EE 290 Theory of Multi-armed Bandits and Reinforcement Learning Lecture 18 - 3/18/2020

Lecture 18: Rmax Exploration

Lecturer: Jiantao Jiao *Scribe: Samyak Parajuli, Dimitris Papadimitriou*

In this lecture, we finish up the discussion between fixed horizon and infinite horizon and then present the Rmax algorithm.

1 Fixed Horizon and Infinite Horizon

Let the value iteration outputs be denoted with $V^{\pi,H}(s) = \mathbb{E}[\sum_{t=1}^{H} \gamma^{t-1} r_t | s_1 = s, \pi]$, where π is a fixed time-invariant policy. Furthermore, let $\pi^\star$ denote the optimal policy for the discounted infinite horizon problem. Then, the following relation holds

$$V^{\pi^\star,H} \leq V^{\star,H}(s), \tag{1}$$

where the left hand side is a particular finite horizon policy and the right hand side is the optimal value for all possible policies. Furthermore,

$$V^{\star,H}(s) \leq V^\star(s) \tag{2}$$

also holds with the right hand side denoting the optimal value for the discounted infinite horizon problem. Due to the truncation effect $V^{\pi^\star,H}(s) \geq V^\star(s) - \gamma^H \frac{R_{\max}}{1-\gamma}$ where the subtracted quantity denotes the expected reward from time step $H+1$ to infinity taking into consideration the discount factor γ. Using (2) we obtain

$$V^\star(s) - V^{\star,H}(s) \leq \gamma^H \frac{R_{\max}}{1-\gamma}, \tag{3}$$

where $V^\star(s)$ is the optimal value and $V^{\star,H}(s)$ is the output of value iteration after H steps.

Lemma 1. $\|V^\star(s) - V^{\pi_f}(s)\|_\infty \leq \frac{2\|f - Q^\star\|_\infty}{1-\gamma}$, *with* $f \in \mathbb{R}^{S\times A}$ *and* $\pi_f(s) = \arg\max_{a\in A} f(s,a)$.

Proof For any $s \in \mathcal{S}$,

$$\begin{aligned} V^\star(s) - V^{\pi_f}(s) &= Q^*(s,\pi^*(s)) - Q^*(s,\pi_f(s)) + Q^*(s,\pi_f(s)) - Q^{\pi_f}(s,\pi_f(s)) \\ &\leq Q^*(s,\pi^*(s)) - f(s,\pi^*(s)) + f(s,\pi_f(s)) - Q^*(s,\pi_f(s)) + \gamma \mathbb{E}_{s'\sim P(s,\pi_f(s))}\left[V^*(s') - V^{\pi_f}(s')\right] \\ &\leq 2\|f - Q^*\|_\infty + \gamma \|V^* - V^{\pi_f}\|_\infty. \end{aligned}$$

□

Claim: Output the non-stationary policy in value iteration $\tilde{\pi} = \pi_Q^{*,H}, \pi_Q^{*,H-1}, \ldots, \pi_Q^{*,1}$ and arbitrary policies for the future then $\|V^* - V^{\tilde{\pi}}\|_\infty \leq \gamma^H \frac{R_{\max}}{1-\gamma}$.

2 Rmax Exploration Algorithm

We will assume that the reward function $R(s,a)$ is completely known, and the initial state distribution is known, but the transition model is unknown. The sampling model is an infinite horizon discounted MDP. We start with an initial policy and perform a rollout $s_1, a_1, \ldots s_t, a_t$ for a finite amount of time. We collect the data, update our policy, obtain another rollout, get a new policy and so on and so forth. We will output a policy such that $V_M(\hat{\pi}) \geq V_M^\star - \epsilon V_{\max}$, where $V_M(\hat{\pi})$ is the value for a particular MDP and $V_{\max} = \frac{R_{\max}}{1-\gamma}$.

