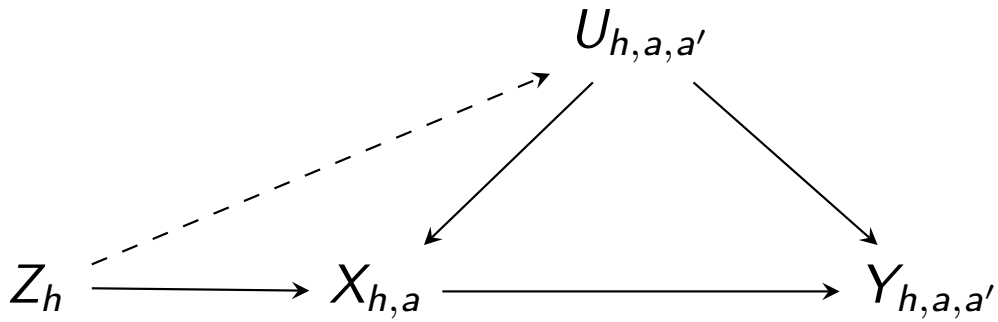

Figure: The relationship between $X_{h,a}$, $Y_{h,a,a'}$, $U_{h,a,a'}$, and Z_h. The arrows indicates the dependency between those variables. The dashed arrow indicates two uncorrelated variables. In this figure, $U_{h,a,a'}$ affects both $X_{h,a}$ and $Y_{h,a,a'}$ directly, Z_h only affects $X_{h,a}$ directly, and Z_h and $U_{h,a,a'}$ are uncorrelated.

Algorithm Overview

With the help of the IV, we have

$$\mathbb{E}[Y_{h,a,a'} \mid Z_h] = \mathbb{F}^{\theta}_{h,a'} \mathbb{E}[X_{h,a} \mid Z_h]. \tag{2.2}$$

Our algorithm has the follows steps:

1. Construct estimators of $\mathbb{E}[Y_{h,a,a'} \mid Z_h]$ and $\mathbb{E}[X_{h,a} \mid Z_h]$ from the dataset.
2. Construct confidence region $\widehat{\Theta}$ of the model parameter so that $\mathbb{E}[Y_{h,a,a'} \mid Z_h]$ and $\mathbb{F}^{\theta}_{h,a'} \mathbb{E}[X_{h,a} \mid Z_h]$ are close enough for $\theta \in \widehat{\Theta}$.
3. Pessimism Planning

$$(\widehat{\pi}, \widehat{\theta}) = \operatorname*{argmax}_{\pi \in \Pi} \operatorname*{argmin}_{\theta \in \widehat{\Theta}} \mathcal{J}(\theta, \pi).$$

Thank you!

Please check our poster @Hall E