Lecture 9
Planning with global access and uniform realizability: Part II

Global access

Can get all the features at all states, can preprocess it

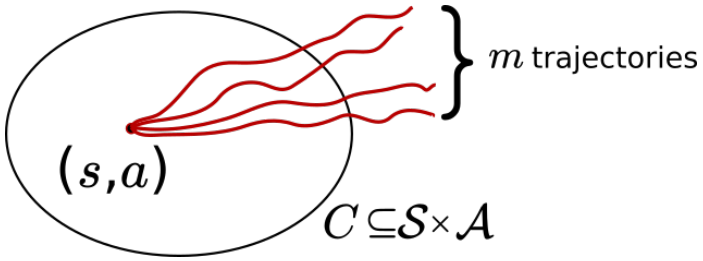
ζ −uniform action-value realizability

$$\sup_{\pi \in \mathrm{DETML}} \inf_{\theta} \|q^{\pi} - \Phi\theta\|_{\infty} \leq \zeta$$

least-squares policy evaluation

Least-squares Policy Evaluation

1. Rollouts from a set $\mathcal{C}$ of well-chosen state-action pairs
2. Average over those
3. Least-squares fit



$$\left(\mathbb{E}\left[\hat{R}_m(z)\right] = q^{\pi}(z)\right)$$

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{z \in \mathcal{C}} \varrho(z) \Big(\langle \theta, \varphi(z) \rangle - \hat{R}_m(z) \Big)^2$$

$$\hat{R}_m(z) = \frac{1}{m} \sum_{j=1}^{m} \sum_{t=0}^{H^{(j)}-1} r_{A_t^{(j)}}(S_t^{(j)})$$

$$\mathbb{P}(H^{(j)} \geq t+1) = \gamma^t$$

First head with $X_1, X_2, \ldots$:
$\mathbb{P}(X_i = \text{Head}) = 1 - \gamma$

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{z \in \mathcal{C}} \varrho(z) \Big(\langle \theta, \varphi(z) \rangle - \hat{R}_m(z) \Big)^2$$

Lemma (LSPE-G extrapolation error control): Fix any full-rank feature-map $\varphi : \mathcal{Z} \to \mathbb{R}^d$ and take the set $\mathcal{C} \subset \mathcal{Z}$ and the weighting function $\varrho \in \Delta_1(\mathcal{C})$ as in the Kiefer-Wolfowitz theorem. Fix an arbitrary policy π and let θ and ε_π such that $q^\pi = \Phi\theta + \varepsilon_\pi$ and assume that immediate rewards belong to the interval $[0, 1]$. Let $\hat{\theta}$ be as in Eq. (6). Then, for any $0 \le \delta \le 1$, with probability $1 - \delta$,

$$\left\| q^\pi - \Phi\hat{\theta} \right\|_\infty \le \|\varepsilon_\pi\|_\infty (1 + \sqrt{d}) + \sqrt{d} \left(\frac{\gamma^H}{1-\gamma} + \frac{1}{1-\gamma} \sqrt{\frac{\log(2|C|/\delta)}{2m}} \right). \tag{7}$$

KW: $|\mathcal{C}| \le d(d+1)/2$, hence to make second term $\le 2\varepsilon$, enough if

$$H \ge H_{\gamma, \varepsilon/\sqrt{d}} \qquad \text{and} \qquad m \ge \frac{d}{(1-\gamma)^2 \varepsilon^2} \log \frac{d(d+1)}{\delta}$$

Total # samples: $|\mathcal{C}|Hm \approx \dfrac{d^3 \log(d/\varepsilon) \log(d/\delta)}{(1-\gamma)^3 \varepsilon^2}$

Note: Can use $|\mathcal{C}| = \tilde{O}(d)$, paying factor of two blow-up factor: $d^3 \Rightarrow d^2$

