Exploration in linear MDPs

A linear MDP $M=(S,A,P,R,H,d_0)$ satisfies: $P(s'|s,a) = \underbrace{\phi(s,a)}_{\text{known}}{}^T \underbrace{\psi(s')}_{\text{unknown}}$, $\phi(s,a) \in \mathbb{R}^d$, $\psi(s') \in \mathbb{R}^d$, $R(s,a) = \phi(s,a)^T \theta_R$.

$[P]_{SA\times S} = [\Phi]_{SA\times d} \cdot [\Psi]_{d\times S}$

Function class: $\mathcal{F} = \{(s,a) \mapsto \phi(s,a)^T\theta : \theta \in \mathbb{R}^d\}$.

Setting yields (a variant of) low Bellman rank, thus statistically solvable by OLIVE. (see [AJKS]), but OLIVE is computationally intractable.

Key observation: $\forall f: S\times A \to \mathbb{R}$, $\mathcal{T}f \in \mathcal{F}$.

$$R(s,a) + \langle P(s,a), V_f \rangle = \phi(s,a)^T \theta_R + \langle \phi^T(s,a)\psi, V_f \rangle$$

"Explore-or-terminate". Alg computes f. explore w/ π_f.

either π_f is near-optimal. or. it explores (s,a) where $\phi(s,a)^T \Lambda_h^{-1} \phi(s,a) = +\infty$.

Alg: $\forall h$, define $\Lambda_h = \mathbb{E}_{D_h}[\phi\phi^T]$.

For $h = H, H-1, H-2, \dots, 1$.

$$\tilde{f}_h = \arg\min_{f\in\mathcal{F}} \mathbb{E}_{D_h}\left[\left(f(s,a) - r - V_{f_{h+1}}(s')\right)^2\right].$$

FQI.

$$\mathbb{E}_{(s_h,a_h)\sim \text{data}}\left[\phi(s_h,a_h)\phi(s_h,a_h)^T\right].$$

$d\times 1$, $1\times d$

feature cov. matrix.

$$f_h(s,a) = \begin{cases} \tilde{f}_h(s,a), & \text{if } \phi(s,a)^T\Lambda_h^{-1}\phi(s,a) < \boxed{+\infty} \\ H & \text{o.w.} \end{cases}$$

V_{max}. "known".

$f = f_1 \circ f_2 \circ f_3 \circ \dots \circ f_H$. explore w/ $\pi_f = \pi_{f_1} \circ \pi_{f_2} \circ \dots \circ \pi_{f_H}$.

Example: data only contains a particular s,a. $\Lambda = \phi(s,a)\phi(s,a)^T$ ← rank 1.

$\phi(s,a)^T \Lambda_h^{-1} \phi(s,a)$

$$\lim_{\lambda \to 0} \phi(s,a)^T \left(\phi(s,a)\phi(s,a)^T + \lambda I\right)^{-1} \phi(s,a) \approx 1.$$

Lemma: $f_h(s,a) = \begin{cases} \mathbb{E}[r + V_{\tilde{f}_{h+1}}(s') \mid s,a] & \leftarrow \text{when } \phi^T(s,a)\Lambda_h^{-1}\phi(s,a) < \epsilon_2 \\ H, & \text{o.w.} \end{cases}$

Comment.

Lemma: (Optimism) $f \geq Q^*$.

At H: $f_H(s,a) = \begin{cases} R(s,a) = Q^*_H(s,a), & \text{if "known"} \\ H \geq Q^*_H(s,a) \end{cases}$

For $h < H$, by induction we have $f_{h+1} \geq Q^* \Rightarrow V_{f_{h+1}} \geq V_{Q^*} = V^*$

$$f_h(s,a) = \begin{cases} \mathbb{E}[r + V_{f_{h+1}}(s') \mid s,a] \geq \mathbb{E}[r + V^*_{h+1}(s') \mid s,a] = Q^*_h(s,a) \\ H \geq Q^*_h(s,a) \ \checkmark \end{cases}$$

Proof of "exploit-or-terminate": if π_f is suboptimal.

$$\varepsilon < J(\pi^*) - J(\pi_f) \leq \mathbb{E}_{s\sim d_0}[\max_a Q^*(s,a)] - J(\pi_f).$$

$$\leq \mathbb{E}_{s\sim d_0}[\max_a f_1(s,a)] - J(\pi_f) \quad \text{(Optimism)}$$

$$= \sum_{h=1}^{H} \mathbb{E}_{\pi_f}[f(s_h,a_h) - r_h - V_f(s_{h+1})].$$

$$\leq \sum_{h=1}^{H} \left|\mathbb{E}_{\pi_f}[f(s_h,a_h) - r_h - V_{\tilde{f}}(s_{h+1})]\right|$$

$f \geq \tilde{f}$.

