Notes on Rmax exploration

Nan Jiang

November 13, 2020

In this note we introduce and analyze Rmax, a PAC exploration algorithm for tabular MDPs [1]. Rmax builds on and simplifies the ideas from E^3, the first PAC-MDP algorithm [2]. For simplicity we will adapt Rmax to the episodic setting and provide PAC guarantee (instead of mistake bound / regret guarantees). See Sham Kakade's thesis [3] for related analyses.

1 Setup

We consider episodic RL problems where the environment is specified by an infinite-horizon discounted MDP $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma, d_0)$, but an episode always terminates in finitely many steps.[1] $\mathcal{S}$ is the finite state space. $\mathcal{A}$ is the finite action space. $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function. $R : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$ is the reward function. $d_0 \in \Delta(\mathcal{S})$ is the initial distribution. For simplicity we assume that R and d_0 are known. Given a policy $\pi : \mathcal{S} \to \mathcal{A}$, the ultimate measure of π's performance is $J_M(\pi) := \mathbb{E}[\sum_{h=1}^{\infty} \gamma^{h-1} r_h \,|\, \pi]$; here the subscript emphasizes the fact that we are ultimately interested in the performance of a policy in the true MDP M. The value of a policy has bounded range $[0, V_{\max}]$ where $V_{\max} = R_{\max}/(1-\gamma)$.

Our goal is to collect data for poly$(|\mathcal{S}|, |\mathcal{A}|, 1/(1-\gamma), 1/\epsilon, 1/\delta)$ episodes and return a policy $\widehat{\pi}$, such that with probability at least $1-\delta$,

$$J_M(\widehat{\pi}) \geq J_M^\star - \epsilon \cdot V_{\max}.$$

Here $J_M^\star := J_M(\pi^\star)$ and $\epsilon \in [0, 1]$ is the relative suboptimality.

2 Algorithm

The Rmax algorithm maintains the following quantities:

1. $n(s, a)$ is the visitation count to each $(s, a) \in \mathcal{S} \times \mathcal{A}$, initialized as 0.
2. $n(s, a, s')$ is the number of times we observe transition tuples $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$.

Rmax takes a threshold parameter m. At any point of execution, define $K := \{(s, a) : s \in \mathcal{S}, a \in \mathcal{A}, n(s, a) = m\}$, which is the *known* set of state-action pairs. The algorithm repeats the following steps until it finds a near-optimal policy (the detection of success is relatively easy so we omit here):

[1] We adopt both discouting and finite horizon assumption only to make notations easier.

1. Build an MDP $\widehat{M}_K$ as follows: for any $s \in \mathcal{S}$, $a \in \mathcal{A}$, $s' \in \mathcal{S}$

$$\widehat{P}_K(s'|s,a) := \begin{cases} n(s,a,s')/n(s,a), & \text{if } (s,a) \in K \\ \mathbb{I}[s'=s], & \text{otherwise} \end{cases}, \qquad \widehat{R}_K(s,a) := \begin{cases} R(s,a), & \text{if } (s,a) \in K \\ R_{\max}, & \text{otherwise} \end{cases}.$$

2. Collect an episode $s_1, a_1, \ldots, s_H, a_H$ by policy $\pi^\star_{\widehat{M}_K}$.

3. If any observed (s_h, a_h) for $1 \le h < H$ has $n(s_h, a_h) < m$, increment both $n(s_h, a_h)$ and $n(s_h, a_h, s_{h+1})$ by 1.

3 Analysis

Definition 1 (State-action occupancy)**.** Define $d^\pi_M(s,a) := (1-\gamma)\sum_{h=1}^\infty \gamma^{h-1}\mathbb{P}_M[s_h = s, a_h = a|\pi]$.

Definition 2 (ℓ_1 difference in transition function)**.** Given MDPs M_1, M_2 that only differ in their transition functions (P_1 and P_2 respectively), define

$$\text{distance}(M_1, M_2) := \max_{s\in\mathcal{S}, a\in\mathcal{A}} \|P_1(s,a) - P_2(s,a)\|_1.$$

Definition 3 (Induced MDP)**.** Define M_K as the "expected version" of $\widehat{M}_K$: for any $s \in \mathcal{S}$, $a \in \mathcal{A}$, $s' \in \mathcal{S}$,

$$P_K(s'|s,a) := \begin{cases} P(s'|s,a), & \text{if } (s,a) \in K. \\ \mathbb{I}[s'=s], & \text{otherwise.} \end{cases}, \qquad R_K(s,a) := \begin{cases} R(s,a), & \text{if } (s,a) \in K. \\ R_{\max}, & \text{otherwise.} \end{cases}$$

Fact 4 (optimism)**.** By construction, for any $\pi : \mathcal{S} \to \mathcal{A}$, $J_{M_K}(\pi) \ge J_M(\pi)$.

