Reinforcement Learning in Finite MDPs: PAC Analysis

A.L. Strehl, L. Li, and M. L. Littman

Presented by: Hamid R. Chinaei

DAMAS Lab, Computer Science and Software Engineering, Laval University

February 29^{th}, 2009

DAMAS
Décision, Adaptation, Multi-AgentS
www.damas.ift.ulaval.ca

UNIVERSITÉ LAVAL

Overview

- Problem
- Framework
- Sample complexity definitions
- R-max algorithm
- Sample complexity of R-max
- Proof
- Discussions

Problem

- RL PAC Analysis
- Intuitively, number of iterations needed, so that the resulted value function is close enough to optimal value with a high probability
- close enough, accuracy, ϵ
- high probability, confidence, $1 - \delta$

$$M :< \mathcal{S}, A, T, R, \gamma >$$

Transition function:

$$\mathcal{T} : S \times A \rightarrow P_{\mathcal{S}}$$

Reward function:

$$\mathcal{R} : S \times A \rightarrow P_{\mathbb{R}}$$

Discount factor:

$$0 \leq \gamma < 1$$

Policy:

$$\pi_t : \{S \times A \times [0, 1]\} \times S \rightarrow A$$

State value of policy for infinite horizon:

$$V_M^{\pi}(s) = \mathbf{E}[\sum_{j=1}^{\infty} \gamma^{j-1} r_j | s]$$

State, action value of policy for infinite horizon:

$$Q_M^{\pi}(s, a) = \mathbf{E}[\sum_{j=1}^{\infty} \gamma^{j-1} r_j | s, a]$$

$$V_M^\pi(c_t) = \mathbf{E}[\sum_{j=1}^{\infty} \gamma^{j-1} r_j | c_t]$$

where $c_t = (s_1, a_1, r_1, \ldots, s_t)$

$$V_M^\pi(c_t, H) = \mathbf{E}[\sum_{j=0}^{H-1} \gamma^j r_{t+j} | c_t]$$

Admissible Heuristic: $U : \mathcal{S} \times A \to \mathbb{R}$

$$\forall s \in \mathcal{S}, a \in \mathcal{A}\ Q^*(s,a) \leq U(s,a) \leq V_{max}$$

$$V_M^\pi = \max_{a \in A} Q_M^\pi(s,a)$$

$$\forall (s,a) \in \mathcal{S} \times A\ 0 \leq U(s,a) \leq V_{max} \leq 1/(1-\gamma)$$

Definition

Let $c = (s_1, a_1, r_1, s_2, a_2, r_2, \ldots)$ be a random path generated by executing an algorithm $\mathcal{A}$ in an MDP M. For any $\epsilon \geq 0$, the sample complexity of exploration, or sample complexity, of $\mathcal{A}$ is the number of timesteps t such that the policy at time t, $\mathcal{A}_t$ satisfies $V^{\mathcal{A}_t} \leq V^*(s_t) - \epsilon$.

Proposition

Let $\beta > 0$ be any real number satisfying $\beta \leq 1/(1-\gamma)$. Suppose that value iteration runs for $\lceil \frac{\ln(1/(\beta(1-\gamma)))}{1-\gamma} \rceil$ iterations, where each initial action value estimate is initiate to some value between 0 and $1/(1-\gamma)$.
Let $Q'(.,.)$ be the resulting action-value estimates. Then we have:

$$\max |Q'(s,a) - Q^*(s,a)| \leq \beta$$

R-max Algorithm, initialization

0: **Inputs:** S, A, γ, m, ε_1, and $U(\cdot,\cdot)$
1: **for all** (s,a) **do**
2: $\quad Q(s,a) \leftarrow U(s,a)$ // *action-value estimates*
3: $\quad r(s,a) \leftarrow 0$
4: $\quad n(s,a) \leftarrow 0$
5: $\quad$ **for all** $s' \in S$ **do**
6: $\quad\quad n(s,a,s') \leftarrow 0$
7: $\quad$ **end for**
8: **end for**

for $t = 1, 2, 3, \cdots$ **do**
  Let s denote the state at time t.
  Choose action $a := \text{argmax}_{a' \in A} Q(s, a')$.
  Let r be the immediate reward and s' the next state after executing action a from state s.
  if $n(s,a) < m$ **then**
    $n(s,a) \leftarrow n(s,a) + 1$
    $r(s,a) \leftarrow r(s,a) + r$ // *Record immediate reward*
    $n(s,a,s') \leftarrow n(s,a,s') + 1$ // *Record immediate next-state*
    if $n(s,a) = m$ **then**
      for $i = 1, 2, 3, \cdots, \left\lceil \frac{\ln(1/(\varepsilon_1(1-\gamma)))}{1-\gamma} \right\rceil$ **do**
        for all $(\bar{s}, \bar{a})$ **do**
          if $n(\bar{s}, \bar{a}) \geq m$ **then**
            $Q(\bar{s}, \bar{a}) \leftarrow \hat{R}(\bar{s}, \bar{a}) + \gamma \sum_{s'} \hat{T}(s' | \bar{s}, \bar{a}) \max_{a'} Q(s', a')$.
          end if
        end for
      end for
    end if
  end if
end for