$$\exists h: \left|\mathbb{E}_{\pi_f}[f(s_h,a_h) - r_h - V_{\tilde{f}}(s_{h+1})]\right| \geq \frac{\varepsilon}{H}.$$

$$\frac{\varepsilon}{H} \le |\cdots| = \left|\mathbb{E}_{\pi_f}\left[f(s_h,a_h) - \mathbb{E}\left[r_h + V_f^{\sim}(s_{h+}) \mid s_h, a_h\right]\right]\right|$$

$$\le \sqrt{\mathbb{E}_{\pi_f}\left[\left(f(s_h,a_h) - (\qquad)\right)^2\right]}$$

$$\le \sqrt{\mathbb{E}_{(s,a)\sim d_h^{\pi_f}}\left[\left(f(s,a) - \mathbb{E}\left[r + V_f^{\sim}(s') \mid s,a\right]\right)^2\right]}$$

$$\sum_{(s,a):\ \text{known}} d_h^{\pi_f}(s,a)\left[\ -\ -\ -\ -\ -\ \right] = 0$$

$$+ \sum_{(s,a):\ \text{unknown}} d_h^{\pi_f}(s,a)\left[\ -\ -\ -\ -\ -\ -\ \right] \to \text{nonzero}$$

$$\le \sum_{(s,a):\ \text{unknown}} d_h^{\pi_f}(s,a) \cdot C = C \cdot \Pr[\text{traj visits unknown } (s,a) \text{ at timestep } h]$$

const. func of V_{max}:

$$\Pr[\text{traj. visits unknown } (s,a) \text{ at step } h] \ge \frac{\varepsilon}{H \cdot C} > 0.$$

unknown: $\phi(s,a)\,\Lambda_h^{-1}\,\phi(s,a) = +\infty$.

Open Problem: Given MDP M, linear function class $\mathcal{F}$. (dim d). s.t. $\forall f \in \mathcal{F}$, $\mathcal{T}f \in \mathcal{F}$.

$\Rightarrow$ (variant of) OLIVE explores sample-efficiently.

Clarification on Linear MDP lectures

Hi all,

I have left a few hand-waving arguments in today's lecture on linear MDPs, and would like to tighten things up a bit.

First of all, as we mentioned, the feature covariance matrix $\Lambda := \mathbb{E}[\phi\phi^\top]$ may be rank-deficient before we have exploratory data, so technically speaking we cannot invert it. As I alluded to in the lecture, one way to define inversion for the purpose of this simplified analysis is: for any real vector $x \in \mathbb{R}^d$ (think of it as $\phi(s,a)$ for some (s,a)) $x^\top \Lambda^{-1} x := \lim_{\lambda \to 0} x^\top (\Lambda + \lambda I)^{-1} x$.

Then, I gave an example where data only comes from 1 single (s,a) pair to show that the above expression can be a good way to define "known" versus "unknown", that is, when $\Lambda = xx^\top$, we have $x\Lambda^{-1}x^\top = 1$. I didn't come up with a proof in class but it turns out to be not very hard: for quadratic forms we can always rotate the orthonormal basis, so wlog we can rotate it such that $x = [1, 0, \ldots, 0]^\top$ (and $\|x\|$ clearly does not matter so I am assuming it has unit norm). Then the expression becomes
$\lim_{\lambda \to 0} [1, 0, \ldots, 0]([1, 0, \ldots, 0]^\top [1, 0, \ldots, 0] + \lambda I)^{-1} [1, 0, \ldots, 0]^\top = \lim_{\lambda \to 0} 1/(1+\lambda) = 1.$
In other words, within the subspace that Λ spans, $x^\top \Lambda^{-1} x$ is always finite. For x that goes out of the spanned subspace, you will get infinity, e.g., let $y = [0, 1, 0, \ldots, 0]$, and $y^\top (xx^\top)^{-1} y = \lim_{\lambda \to 0} 1/\lambda = \infty$. Remember that we are getting infinities here because we are doing a simplified analysis, where in every iteration we can collect infinite amount of data using the current exploration policy (similar to the simplified OLIVE analysis). In the full analysis you will never run into these infinities, as λ will be set to a finite value (instead of taking the limit of $\lambda \to 0$), and the threshold for determining "known" vs "unknown" will also be finitely large (like $O(1/\lambda)$) instead of ∞.

Finally, this lecture may be easier to understand if you compare it with the Rmax analysis; they are very similar, and linear MDP strictly generalizes tabular MDPs. It will be also helpful if you have seen rank-deficient linear regression before (as in e.g., online learning). For that I recommend the section of Lihong Li's thesis on KWIK linear regression (https://rucore.libraries.rutgers.edu/rutgers-lib/26345/pdf/1/, Section 5.2.3), which is relatively easy to read if you just want to get the high-level idea. Retrospectively it would have been better if I spend half of a lecture on related material to give more background on this. (This is new material I added to the course for the first time in this semester, so it's a little experimental.)

Best,
Nan

lecture

Updated Just now by Nan Jiang

followup discussions *for lingering questions and comments*