Lemma 1. *For any fixed (s,a), let $\widehat{p}$ be the empirical frequency of $P(s,a)$ based on m i.i.d. samples of $s' \sim P(s,a)$. With probability at least $1-\delta$,*

$$\|\widehat{p} - P(s,a)\|_1 \le \sqrt{\frac{2}{m}\log\frac{2\cdot(2^{|\mathcal{S}|}-2)}{\delta}}.$$

Proof. See Section 2.2 in Note 3. □

Lemma 2. *Suppose MDPs M_1 and M_2 only differ in dynamics. Then $\|V^\star_{M_1} - V^\star_{M_2}\|_\infty \le$ distance$(M_1, M_2) \cdot \frac{V_{\max}}{2(1-\gamma)}$.*

Table 1: Relationship between M, M_K, and $\widehat{M}_K$.

	M	M_K	$\widehat{M}_K$
Known (K)	$= M$	$= M$	$\approx M$
Unknown	$= M$	self-loop	self-loop

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

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

Therefore,

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

Solving for the inequality yields the result. □

Lemma 3 (Simulation lemma)**.** *Suppose M_1 and M_2 only differ in dynamics. Then $\forall \pi : \mathcal{S} \to \mathcal{A}$,*

$$|J_{M_1}(\pi) - J_{M_2}(\pi)| \le \textit{distance}(M_1, M_2) \cdot \frac{V_{\max}}{2(1-\gamma)}.$$

We already proved it in Note 3 (Lemma 1 with $\epsilon_R = 0$; a γ factor is dropped for convenience).

Lemma 4 (Induced Inequality)**.** *Suppose MDPs M_1 and M_2 agree exactly on $K \subseteq \mathcal{S}\times\mathcal{A}$ in terms of reward and dynamics. Let $\textit{escape}_K(\tau)$ be 1 if the trajectory τ visits some $(s,a) \notin K$, and 0 otherwise. $\forall \pi : \mathcal{S} \to \mathcal{A}$,*

$$|J_{M_1}(\pi) - J_{M_2}(\pi)| \le V_{\max} \cdot \mathbb{P}_{M_1}[\textit{escape}_K(\tau) \,\big|\, \pi].$$

Proof. Let $R_M(\tau)$ denote the sum of discounted rewards in τ according to the reward function of M. We can write $v^\pi_{M_1} = \sum_\tau \mathbb{P}_{M_1}[\tau|\pi] R_{M_1}(\tau)$ (and similarly for M_2). For τ such that $\text{escape}_K(\tau) = 1$, define $\text{pre}_K(\tau)$ as the prefix of τ where every state-action is in K except for the last one (which escapes). Similarly define $\text{suf}_K(\tau)$ as the remainder of the episode. Let $R(\text{pre}_K(\tau))$ be the sum of discounted rewards within the prefix (or suffix), and $\mathbb{P}_{M_1}[\text{pre}_K(\tau)|\pi]$ be the marginal probability of the prefix assigned by M_1 under policy π.

Below we upper bound $J_{M_1}(\pi) - J_{M_2}(\pi)$; the other direction (upper bounding $J_{M_2}(\pi) - J_{M_1}(\pi)$) is similar and hence omitted.

$$\begin{aligned}
J_{M_1}(\pi) &= \sum_{\tau:\text{escape}_K(\tau)=1} \mathbb{P}_{M_1}[\tau|\pi] \left(R_{M_1}(\text{pre}_K(\tau)) + R_{M_1}(\text{suf}_K(\tau))\right) + \sum_{\tau:\text{escape}_K(\tau)=0} \mathbb{P}_{M_1}[\tau|\pi] R_{M_1}(\tau) \\
&\le \sum_{\tau:\text{escape}_K(\tau)=1} \mathbb{P}_{M_1}[\tau|\pi] \left(R_{M_1}(\text{pre}_K(\tau)) + V_{\max}\right) + \sum_{\tau:\text{escape}_K(\tau)=0} \mathbb{P}_{M_1}[\tau|\pi] R_{M_1}(\tau) \\
&\le \sum_{\text{pre}_K(\tau)} \mathbb{P}_{M_1}[\text{pre}_K(\tau)|\pi] \left(R(\text{pre}_K(\tau)) + V_{\max}\right) + \sum_{\tau:\text{escape}_K(\tau)=0} \mathbb{P}_{M_1}[\tau|\pi] R_{M_1}(\tau).
\end{aligned}$$

The last step uses the fact that for any τ that shares the same $\text{pre}_K(\tau)$, we can combine their probabilities (because $R(\text{pre}_K(\tau)) + V_{\max}$ does not depends on the suffix), and we get the marginal probability of the prefix. We lower bound $v^{\pi}_{M_2}$ similarly by relaxing $R(\text{suf}_K(\tau))$ to 0, and obtain

$$J_{M_2}(\pi) \geq \sum_{\text{pre}_K(\tau)} \mathbb{P}_{M_2}[\text{pre}_K(\tau)|\pi]\, R_{M_2}(\text{pre}_K(\tau)) + \sum_{\tau:\text{escape}_K(\tau)=0} \mathbb{P}_{M_2}[\tau|\pi] R_{M_2}(\tau).$$

