IE 498: Online Learning and Decision Making, Fall 2019 10-24-2019

Lecture-18: Tighter Regret for Model Based Reinforcement Learning

Lecturer: Yuan Zhou *Scribe: Yifan Hu, Jafar Abbaszadeh Chekan*

1 Recap

In the last lecture, for an episodic MDP $m = (S, A, P, R, H)$ with transition matrix:

$$P_h : S \times A \times S \to [0, 1],\ P_h(s_h + 1|s_h, a_h),$$

we derived an upper bound on the regret, such that

$$R_k \lesssim |S|H^{1.5}\sqrt{|A|T\log(|S||A|T)} + H^2|S|^2|A|\log^2(|S||A|T), \tag{1}$$

where $\lesssim$ is less or equal to with respect to ignoring the logarithmic factors, H is the number of periods per episode, K is the number of episodes, $T := KH$ is the total number of periods, S and A are the finite state and action space, respectively, $|\cdot|$ represents the cardinality of a set. Note that we will use the space and its cardinality interchangeably.When T grows super large, the first term in (1) dominates. As a result, the upper bound of regret bound depends linearly on the cardinality of the state space.

Since the state space could be really large, the bounds might not scale well. In this lecture, we introduce a method to achieve a regret bound with $\sqrt{|S|}$ dependency on the state space S.

2 Method

Goal: In this lecture, we are interested to derive a new upper bound for the regret such that the parameter S appears in the radical, namely $\mathcal{O}(\sqrt{|S|})$. As we have shown in the last lecture Claim 2:

$$R_K \le \sum_{k=1}^{K} \mathbb{E}\tilde{\delta}_{k,l} + \delta T,$$

where $\tilde{\delta}_{k,h} := (\tilde{V}_{k,h} - V_k^{\pi_k})(x_{k,h})$. To achieve that, we define the event

$$E := \left\{ |\langle \hat{P}_{k,h}(s,a) - P_h(s,a), V^*_{h+1}\rangle| \le b_{k,h}(s,a), \forall k, h, s, a,\ \text{where } b_{k,h}(s,a) \le CH\sqrt{\frac{\ln(SAT/\delta)}{n_{k,h}(s,a)}} \right\}.$$

Recall Claim 3 in the last lecture:

$$\mathbb{E}[\tilde{\delta}_{k,h}] = \mathbb{E}\left[\langle (\hat{P}_{k,h} - P_h)(x,a), \tilde{V}_{k,h+1}\rangle + b_{k,h}(x,a)\right] + \mathbb{E}\tilde{\delta}_{k,h+1},$$

where $s = s_{k,h}$, $a = a_{k,h} = \pi_{k,h}(s,\cdot)$. In the previous lecture, we bound each term in $\Delta_{k,h} := \langle (\hat{P}_{k,h} - P_h)(x,a), \tilde{V}_{k,h+1}\rangle$ separately, each bounded by $\mathcal{O}\sqrt{|S|}$. Although $\tilde{V}_{k,h+1}$ is a random variable, after iterating

for some time, intuitively its value should be close to a constant. If we bound $(*)$ as a whole with a more delicate inequality, we might be able to achieve a $\mathcal{O}\sqrt{|S|}$ bound.

To bound $(*)$, recall Claim 4 in the last lecture:

$$\mathbb{E}\Delta_{k,h} \lesssim \delta_H + H\sqrt{\frac{S}{n}\log\frac{SAT}{\delta}} + \frac{HS}{n}\log\frac{SAT}{\delta},$$

where $n = n_{k,h}(s)$. To come up a better bound on $\mathbb{E}\Delta_{k,h}$, we decompose it as

$$\mathbb{E}\Delta_{k,h} = \mathbb{E}\langle(\hat{P}_{k,h} - P_h)(x,a), V^*_{h+1}\rangle + \mathbb{E}\langle(\hat{P}_{k,h} - P_h)(x,a), \tilde{V}_{k,h+1} - V^*_{h+1}\rangle.$$