Progress Lemma with Approximation Errors

Lemma (Geometric progress lemma with approximate policy improvement): Consider a memoryless policy π and its corresponding value function v^π. Let π' be any policy and define $\varepsilon : \mathcal{S} \to \mathbb{R}$ via

$$Tv^\pi = T_{\pi'} v^\pi + \varepsilon\,.$$

Then,

$$\|v^* - v^{\pi'}\|_\infty \leq \gamma \|v^* - v^\pi\|_\infty + \frac{1}{1-\gamma}\,\|\varepsilon\|_\infty.$$

Approximate Policy Iteration

$$Tv^{\pi_k} = T_{\pi_{k+1}} v^{\pi_k} + \varepsilon_k$$

Theorem (Approximate Policy Iteration): Let $(\pi_k)_{k\geq 0}$, $(\varepsilon_k)_k$ be such that (11) holds for all $k \geq 0$. Then, for any $k \geq 1$,

$$\|v^* - v^{\pi_k}\|_\infty \leq \frac{\gamma^k}{1-\gamma} + \frac{1}{(1-\gamma)^2} \max_{0\leq s\leq k-1} \|\varepsilon_s\|_\infty . \tag{12}$$

Approximate Policy Iteration 2

$$q_k = q^{\pi_k} + \varepsilon'_k, \qquad M_{\pi_k} q_k = M q_k\,, \quad k = 0, 1, \ldots\,. \tag{13}$$

Corollary (Approximate Policy Iteration with Approximate Action-value Functions): The sequence defined in (13) is such that

$$\|v^* - v^{\pi_k}\|_\infty \leq \frac{\gamma^k}{1-\gamma} + \frac{2}{(1-\gamma)^2} \max_{0 \leq s \leq k-1} \|\varepsilon'_s\|_\infty\,.$$

Least-squares Policy Iteration with G-optimal design (MC-LSPI)

Given the feature map φ, find $\mathcal{C}$ and ρ as in the Kiefer–Wolfowitz theorem

Let $\theta_{-1} = 0$

For $k = 0, 1, 2, \ldots, K-1$ do

    Roll out with policy $\pi := \pi_k$ for H steps to get the targets $\hat{R}_m(z)$ where $z \in \mathcal{C}$ and $\pi_k(s) = \arg\max_a \langle \theta_{k-1}, \varphi(s,a) \rangle$

    Solve the weighted least-squares problem given by Eq. (4) to get θ_k.

Return θ_{K-1}

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{z \in \mathcal{C}} \varrho(z) \Big(\langle \theta, \varphi(z) \rangle - \hat{R}_m(z) \Big)^2 \qquad (4)$$

$$\hat{R}_m(z) = \frac{1}{m} \sum_{j=1}^{m} \sum_{t=0}^{H-1} \gamma^t r_{A_t^{(j)}}(S_t^{(j)})$$

Theorem (LSPI performance): Fix an arbitrary full rank feature-map $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ and let $K, m, H \geq 1$. Assume that B2$_\varepsilon$ holds. Then, for any $0 \leq \zeta \leq 1$, with probability at least $1 - \zeta$, the policy π_K which is greedy with respect to $\Phi\theta_{K-1}$ is δ-suboptimal with

$$\delta \leq \underbrace{\frac{2(1+\sqrt{d})}{(1-\gamma)^2}\varepsilon}_{\text{approx. error}} + \underbrace{\frac{\gamma^{K-1}}{1-\gamma}}_{\text{iter. error}} + \underbrace{\frac{2\sqrt{d}}{(1-\gamma)^2}\left(\gamma^H + \sqrt{\frac{\log(d(d+1)K/\zeta)}{2m}}\right)}_{\text{pol.eval. error}}.$$

In particular, for any $\varepsilon' > 0$, choosing K, H, m so that

$$\begin{aligned} K &\geq H_{\gamma,\gamma\varepsilon'/2} \\ H &\geq H_{\gamma,(1-\gamma)^2\varepsilon'/(8\sqrt{d})} \qquad \text{and} \\ m &\geq \frac{32d}{(1-\gamma)^4(\varepsilon')^2}\log((d+1)^2K/\zeta) \end{aligned}$$

policy π_K is δ-optimal with

$$\delta \leq \frac{2(1+\sqrt{d})}{(1-\gamma)^2}\varepsilon + \varepsilon',$$

while the total computation cost is $\mathrm{poly}(\frac{1}{1-\gamma}, d, \mathrm{A}, \frac{1}{(\varepsilon')^2}, \log(1/\zeta))$.