Known state-action MDP

$$M =< \mathcal{S}, A, T, R, \gamma >, \ Q(s,a)$$

- K: Known state-action
- Known state-action MDP: $M_K =< S \cup \{z_{s,a} | (s,a) \notin K\}, A, T_K, R_K, \gamma >$
- $\forall (s,a) \notin K$, add a new state $z_{s,a}$ with self-loops. $T(z_{s,a}|z_{s,a},.) = 1$
- $\forall (s,a) \in K \ R_K(s,a) = R(s,a) \ T_K(.|s,a) = T(.|s,a)$
- $\forall (s,a) \notin K, R_K(s,a) = Q(s,a)(1-\gamma) \ T_K(z_{s,a}|s,a) = 1$
- $\forall z_{s,a} \ R_K(z_{s,a},.) = Q(s,a)(1-\gamma)$

Theorem 10 *Let $\mathcal{A}(\varepsilon,\delta)$ be any greedy learning algorithm such that, for every timestep t, there exists a set K_t of state-action pairs that depends only on the agent's history up to timestep t. We assume that $K_t = K_{t+1}$ unless, during timestep t, an update to some state-action value occurs or the escape event A_K happens. Let M_{K_t} be the known state-action MDP and π_t be the current greedy policy, that is, for all states s, $\pi_t(s) = \text{argmax}_a Q_t(s,a)$. Furthermore, assume $Q_t(s,a) \leq V_{\max}$ for all t and (s,a). Suppose that for any inputs ε and δ, with probability at least $1-\delta$, the following conditions hold for all states s, actions a, and timesteps t: (1) $V_t(s) \geq V^*(s) - \varepsilon$ (optimism), (2) $V_t(s) - V^{\pi_t}_{M_{K_t}}(s) \leq \varepsilon$ (accuracy), and (3) the total number of updates of action-value estimates plus the number of times the escape event from K_t, A_K, can occur is bounded by $\zeta(\varepsilon,\delta)$ (learning complexity). Then, when $\mathcal{A}(\varepsilon,\delta)$ is executed on any MDP M, it will follow a 4ε-optimal policy from its current state on all but*

$$O\left(\frac{V_{\max}\zeta(\varepsilon,\delta)}{\varepsilon(1-\gamma)}\ln\frac{1}{\delta}\ln\frac{1}{\varepsilon(1-\gamma)}\right)$$

timesteps, with probability at least $1-2\delta$.

Theorem 11 *Suppose that* $0 \leq \varepsilon < \frac{1}{1-\gamma}$ *and* $0 \leq \delta < 1$ *are two real numbers and* $M = \langle \mathcal{S}, \mathsf{A}, T, \mathcal{R}, \gamma \rangle$ *is any MDP. There exists inputs* $m = m(\frac{1}{\varepsilon}, \frac{1}{\delta})$ *and* ε_1*, satisfying* $m(\frac{1}{\varepsilon}, \frac{1}{\delta}) = O\left(\frac{(S+\ln(SA/\delta))V_{\max}^2}{\varepsilon^2(1-\gamma)^2}\right)$ *and* $\frac{1}{\varepsilon_1} = O(\frac{1}{\varepsilon})$*, such that if R-MAX is executed on M with inputs m and* ε_1*, then the following holds. Let* $\mathcal{A}_t$ *denote R-MAX's policy at time t and* s_t *denote the state at time t. With probability at least* $1-\delta$*,* $V_M^{\mathcal{A}_t}(s_t) \geq V_M^*(s_t) - \varepsilon$ *is true for all but*

$$O\left(\frac{|\{(s,a) \in \mathcal{S} \times \mathsf{A} | U(s,a) \geq V^*(s) - \varepsilon\}|}{\varepsilon^3(1-\gamma)^3}\left(S + \ln\frac{SA}{\delta}\right)V_{\max}^3 \ln\frac{1}{\delta}\ln\frac{1}{\varepsilon(1-\gamma)}\right)$$

timesteps t.

Approximating Q by approximating dynamics

Lemma 12 *(Strehl and Littman, 2005) Let $M_1 = \langle \mathcal{S}, \mathrm{A}, T_1, R_1, \gamma \rangle$ and $M_2 = \langle \mathcal{S}, \mathrm{A}, T_2, R_2, \gamma \rangle$ be two MDPs with non-negative rewards bounded by 1 and optimal value functions bounded by $V_{\max}$. Suppose that $|R_1(s,a) - R_2(s,a)| \leq \alpha$ and $\|T_1(s,a,\cdot) - T_2(s,a,\cdot)\|_1 \leq 2\beta$ for all states s and actions a. There exists a constant $C > 0$ such that for any $0 \leq \varepsilon \leq 1/(1-\gamma)$ and stationary policy π, if $\alpha = 2\beta = C\varepsilon(1-\gamma)/V_{\max}$, then*