The first term could be bounded by Hoeffding inequality and satisfaction of event E. As for the second term, we know the term $\tilde{V}_{k,h+1} \to V^*$ as $t \to +\infty$. We denote the second term as $(*) := \mathbb{E}\langle(\hat{P}_{k,h} - P_h)(x,a), \tilde{V}_{k,h+1} - V^*_{h+1}\rangle$. To upper bound $(*)$, we have

$$\begin{aligned}
(*) &\le \mathbb{E}\left[\left\langle(\hat{P}_{k,h} - P_h)(x,a), \tilde{V}_{k,h+1} - V^*_{h+1}\right\rangle \middle| E\right]\mathbb{P}(E) + 2H\mathbb{P}(\bar{E}) \\
&\le \mathbb{E}\left[\left\langle\left|(\hat{P}_{k,h} - P_h)(x,a)\right|, \tilde{V}_{k,h+1} - V^*_{h+1}\right\rangle \middle| E\right]\mathbb{P}(E) + 2H\delta \\
&\le \mathbb{E}\left[\left\langle\left|(\hat{P}_{k,h} - P_h)(x,a)\right|, \tilde{V}_{k,h+1} - V^*_{h+1}\right\rangle \middle| E\right] + 2H\delta \\
&\le \mathbb{E}\sum_{y\in S}\left|(\hat{P}_{k,h} - P_h)(y|x,a)\right|\left(\tilde{V}_{k,h+1} - V^*_{h+1}\right)(y) + 2H\delta;
\end{aligned} \tag{2}$$

Note that the absolute value is pointwise absolute value. From the first inequality to the second one, we use that fact that given event E that $\tilde{V}_{k,h+1} \ge V^{\pi_h}_{h+1}$. Recalling the event F:

$$F \triangleq \{\forall\, s,y,a,k,h,\ n = n_{k,h}(s,a), \mu = P_h(y|s,a), |\hat{(P_{k,h}} - P_h)(y|s,a)| \le C_1\sqrt{\frac{\mu}{n}\log\frac{SAT}{\delta}} + \frac{\log(SAT/\delta)}{n}\}$$

when F holds, the right hand side of the last inequality (2) can be upper-bounded by Bernstein inequality,

$$(*) \le \mathbb{E}\sum_{y\in S}\left(\sqrt{\frac{p(y)}{n}\log\frac{SAT}{\delta}} + \frac{\log(SAT/\delta)}{n}\right)\left(\tilde{V}_{k,h+1} - V^*_{h+1}\right)(y) + 2H\delta \tag{3}$$

Using AM-GM inequality (more specifically e.g. $\sqrt{ab} \le \frac{a+b}{2}$ property), we have

$$\sqrt{\frac{p(y)}{n}\log\frac{SAT}{\delta}} + \frac{\log(SAT/\delta)}{n} \le \frac{p(y)}{H} + \frac{H}{n}\log\frac{SAT}{\delta}.$$

Note that here H could be replaced by H^2 or even higher order. As a result, we have,

$$\begin{aligned}
(*) &\le \mathbb{E}\sum_{y\in S}\left(\frac{p(y)}{H} + \frac{H}{n}\log\frac{SAT}{\delta}\right)\left(\tilde{V}_{k,h+1} - V^*_{h+1}\right)(y) + 2H\delta \\
&\le \frac{1}{H}\mathbb{E}\tilde{\delta}_{k,h+1} + |S|H^2\log\frac{SAT}{\delta} + 2H\delta
\end{aligned} \tag{4}$$

Claim 5. We have

$$\mathbb{E}\tilde{\delta}_{k,h} \le \left(1 + \mathcal{O}\left(\frac{1}{H}\right)\right)\mathbb{E}\tilde{\delta}_{k,h+1} + \mathbb{E}\left[b_{k,h}(s,a) + \mathcal{O}\left(|S|H^2\log\frac{SAT}{\delta} + 2H\delta\right)\right] \tag{5}$$

Claim 6. We have

$$\mathbb{E}\tilde{\delta}_{k,1} \lesssim \mathbb{E}\left[b_{k,h}(s,a) + |S|H^2 \log\frac{SAT}{\delta} + 2H\delta\right] \tag{6}$$

As a result, invoking event E and Claim 6 we could bound the regret by

$$\begin{aligned} R_K &\lesssim \delta_T^2 + \mathbb{E}\sum_{k=1}^{K}\sum_{h=1}^{H}\left(H\sqrt{\frac{\log(SAT/\delta)}{n_{k,h}(s_{k,h},a_{k,h})}} + \frac{|S|H^2}{m}\log\frac{SAT}{\delta}\right) \\ &\lesssim \delta_T^2 + \mathbb{E}\sum_{s,a,h}\left(\sum_{m=1}^{n_{k,h}(s,a)} H\sqrt{\frac{\log(SAT/\delta)}{m}} + \frac{|S|H^2}{m}\log\frac{SAT}{\delta}\right) \\ &\lesssim \delta_T^2 + \mathbb{E}\sum_{s,a,h}\left(H\sqrt{n_{k,h}(s,a)\log(SAT/\delta)} + +|S|H^2\log\frac{SAT}{\delta}\right) \\ &\lesssim \delta_T^2 + \mathbb{E}\sqrt{SAH}H\sqrt{\sum_{s,a,h} n_{k,h}(s,a)\log(SAT/\delta)} + |S|^2H^3A\log^2\frac{SAT}{\delta}. \end{aligned} \tag{7}$$

The second inequality comes from a exchange of summation from episodes and periods to number of times a state and action pair are visited. And, conclusively, we have the following regret bound:

$$R_K \lesssim \delta T^2 + H^{1.5}\sqrt{SAT\log\frac{SAT}{\delta}} + S^2AH^3\log^2\frac{SAT}{\delta} \tag{8}$$

which is donimated by the second term and the upper bound of regret is $\mathcal{O}(H^{1.5}\sqrt{SAT\log(SAT/\delta)})$.