2.1 Algorithm

The Rmax algorithm takes as input a threshold parameter m.We denote with $n(s,a)$ the visitation count for (s,a). We denote with $n(s,a,s')$ the visitation count for (s,a,s'). The known set K is defined as $K \triangleq \{(s,a)|n(s,a)=m\}$. Intuitively, if we have a state action pair that we've visited many times, we should have a good idea of the transition from that pair.

Step 1) Build an MDP $\hat{M}_k$ with transitions

$$\hat{P}_k(s'|s,a) = \begin{cases} \frac{n(s,a,s')}{n(s,a)} & \text{if } (s,a) \in k, \\ I(s'=s) & o.w. \end{cases}$$

and a reward function

$$\hat{R}_k(s,a) = \begin{cases} R(s,a) & \text{if } (s,a) \in K, \\ R_{\max} & o.w. \end{cases}.$$

Step 2) Rollout policy $\pi^\star_{\hat{M}_k}$ and collect new trajectory $s_1, a_1, s_2, a_2, \ldots$.
Step 3) For each time step h: if the count of the (state, action, next state) tuple is less than m then we increment by 1, otherwise we continue. Before digging into the details of $Rmax$, we first define some notation.

Definition 2. *Suppose MDPs M_1, M_2 are only different in dynamics and denote the transition functions as P_1, P_2. Then the dist$(M_1, M_2) = \max_{s\in S, a\in A} \|P_1(s,a)) - P_2(s,a))\|_1$ where $P_i(s,a) \in \mathbb{R}^S$, $i=1,2$ and the ℓ_1 norm represents the summation of absolute values.*

Definition 3. *The induced MDP M_k is defined as*

$$P_k(s'|s,a) = \begin{cases} P(s'|s,a) & \textit{if } (s,a) \in K \\ I(s',s) & o.w. \end{cases}$$

and

$$R_k(s,a) = \begin{cases} R(s,a) & \textit{if } (s,a) \in K \\ R_{\max} & o.w. \end{cases}$$

.
When m is large enough, $\hat{M}_k$ should be close to M_k.

Lemma 4. *For fixed (s,a), let $\hat{p}$ be the empirical distribution of m iid samples from $p(s,a)$. Then w.p. $\geq 1-\delta$,*

$$\|\hat{p} - p(s,a))\|_1 \lesssim \sqrt{\frac{1}{m}(s + \log(\frac{1}{\delta}))},$$

where s is the support size of the distribution $p(s,a)$.

Proof Note that for any vector $v \in \mathbb{R}^s$, $\|v\|_1 = \sup_{u\in\{-1,1\}^s} u^T v$. $u^T\hat{p}$ is the average of i.i.d random variables with bounded range, so we can use Hoeffding's inequality and union bound over all u to get:

$$\max_{s,a} \max_{u\in\{-1,1\}^s} u^T(\hat{p} - p(s,a)) = \max_{s,a} \|\hat{p} - p(s,a)\|_1 \lesssim \sqrt{\frac{1}{m}(s + \log(\frac{1}{\delta}))}.$$

□

Lemma 5. *If M_1, M_2 only differ in transitions, then $\|V^\star_{M_1} - V^\star_{M_2}\|_\infty \leq$ dist$(M_1, M_2)\frac{V_{\max}}{2(1-\gamma)}$*

Proof Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be the Bellman update operators of M_1 and M_2 respectively.