$$|Q_1^\pi(s,a) - Q_2^\pi(s,a)| \leq \varepsilon.$$

Reward approximation

Lemma 13 *Suppose that $r[1], r[2], \ldots, r[m]$ are m rewards drawn independently from the reward distribution, $\mathcal{R}(s,a)$, for state-action pair (s,a). Let $\hat{R}(s,a)$ be the empirical (maximum-likelihood) estimate of $\mathcal{R}(s,a)$. Let δ_R be any positive real number less than 1. Then, with probability at least $1-\delta_R$, we have that $|\hat{R}(s,a) - \mathcal{R}(s,a)| \leq \varepsilon^R_{n(s,a)}$, where*

$$\varepsilon^R_m := \sqrt{\frac{\ln(2/\delta_R)}{2m}}.$$

Transition approximation

Lemma 14 *Suppose that $\hat{T}(s,a)$ is the empirical transition distribution for state-action pair (s,a) using m samples of next states drawn independently from the true transition distribution $T(s,a)$. Let δ_T be any positive real number less than 1. Then, with probability at least $1-\delta_T$, we have that $\|T(s,a)-\hat{T}(s,a)\|_1 \leq \varepsilon^T_{n(s,a)}$ where*

$$\varepsilon^T_m = \sqrt{\frac{2[\ln(2^S-2)-\ln(\delta_T)]}{m}}.$$

Event A1

For all stationary policies π, timesteps t and states s during execution of the R-max algorithm on some MDP M:

$$|V^{\pi}_{M_{k_t}} - V^{\pi}_{\hat{M}_{k_t}}| \leq \epsilon 1$$

Theorem 11 *Suppose that* $0 \leq \varepsilon < \frac{1}{1-\gamma}$ *and* $0 \leq \delta < 1$ *are two real numbers and* $M = \langle \mathcal{S}, \mathrm{A}, T, \mathcal{R}, \gamma \rangle$ *is any MDP. There exists inputs* $m = m(\frac{1}{\varepsilon}, \frac{1}{\delta})$ *and* ε_1*, satisfying* $m(\frac{1}{\varepsilon}, \frac{1}{\delta}) = O\left(\frac{(S+\ln(SA/\delta))V_{\max}^2}{\varepsilon^2(1-\gamma)^2}\right)$ *and* $\frac{1}{\varepsilon_1} = O(\frac{1}{\varepsilon})$*, such that if R-MAX is executed on M with inputs m and* ε_1*, then the following holds. Let* $\mathcal{A}_t$ *denote R-MAX's policy at time t and* s_t *denote the state at time t. With probability at least* $1-\delta$*,* $V_M^{\mathcal{A}_t}(s_t) \geq V_M^*(s_t) - \varepsilon$ *is true for all but*

$$O\left(\frac{|\{(s,a) \in \mathcal{S} \times \mathrm{A} | U(s,a) \geq V^*(s) - \varepsilon\}|}{\varepsilon^3(1-\gamma)^3}\left(S + \ln\frac{SA}{\delta}\right)V_{\max}^3 \ln\frac{1}{\delta}\ln\frac{1}{\varepsilon(1-\gamma)}\right)$$

timesteps t.

$$V_t(s) \geq V^*_{\hat{M}_{K_t}}(s) - \epsilon 1 \geq V^*_{M_{K_t}}(s) - 2\epsilon 1 \geq V^*(s) - 2\epsilon 1$$

Proof

- Condition 1

$$V_t(s) \geq V^*_{\hat{M}_{K_t}}(s) - \epsilon 1 \geq V^*_{M_{K_t}}(s) - 2\epsilon 1 \geq V^*(s) - 2\epsilon 1$$

- Condition 2 follows from Event A1
- Condition 3, learning complexity $\zeta(\epsilon, \delta) \leq |\{(s,a)|U(s,a) \geq V^*(s) - \epsilon\}|m$
  - (s,a) such that $U(s,a) < V^*(s) - \epsilon$ will never be experienced, with high probability
  - There is always actions a' $s.t.$ $Q_t(s,a') > V^*(s) - \epsilon$
  - When escape occurs, some $(s,a) \notin K$ is experienced
  - When (s,a) experience m times, it becomes part of, and never leaves set K
  - To guaranty that Event A1 occurs with probability at leas $1 - \delta$, Lemma 15 is used

Discussions

- R-max sample Complexity

$$\tilde{O}((S^2A)/(\epsilon^3(1-\gamma)^6))$$

$$\tilde{O}(\frac{V_{max}^3 S|(s,a) \in S \times A|U(s,a) \geq V^*(s) - \epsilon||}{\epsilon^3(1-\gamma^3)})$$

- Should we care about sample complexity in RL
- How choose an RL algorithm for some application
- Can we have more intuitive ways of comparing RL algorithms

Thanks!

Questions?