3 Policy Learning for MDP with low Bellman Rank

There are huge gaps between reinforcement learning theory to applications. Practically, the problem might have huge state space. There exists worst cases for reinforcement learning problem with large state space and no structure on the model, such that the regret bound depends linearly on the size of state space. However, some of the practically problem could be solved quite well. In this section, we are going to consider the policy learning algorithms from theory point of view, which is the most powerful algorithm one could have a theoretical proof.

In this section, We will consider the policy learning for MDP with low Bellman rank. Sample complexity, instead of regret bound, is used as the performance measure. Note that sample complexity is a weaker measure since low regret almost surely implies low sample complexity.

3.1 Setting

We consider a episodic layered MDP, with state space $S := S_1 \bigcup S_2 \bigcup \ldots \bigcup S_H$, with a known and deterministic reward function $r : S \times A \to [0,1]$. We define a hypothesis class $\mathcal{F} = \{f : S \times A \to R\}$. Note that the hypothesis in the class corresponds to some policies. With the realizability assumption and Bellman operator, we could define the Bellman error.

Assumption:(realizability) $Q^* \in \mathcal{F}$.

Recall: The Bellman operator $\mathcal{T}$ such that

$$\mathcal{T}f(s,a) = \mathbb{E}_{s'\sim p(s,a)}V_f(s') + r(s,a),$$

where $V_f(s') = \max_{a'} f(s', a')$.

Now we could define the Bellman error as follows:

Definition:(Bellman error) $\forall h \in [H]$, an roll-in policy π, hypothesis $f \in \mathcal{F}$, the Bellman error ϵ_f is defined as

$$\epsilon(f) := \sum_{h=1}^{H} \epsilon_h(f, \pi_f),$$

where

$$\mathcal{E}_h(\mathcal{F}, \pi_f) \stackrel{\Delta}{=} \mathop{\mathbb{E}}_{\substack{a_h \sim \pi_{\mathcal{F}}(s_h) \\ s_{h+1} \sim P(s_h, a_h) \\ a_{h+1} \sim \pi_{\mathcal{F}}(s_{h+1})}} \left[f(s_h, a_h) - r(s_h, a_h) - f(s_{h+1}, a_{h+1}) \right] \tag{9}$$

Based on the definition, we have the following claims:

Claim 1: Q^* is consistent, namely $\epsilon(Q^*) = 0$.

Claim 2: $\forall f \in \mathcal{F}$,

$$\epsilon(f) := \underbrace{V_f}_{\text{Claimed Value}} - \underbrace{V^{\pi_f}}_{\text{real value obtained by greedy policy}}$$

Then we define the Bellman Rank

Definition(Bellman Rank) An MDP and a hypothesis class $\mathcal{F}$ admits Bellman rank M and norm ζ, if $\forall h$, $\exists \xi_h : \mathcal{F} \to R^M$, $V_h : \mathcal{F} \to R^M$, such that

$$\|\xi_h(f)\|_2 \leq \zeta; \quad \|V_h(f)\|_2 \leq \zeta; \forall f \in \mathcal{F},$$

and $\forall f, g \in \mathcal{F}$,

$$\epsilon_h(f, \pi_g) = \langle \xi_h(f), V_h(g) \rangle$$

.

Claim 3: If an MDP has low rank transition matrix, it means that for $\forall h$, $\mathcal{T} \in R^{(S \times A) \times S}$, $\mathcal{T} = \mathcal{T}_1 \cdot \mathcal{T}_2$, where $\mathcal{T}_\infty \in R^{(S \times A) \times M}$, $\mathcal{T}_\in \in R^{M \times S}$. Then the MDP and any $\mathcal{F}$ admits Bellman rank M and norm $\zeta = 2\sqrt{M}$.

Note that this low Bellman rank is also related to a class of problems called MDP with rich observations, it is a special MDP with a few hidden states and large number of observable states, the reward function depends on the action and the hidden states pair instead of the action and the observable state pair, see [1] for more details.

References

[1] 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.