IE 498: Online Learning and Decision Making, Fall 2019 Nov 4th

Otc 29: OLIVE Algorithm

Lecturer: Yuan Zhou *Scribe: Jinglin Chen, Tiancheng Qin*

In the previous lecture, we have shown a regret bound for the model-based RL problem. In this lecture, we switch to the value-based RL. In the value-based setting, there also exist tons of literature about the PAC or regret bound, and some of them attain the optimal rate. It seems that the theoretical foundation is pretty solid. However, on the other side, we know that almost all the existing RL algorithms (e.g. DQN (Mnih et al. [2015])) are brittle in the complex experimental experiment. Their results often chatter or even diverge, and are very sensitive to the hyperparameters or neural network structures. So, what is missing here?

The answer is the function approximation. Most known theoretical results consider the tabular case and the complexity has dependence on $|\mathcal{S}|$. In empirical, we usually have very large or even infinite state space $|\mathcal{S}|$, which makes the tabular algorithms intractable. Therefore, it is crucial for us to consider the theoretical guarantee for the RL problem with function approximation. In the value-based setting, our goal is to find a policy π that approximates π^*, and we are given an action-value function class $\mathcal{F}$. It suffices to find a $f \in \mathcal{F}$ that approximate Q^*, since the error between V^{π_f} (the greedy policy w.r.t. f) and V^* can be upper bounded by a factor multiplies $\|f - Q^*\|$. The (finite) function class $\mathcal{F}$ is much smaller than the all possible action-value function and we hope to pay only $\log|\mathcal{F}|$ dependence. In general, if we don't make assumption on the complexity of the environment, we will suffer intolerable $\exp(H)$ lower bound. Therefore, Jiang et al. [2017] proposed the Bellman rank M to capture such complexity of the environment and showed the Bellman rank is natural and small in many settings. They also proposed the OLIVE algorithm (Jiang et al. [2017]), which can learn efficiently. Specifically, OLIVE achieves a sample complexity upper bound that is polynomial to $H, |\mathcal{A}|, M, \log|\mathcal{F}|$ and some other standard parameters in statistical learning theory. We will discuss the OLIVE algorithm in the following.

Algorithm 1 OLIVE Algorithm

1. Initiate $\mathcal{F}_1 = \mathcal{F}$.
2. For $t = 1, 2, 3, \ldots$ Do
   - 2a. Select $f_t = \arg\max_{f \in \mathcal{F}_t}\{V_f\}$.
   - 2b. Run policy π_{f_t} for $\frac{cH^4}{\epsilon^2}\log(|\mathcal{F}_t|/\delta)$ times, and estimate $\hat{\mathcal{E}}_h^{(t)}(f_t, \pi_{f_t})$ and $\hat{\mathcal{E}}^{(t)}(f_t, \pi_{f_t})$.
   - 2c. If $\hat{\mathcal{E}}^{(t)}(f_t, \pi_{f_t}) < \epsilon/2$, Then return f_t.
   - 2d. Otherwise let $h_t = \arg\max_h |\hat{\mathcal{E}}_h^{(t)}(f_t, \pi_{f_t})|$. $(\max_h |\hat{\mathcal{E}}_h^{(t)}(f_t, \pi_{f_t})| \geq \epsilon/(2H))$
   - 2e. Collect $n = c\frac{A^2H^2}{\phi^2}\log|\mathcal{F}_t|/\delta$ trajectories; $s_1, \ldots, s_{h_t} \sim \pi_{f_t}, a_{h_t} \sim \text{Unif}(A), s_{h_t+1} \sim P_{h_t}(s_{h_t}, a_{h_t})$.
   - 2f. Estimate $\forall f \in \mathcal{F}_t, \hat{\mathcal{E}}_h^{(t)}(f, \pi_{f_t}) \triangleq \frac{1}{n}\sum_{i=1}^n A \cdot \mathbb{1}[a_{h_t}^i = \pi_f(s_{h_t}^i)](f(s_{h_t}^i, a_{h_t}^i) - r_{h_t}^i - \max_a\{f(s_{h_t+1}^i, a)\})$
   - 2g. Let $\mathcal{F}_{t+1} = \{f \in \mathcal{F}_t | \hat{\mathcal{E}}_{h_t}^{(t)}(f, \pi_{f_t}) \leq \phi/2\}$.

Definition 1. *For a given MDP with infinite states and horizon $[H]$, we define ζ to be the norm of the Bellman rank M if: $\forall h \in [H]$, $\exists \xi_h : \mathcal{F} \to \mathbb{R}^M$, $\nu_h : \mathcal{F} \to \mathbb{R}^M$, $\|\xi_h\|_2 \leq \zeta$, $\|\nu_h\|_2 \leq \zeta$, s.t. $\forall f, g \in$*

$\mathcal{F}, \mathcal{E}_h(f, \pi_g) = \langle \xi_h(f), \nu_h(g) \rangle$.