Offline planning

- This was called “global planning”
- Idea: Use the simulator to get a policy. Then keep the policy, and run with it.
- No simulator needed while using the policy.
- LSPI is offline planner with global access

From global to local access

https://arxiv.org/abs/2108.05533

Efficient Local Planning with Linear Function Approximation

Dong Yin, Botao Hao, Yasin Abbasi-Yadkori, Nevena Lazić, Csaba Szepesvári

Confident MC-LSPI

Algorithm 2 CONFIDENT MC-LSPI

Input: initial state ρ, initial policy parameter w_0, number of iterations K, regularization coefficient λ, threshold τ, discount γ, number of rollouts m, length of rollout n.

$\mathcal{C} \leftarrow \emptyset$ *// Initialize core set.*

for $a \in \mathcal{A}$ **do**

  if $\mathcal{C} = \emptyset$ or $\phi(\rho, a)^\top (\Phi_\mathcal{C}^\top \Phi_\mathcal{C} + \lambda I)^{-1} \phi(\rho, a) > \tau$ **then**

    $\mathcal{C} \leftarrow \mathcal{C} \cup \{(\rho, a, \phi(\rho, a), \text{none})\}$

  end if

end for

$z_q \leftarrow \text{none}, \ \forall z \in \mathcal{C}$ *// Policy iteration starts.* $(*)$

for $k = 1, \ldots, K$ **do**

  for $z \in \mathcal{C}$ **do**

    $\text{status}, \text{result} \leftarrow \text{CONFIDENTROLLOUT}(m, n, w_{k-1}, \gamma, z_s, z_a, \Phi_\mathcal{C}, \lambda, \tau)$

    if status $=$ done, **then** $z_q \leftarrow$ result; **else** $\mathcal{C} \leftarrow \mathcal{C} \cup \{\text{result}\}$ and **goto** line $(*)$

  end for

  $w_k \leftarrow (\Phi_\mathcal{C}^\top \Phi_\mathcal{C} + \lambda I)^{-1} \Phi_\mathcal{C}^\top q_\mathcal{C}$

end for

return w_{K-1}, or equivalently, the policy $\pi_{w_{K-1}}$ in the form of Eq. (3.1).

Algorithm 1 CONFIDENTROLLOUT

Input: number of rollouts m, length of rollout n, policy parameter w, discount γ, initial state s_0, initial action a_0, feature matrix $\Phi_{\mathcal{C}}$, regularization coefficient λ, threshold τ.
for $i = 1, \ldots, m$ **do**
  $s_{i,0} \leftarrow s_0$, $a_{i,0} \leftarrow a_0$, query the simulator, obtain reward $r_{i,0} \leftarrow r(s_{i,0}, a_{i,0})$, next state $s_{i,1}$.
  for $t = 1, \ldots, n$ **do**
    Query the simulator with $(s_{i,t}, a)$, $\forall a \in \mathcal{A}$.
    Obtain rewards $r(s_{i,t}, a)$, feature vectors $\phi(s_{i,t}, a)$, next states $s'_{i,t}(a)$, $\forall a \in \mathcal{A}$.
    if there exists $a' \in \mathcal{A}$ such that $\phi(s_{i,t}, a')^\top (\Phi_{\mathcal{C}}^\top \Phi_{\mathcal{C}} + \lambda I)^{-1} \phi(s_{i,t}, a') > \tau$ **then**
      status $\leftarrow$ uncertain, result $\leftarrow (s_{i,t}, a', \phi(s_{i,t}, a'),$ none$)$
      return status, result
    end if
    $a_{i,t} \leftarrow \arg\max_{a \in \mathcal{A}} w^\top \phi(s_{i,t}, a)$, $r_{i,t} \leftarrow r(s_{i,t}, a_{i,t})$, $s_{i,t+1} \leftarrow s'_{i,t}(a_{i,t})$.
  end for
end for
status $\leftarrow$ done, result $\leftarrow \frac{1}{m} \sum_{i=1}^{m} \sum_{t=0}^{n} \gamma^t r_{i,t}$
return status, result

Theorem

Under uniform ζ realizability, w.p. $\geq 1-\delta$, Confident MC-LSPI obtains an

$$\sqrt{d}\,\zeta\,H^2+\varepsilon$$

optimal policy with at most

$$O\left(\mathrm{poly}\left(d,A,B,H,1/\varepsilon,\log\left(1/\delta\right)\right)\right)$$

simulation calls, while the total computation cost is also polynomial in the same factors.