$$\begin{aligned}
&\|V^*_{M_1} - \mathcal{T}_2 V^*_{M_1}\|_\infty = \|\mathcal{T}_1 V^*_{M_1} - \mathcal{T}_2 V^*_{M_1}\|_\infty \\
&= \gamma \max_{s,a \in S \times A} |\mathbb{E}_{s' \sim P_1(s,a)}[V^*_{M_1}(s')] - \mathbb{E}_{s' \sim P_2(s,a)}[V^*_{M_1}(s')]| \\
&= \gamma \max_{s,a \in S \times A} \langle P_1(s,a) - P_2(s,a), V^*_{M_1} - V_{\max}/2 \cdot \mathbf{1}_{|S|\times 1} \rangle \\
&\leq \gamma \max_{s,a \in S \times A} \|P_1(s,a) - P_2(s,a)\|_1 \|V^*_{M_1} - V_{\max}/2 \cdot \mathbf{1}_{|S|\times 1}\|_\infty \\
&\leq \text{dist}(M_1, M_2) V_{\max}/2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|V^*_{M_1} - V^*_{M_2}\|_\infty &= \|V^*_{M_1} - \mathcal{T}_2 V^*_{M_1} + \mathcal{T}_2 V^*_{M_1} - \mathcal{T}_2 V^*_{M_2}\|_\infty \\
&\leq \text{dist}(M_1, M_2) \cdot V_{\max}/2 + \|\mathcal{T}_2 V^*_{M_1} - \mathcal{T}_2 V^*_{M_2}\|_\infty \\
&\leq \text{dist}(M_1, M_2) \cdot V_{\max}/2 + \gamma \|V^*_{M_1} - V^*_{M_2}\|_\infty.
\end{aligned}$$

□

Lemma 6. *If M_1, M_2 only differ in transitions then $\forall \pi : \mathcal{S} \to \mathcal{A}$:*

$$|V_{M_1}(\pi) - V_{M_2}(\pi)| \leq \textit{dist}(M_1, M_2) \frac{V_{\max}}{2(1-\gamma)}.$$

Proof

$$\begin{aligned}
|V_{M_1}(\pi) - V_{M_2}(\pi)| &= |R_1(s,\pi) + \gamma \langle P_1(s,\pi), V^\pi_{M_1} \rangle - R_2(s,\pi) - \langle P_2(s,\pi), V^\pi_{M_2} \rangle| \\
&\leq \gamma |\langle P_1(s,\pi), V^\pi_{M_1} \rangle - \langle P_2(s,\pi), V^\pi_{M_1} \rangle + \langle P_2(s,\pi), V^\pi_{M_1} \rangle - \langle P_2(s,\pi), V^\pi_{M_2} \rangle| \\
&\leq \gamma |\langle P_1(s,\pi) - P_2(s,\pi), V^\pi_{M_1} \rangle| + \gamma \|V^\pi_{M_1} - V^\pi_{M_2}\|_\infty \\
&= \gamma |\langle P_1(s,\pi) - P_2(s,\pi), V^\pi_{M_1} - \frac{R_{\max}}{2(1-\gamma)} \mathbf{1} \rangle| + \gamma \|V^\pi_{M_1} - V^\pi_{M_2}\|_\infty \\
&\leq \gamma \|P_1(s,\pi) - P_2(s,\pi)\|_1 \|V^\pi_{M_1} - \frac{R_{\max}}{2(1-\gamma)}\|_\infty + \gamma \|V^\pi_{M_1} - V^\pi_{M_2}\|_\infty \\
&\leq \gamma \frac{\text{dist}(M_1, M_2) R_{\max}}{2(1-\gamma)} + \gamma \|V^\pi_{M_1} - V^\pi_{M_2}\|_\infty \\
&\leq \text{dist}(M_1, M_2) \frac{V_{\max}}{2(1-\gamma)}.
\end{aligned}$$

The above holds for all $s \in S$ and hence we can take the infinity norm on the left hand side to obtain the final result. We also subtract the quantity $\frac{R_{\max}}{2(1-\gamma)}\mathbf{1}$ to center the range of $V^\pi_{M_1}$, exploiting the fact that P_1 and P_2 are valid probability distributions. Finally, we use the results from Lemma 5, and we neglect a γ term for clarity to obtain the last expression. □

Intuitively, what theses lemmas say is that having two MDPs which differ only in transitions, which are close to each other, then the evaluations of the policy cannot be very different since reward functions are the same.