We assume that $Q^* \in \mathcal{F}$, and our goal is to calculate an (ϵ, δ)-PAC approximation of Q^* or equivalently V^*.

Theorem 1. *With probability at least $1-\delta$, we can find $\hat{f} \in \mathcal{F}$ so that $V^{\pi_{\hat{f}}} \geq V^* - \epsilon$ within poly$(H, |\mathcal{A}|, M, \log|\mathcal{F}|, 1/\epsilon, \ln(1/\delta))$ samples.*

Remark 1. *The realizability assumption that $Q^\star \in \mathcal{F}$ is natural. Otherwise, the sample complexity bound should contain an approximation error term. For simplicity, we only show the exact realizability case. Interesting readers can refer to Jiang et al. [2017] for extended results on approximate case.*

Proof sketch: The idea of the proof is to eliminate the sub-optimal function in $\mathcal{F}$, and stop until we find ϵ-optimal function. Claim 2 below tells us that the Q^* will never be eliminated and Claim 3 shows us that the function returned by the algorithm is ϵ-optimal. To guarantee the desired PAC bound, we still need to addition two things. One is such event E happens with a high probability (Claim 1). The other thing is that the elimination stops within polynomial time (Claim 4). To obtain Claim 4, we will need to apply the Claim 5.

In the following, we will give a formal statement of the claims and their proofs. Finally, we will prove the theorem.

To begin with, let's consider our favorable event $E : \forall t = 1, \ldots, H$, the estimation error at Step 2b is no larger than $\frac{\epsilon}{4H}$, and the estimation error at Step 2g is no larger than $\frac{\phi}{4}$.

Claim 1. $\mathbb{P}[E] \geq 1 - \delta$.

Proof. By the concentration inequality (Hoeffding). □

Claim 2. *$Q*$ will never be eliminated, i.e. $Q^* \in \mathcal{F}_t, \forall t$.*

Proof. $\hat{\mathcal{E}}^{(t)}_{h_t}(Q^*, \pi_{f_t}) = 0$ always holds so $Q^* \in \mathcal{F}_{t+1}$. □

Claim 3. *When f_t is returned, we have $V^{\pi_{f_t}} \geq V^* - \epsilon$.*

Proof.

$$\begin{aligned} V^{\pi_{f_t}} &= V_{f_t} - \mathcal{E}(f_t, \pi_{f_t}) \\ &= V_{f_t} - \hat{\mathcal{E}}^{(t)}(f_t, \pi_{f_t}) + [\hat{\mathcal{E}}^{(t)}(f_t, \pi_{f_t}) - \mathcal{E}(f_t, \pi_{f_t})] \\ &\geq V^* - \epsilon/2 - \epsilon/(4H) \times H \\ &\geq V^* - \epsilon. \end{aligned}$$

□

Claim 4. *$\forall h = 1, \ldots, H$, the elimination step for $h_t = h$ is run by $\lesssim M \log(MH\zeta/\epsilon)$ times.*

Proof. Fix h, let $t_1, t_2, \ldots,$ be the rounds for $h_t = h$.

Let B_1 be $\text{Ball}_\zeta(\vec{O})$ in $\mathbb{R}^M$ and for $i \geq 2$, let $\widetilde{B_i} = \{\vec{u} \in B_{i-1} : |\langle \vec{u}, \nu_h(f_{t_{i-1}}) \rangle| \leq \phi\}$. Also, we use B_i to represent the minimum (volume) ellipsoid covering $\widetilde{B_i}$.

Firstly, we verify via induction that $\forall f \in \mathcal{F}_{t_i}, \xi_h(f) \in \widetilde{B_i} \subseteq B_i$.

By Step 2d, we have that $\forall i, \exists \vec{p_i} = \nu_h(f_{t_i}),$ s.t. $\exists \vec{u} \in B_i, \vec{p_i}^\top \vec{u} > \frac{\epsilon}{4H}$.

By Claim 5, when $\phi \leq \frac{\epsilon}{4H} \cdot \frac{1}{1000\sqrt{M}}$, we have $vol(B_{i+1}) \leq 1/2 \cdot vol(B_i)$.