Here B is a bound on the 2-norm of the parameter vectors of policies.

Can we do better?

Theorem (Query lower bound: large action sets): For any $\varepsilon > 0, 0 < \delta \leq 1/2$, positive integer d and for any (δ, ε)-sound online planner $\mathcal{P}$ there exists a "featurized-MDP" (M, φ) with rewards in $[0, 1]$ with $\varepsilon^*(M, \Phi) \leq \varepsilon$ such that when interacting with a simulator of (M, φ), the expected number of queries used by $\mathcal{P}$ is at least

$$\Omega\left(\exp\left(\frac{1}{32}\left(\frac{\sqrt{d}\varepsilon}{\delta}\right)^2\right)\right).$$

..just another needle in the haystack argument..

and when $A = O(1)$?

Theorem (Query lower bound: small action sets, fixed-horizon objective): For $\varepsilon > 0$, $0 < \delta \leq 1/2$ and positive integer d, let

$$u(d, \varepsilon, \delta) = \left\lfloor \exp\left(\frac{d(\frac{\varepsilon}{2\delta})^2}{8} \right) \right\rfloor .$$

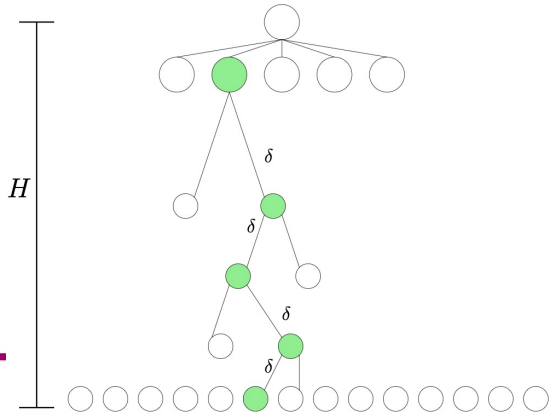
Then, for any $\varepsilon > 0$, $0 < \delta \leq 1/2$, positive integers A, H, d such that $d \leq \mathrm{A}^H$ and for any online planner $\mathcal{P}$ that is (δ, ε)-sound for MDPs with at most A actions and the H-step criterion, there exists a "featurized-MDP" (M, φ) with A actions and rewards in $[0, 1]$ such that when interacting with a simulator of (M, φ), the expected number of queries used by $\mathcal{P}$ is at least

$$\tilde{\Omega}\left(\frac{u(d, \varepsilon, \delta)}{d(\varepsilon/\delta)^2} \right)$$

provided that $\mathrm{A}^H > u(d, \varepsilon, \delta)$ ("large horizons"), while it is

$$\tilde{\Omega}\left(\frac{\mathrm{A}^H}{H} \right)$$

otherwise ("small horizon").



Questions from slack

Ehsan Imani 3 days ago

ML textbooks usually motivate PCA by noting that oftentimes real-world data is mostly within a small linear subspace of our d-dimensional space. Would an assumption like this rule out the pathological case of nearly-orthogonal vectors that lead to the lower bound?

+9

Jiamin He 11 hours ago

The approximation error appears in the bound of LSPI,

$$2\left(1+\sqrt{d}\right)\varepsilon/\left(1-\gamma\right)^{2},$$

can not be controlled by the algorithm. Looks like it is also a major concern of some people according to the endnote. Why shouldn't we be worried about it? Is it possible that controlling the size of the ball (epsilon) may also increase the dimension d? Would a large value of γ be a problem here?

$+\infty$?