Lemma 7. *Suppose M_1, M_2 agree on $K \subset S \times A$ in terms of rewards and dynamics. Then*

$$|V_{M_1}(\pi) - V_{M_2}(\pi)| \leq V_{\max} P_{M_1}(\textit{under } \pi \textit{ trajectory goes out of } K).$$

Proof Let $R_M(\tau)$ denote the sum of discounted rewards in a trajectory τ, according to the reward function of M. We write $v^\pi_{M_1} = \sum_\tau P_{M_1}[\tau|\pi]R_{M_1}(\tau)$ and $v^\pi_{M_2} = \sum_\tau P_{M_2}[\tau|\pi]R_{M_2}(\tau)$. We consider the trajectories τ for which $escape_K(\tau)$ equals 1, where $escape_K(\tau)$ is 1 if the arbitrary τ visits some $(s,a) \notin K$ and 0 otherwise. We define $\text{pre}_K(\tau)$ as the "prefix" of τ where every state action pair is in the known set except the last one. We also define $\text{suf}_K(\tau)$ which is the remainder of the episode. Let $R(\text{pre}_K(\tau))$ be the sum of discounted rewards within the prefix (or suffix), and $P_{M_1}[\text{pre}_K(\tau)|\pi]$ be the marginal probability of the prefix assigned by M_1 under policy π. We can now upper bound $V_{M_1}(\pi) - V_{M_2}(\pi)$ (the other direction is similar) as follows:

$$\begin{aligned}
V_{M_1}(\pi) &= \sum_{\tau:\, escape_K(\tau)=1} P_{M_1}[\tau|\pi](R_{M_1}(\text{pre}_K(\tau)) + R_{M_1}(\text{suf}_K(\tau))) + \sum_{\tau:\, escape_K(\tau)=0} P_{M_1}[\tau|\pi]R_{M_1}(\tau) \\
&\leq \sum_{\tau:\, escape_K(\tau)=1} P_{M_1}[\tau|\pi](R_{M_1}(\text{pre}_K(\tau)) + V_{\max}) + \sum_{\tau:\, escape_K(\tau)=0} P_{M_1}[\tau|\pi]R_{M_1}(\tau) \\
&\leq \sum_{\text{pre}_K(\tau)} P_{M_1}[\text{pre}_K(\tau)|\pi](R(\text{pre}_K(\tau)) + V_{\max}) + \sum_{\tau:\, escape_K(\tau)=0} P_{M_1}[\tau|\pi]R_{M_1}(\tau).
\end{aligned}$$

The last inequality comes from the fact that for any τ that shares the same prefix, we can combine their probabilities because $R(\text{pre}_K(\tau)) + V_{\max}$ does not depend on the suffix which gives us the marginal probability of the prefix. We lower bound $V^\pi_{M_2}$ by relaxing $R(\text{suf}_K(\tau))$ to 0 and we get

$$V_{M_2}(\pi) \geq \sum_{\text{pre}_K(\tau)} P_{M_2}[\text{pre}_K(\tau)|\pi]R_{M_2}(\text{pre}_K(\tau))) + \sum_{\tau:\, escape_K(\tau)=0} P_{M_2}[\tau|\pi]R_{M_2}(\tau).$$

We observe that when $escape_K(\tau) = 0$, $P_{M_1}[\tau|\pi] = P_{M_2}[\tau|\pi]$ and $R_{M_1}(\tau) = R_{M_2}(\tau)$ because since the trajectory does not go out of K, M_1 and M_2 will assign the same probability to τ. Therefore, when we subtract the above two inequalities we get

$$V_{M_1}(\pi) - V_{M_2}(\pi) \leq \sum_{\text{pre}_K(\tau)} P_{M_1}[\text{pre}_K(\tau)|\pi]V_{\max}.$$

We then get the result by noticing that the sum of probabilities is equal to $P_{M_1}(\text{under } \pi\ escape_K(\tau))$.

□

References

[1] "UIUC, CS 598, lecture notes 1," https://nanjiang.cs.illinois.edu/files/cs598/note1.pdf.

[2] "UIUC, CS 598, lecture notes 3," https://nanjiang.cs.illinois.edu/files/cs598/note3.pdf.

[3] "UIUC, CS 598, lecture notes 7," https://nanjiang.cs.illinois.edu/files/cs598/note7.pdf.