On the other hand, $vol(B_i) \geq (\phi/\zeta)^M B_M$. Combining these two inequalities, we get the number of eliminations $\leq \log_2 \frac{\zeta^M B_M}{(\phi/\zeta)^M B_M} = M \log \phi = M \log(MH\zeta/\epsilon)$. □

Claim 5. *Let $B = \{\vec{u} : \|M\vec{u}\|_2 \leq 1\}$ be an ellipsoid in $\mathbb{R}^d$. Suppose $\exists \vec{p} \in \mathbb{R}^d, \vec{w} \in B$, s.t. $\vec{p}^\top \vec{w} \geq \epsilon$. Also, we let $B' = \{\vec{u} \in B : |\vec{p}^\top \vec{u}| \leq \epsilon\delta\}$, where $\delta \leq \frac{1}{1000\sqrt{d}}$ and let B'' be the min covering of such B'. Then we have $vol(B'') \leq vol(B)/2$.*

Proof. Firstly, we have the inequality $\vec{p}^\top \vec{w} \geq \epsilon \Rightarrow |\langle M^{-1}\vec{p}, M\vec{w}\rangle| \geq \epsilon \Rightarrow \|M^{-1}\vec{p}\| \geq \epsilon$.

Let Q be the ellipsoid $Q = \{M\vec{u} : \vec{u} \in B'\} = \{M\vec{u} : |\langle M^{-1}\vec{p}, M\vec{u}\rangle| \leq \epsilon\delta, \|M\vec{u}\| \leq 1\} = \{\vec{q} : |\langle M^{-1}\vec{p}, \vec{q}\rangle| \leq \epsilon\delta, \|\vec{q}\| \leq 1\}$.

Let $\vec{r}$ be the normalized vector, $\vec{r} = \frac{M^{-1}\vec{p}}{\|M^{-1}\vec{p}\|}$ and let $Q' = \{\vec{q} : \langle \vec{r}, \vec{q}\rangle \leq \delta, \|q\| \leq 1\}$. We have that $Q \subset Q'$. Then we use Q'' to represent the min covering ellipsoid of Q'. We have

$$B' = M^{-1}Q \subset M^{-1}Q' \subset M^{-1}Q'' \Rightarrow vol(B'') \leq vol(M^{-1}Q'').$$

This means that

$$\frac{vol(B'')}{vol(B)} \leq \frac{vol(M^{-1}Q'')}{vol(M^{-1}Q)} \leq \frac{vol(Q'')}{vol(Q)}.$$

We can select M, rotate Q to B_d, and let $\vec{r} = (1, 0, \ldots, 0)^\top$ as shown Figure 1.

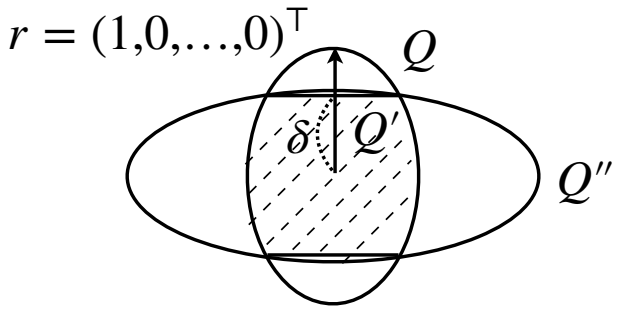

Figure 1: Construction

Assume that $Q' = \{(q_1, \ldots, q_d)^\top, |q_1| \leq \delta, \|q\|_2 \leq 1\}$. We can construct Q'' with the form $Q'' = \{\left(\frac{q_1}{H}\right)^2 + \sum_{i=2}^d \left(\frac{q_i}{R}\right)^2 \leq 1\}$. To guarantee $Q' \subset Q''$, we need $\left(\frac{q_1}{H}\right)^2 + \frac{1-\delta^2}{R^2} \leq 1$. We can pick $H = \frac{1}{3}$ and $R = 1 + 4\delta^2$, and check that such inequality holds. This implies that

$$vol(Q'') = B_d \cdot H \cdot R^{d-1} = B_d \cdot \frac{1}{3} \cdot (1 + 4\delta^2)^{d-1} \leq \frac{1}{2} B_d.$$

Therefore

$$\frac{vol(B'')}{vol(B)} \leq \frac{vol(Q'')}{vol(Q)} \leq \frac{B_d/2}{B_d} = \frac{1}{2},$$

which completes the proof. □

Proof of the theorem: By Claim 4, we know that the total number of episodes is $O(MH)$. In 2b and 2g, we both need polynomial samples (w.r.t. $H, |\mathcal{A}|, M, \log|\mathcal{F}|$). Therefore, we only need polynomial samples in total. By Claim 1, such event E happens with high probability ($\mathbb{P}(E) \geq 1-\delta$) and Claim 3 tells us that the returned policy is within ϵ range of the optimal policy. So we complete the proof.

References

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1704–1713. JMLR. org, 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.