Observe that when $\text{escape}_K(\tau) = 0$, $\mathbb{P}_{M_1}[\tau|\pi] = \mathbb{P}_{M_2}[\tau|\pi]$ and $R_{M_1}(\tau) = R_{M_2}(\tau)$. This can be verified by expanding $\mathbb{P}_{M_1}$ for $\tau = s_1, a_1, \ldots, s_H, a_H$, which is

$$d_0(s_1)\pi(a_1|s_1)P_1(s_2|s_1,a_1)\pi(a_2|s_2)\ldots, P_1(s_H|s_{H-1},a_{H-1})\pi(a_H|s_H).$$

Since τ does not escape, $P_1(s_{h+1}|s_h,a_h) = P_2(s_{h+1}|s_h,a_h)$ by definition, so M_1 and M_2 assigns the same probability to τ. Similarly we have $\mathbb{P}_{M_1}[\text{pre}_K(\tau)|\pi] = \mathbb{P}_{M_2}[\text{pre}_K(\tau)|\pi]$. Subtracting the above two inequalities,

$$J_{M_1}(\pi) - J_{M_2}(\pi) \leq \sum_{\text{pre}_K(\tau)} \mathbb{P}_{M_1}[\text{pre}_K(\tau)|\pi]\, V_{\max}.$$

The result follows by noticing that the sum of probabilities here is simply $\mathbb{P}_{M_1}[\text{escape}_K(\tau)]$. □

Sample complexity of Rmax We show that in each episode either $\pi^{\star}_{\widehat{M}_K}$ is ϵ-optimal, or there is significant probability in increasing the counter of some $n(s,a)$ if we set m appropriately. Note that $\widehat{M}_K$ will be a good approximation of M_K for all states and actions: for $(s,a) \in K$ we can set m large enough so that the empirical estimate is accurate, and for $(s,a) \notin K$ the two MDPs agree exactly anyway. With this in mind, consider the situation when $\pi^{\star}_{\widehat{M}_K}$ is ϵ-suboptimal:

$$\begin{aligned}
\epsilon\, V_{\max} &< J^{\star}_M - J_M(\pi^{\star}_{\widehat{M}_K}) \leq J_{M_K}(\pi^{\star}) - J_M(\pi^{\star}_{\widehat{M}_K}) && \text{(optimism)}\\
&\leq J^{\star}_{M_K} - J_M(\pi^{\star}_{\widehat{M}_K})\\
&\leq J^{\star}_{\widehat{M}_K} + \text{distance}(M_K, \widehat{M}_K) \cdot \tfrac{V_{\max}}{2(1-\gamma)} - J_M(\pi^{\star}_{\widehat{M}_K}) && \text{(apply Lemma 2 on } M_K \text{ and } \widehat{M}_K)\\
&\leq J_{M_K}(\pi^{\star}_{\widehat{M}_K}) + \text{distance}(M_K, \widehat{M}_K) \cdot \tfrac{V_{\max}}{(1-\gamma)} - J_M(\pi^{\star}_{\widehat{M}_K}).\\
& && \text{(apply Lemma 3 on } M_K, \widehat{M}_K, \text{ and } \pi^{\star}_{\widehat{M}_K})\\
&\leq V_{\max}\, \mathbb{P}_M[\text{escape from } K \mid \pi^{\star}_{\widehat{M}_K}] + \text{distance}(M_K, \widehat{M}_K) \cdot \tfrac{V_{\max}}{(1-\gamma)}. && \text{(Lemma 4)}
\end{aligned}$$

We will later use Lemma 1 to guarantee that for all $(s,a) \in K$, $\|P(s,a) - \widehat{P}_K(s,a)\|_1 \leq \eta$ for some small η. Since M_K and $\widehat{M}_K$ are identical on $(s,a) \notin K$, we immediately have $\text{distance}(M_K, \widehat{M}_K) \leq \eta$.

In particular, we would like to set $\eta = \epsilon(1-\gamma)/2$, so that the escaping probability is at least $\epsilon/2$. This η can be guaranteed via Lemma 1 by setting $m = \tilde{O}(\frac{|\mathcal{S}|}{\epsilon^2(1-\gamma)^2} \log \frac{1}{\delta})$,[2] and "on average" we increase the counter by $\epsilon/2$ in each episode, so the number of episodes is

$$\tilde{O}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|}{\epsilon^3(1-\gamma)^2} \log \frac{1}{\delta}\right).$$

[2]There is a union bound over all states and actions here, which only incurs logarithmic dependence on $|\mathcal{S} \times \mathcal{A}|$ and is suppressed by $\tilde{O}$.

Note that the above treatment is not accurate as the growth of the counters is random, and we need to use concentration of measure to show that they will grow to m within the sample complexity bound with high probability. Note that we cannot use Hoeffding's here, as later random variables (which counter will grow, how much the growth is, etc.) highly depend on earlier variables. For a relatively rigorous treatment (which involves Azuma's inequality for martingales), see [4, Appendix E].

References

[1] Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *The Journal of Machine Learning Research*, 3:213–231, 2003.

[2] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.

[3] Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University of College London, 2003.

[4] Nan Jiang. PAC reinforcement learning with an imperfect model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 3334–3341, 2018.