RL Theory[1]: Lecture 2 (Chapter 1)

Shivam Garg, RLAI@UAlberta

8th September 2020

[1] based on https://rltheorybook.github.io/

Adminstrivia (to set the mood right again)

- My goal with these lectures is that **all of you** understand most of the material.
- I will go slowly and try to be shameless about it. However, I have this problem where I start picking pace without realizing.
- Do stop me at any time if something is unclear (be it my accent, a mistake, some skipped steps).
- Especially today! I think today would be a little dense :D

Bellman Optimality Operator $\mathcal{T}$

- $\mathcal{T} : \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}\times\mathcal{A}|}$

$$V_Q(s) = \max_a Q(s,a)$$

$$\mathcal{T}Q := (1-\gamma)r + \gamma P V_Q.$$

$$\mathcal{T}Q(s,a) = (1-\gamma)\, r(s,a) + \gamma \sum_{s'} p(s'|s,a) \max_a Q(s',a)$$

- Therefore, the $Q^* \to V^*$ equation can be written as

$$Q^*(s,a) = (1-\gamma)r(s,a) + \gamma\,\mathbb{E}_{s'\sim P(\cdot|s,a)}[V^*(s')]$$
$$\Rightarrow \qquad Q^* = \mathcal{T}Q^*.$$

Q–Value Iteration

- Q–value iteration:

$$\boxed{Q^{(k+1)} = \mathcal{T} Q^{(k)}.}$$

$$Q^{(0)} \xrightarrow{\mathcal{T}} Q^{(1)} \xrightarrow{\mathcal{T}} Q^{(2)} \cdots \rightsquigarrow Q^{*}$$

- What are iterative equations?
- E.g.: $x_{k+1} = 4x_k + 3$. (divergent for $x_0 = 0$) $0, 3, 15, \cdots$ (divergent)
- E.g.: $\boxed{y_{k+1} = 0.1y_k + 1}$ (convergent for $y_0 = 0$) $1, 1.1, 1.11, \ldots$

$\to 1.11111$

Q–Value Iteration Converges!

Theorem (Q–Value Iteration Convergence)

- *Set* $\boxed{Q^{(0)} = 0.}$
- *Obtain* $Q^{(k+1)} = \mathcal{T}Q^{(k)}$ *for* $k = 0, 1, 2, \ldots$.
- *Let* $\pi^{(k)} = \pi_{Q^{(k)}}$.
- *Then for* $k \geq \frac{1}{1-\gamma}\log\left(\frac{2}{\epsilon(1-\gamma)}\right)$,

$$V^{\pi^{(k)}} \geq V^* - \epsilon\mathbb{1}.$$

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$Q(s, a_1) = 0$

$Q(s_1, a_2) = 0$

$\vdots$

$\pi^{(k)}(s) = \arg\max_a Q^{(k)}(s,a)$

$V^* - V^{\pi^{(k)}} \leq \epsilon \mathbb{1}$

Proof: Q–Value Iteration Convergence

- To prove: For $k > \frac{1}{1-\gamma} \log\left(\frac{2}{\epsilon(1-\gamma)}\right)$, $V^{\pi^{(k)}} \geq V^* - \epsilon \mathbb{1}$ holds.

- Bellman Optimality Operator $\mathcal{T}$ is a Contraction

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \leq \gamma \|Q - Q'\|_\infty.$$

- Bounds on $\|Q^{(k)} - Q^*\|_\infty$

$$\|Q^{(k)} - Q^*\|_\infty \leq e^{-(1-\gamma)k}.$$

- Bounding the Suboptimality of π_Q (Singh & Yee, 1994)

$$V^{\pi_Q} \geq V^* - \frac{2}{1-\gamma}\|Q - Q^*\|_\infty \mathbb{1}.$$

(Part 1): Bellman Optimality Operator $\mathcal{T}$ is a Contraction

$$Q' = \begin{bmatrix} s_1, a_1 \\ s_1, a_2 \\ s_1, a_3 \end{bmatrix} \quad \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

$\max_{a'} Q'(s, a') = 3$

$Q'(s, a^*) = 1$

$\| P x \|_\infty \le \| x \|_\infty$

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty \le \gamma \|Q - Q'\|_\infty$$

$\begin{bmatrix} \quad \end{bmatrix}_P \begin{bmatrix} \quad \end{bmatrix}_x = \begin{bmatrix} \quad \end{bmatrix}$ $\sum p_i x_i \le \sum p_i \max x_i = \max x_i$

$$\|\mathcal{T}Q - \mathcal{T}Q'\|_\infty = \| \cancel{(1-\gamma)r} + \gamma P V_Q - \cancel{(1-\gamma)r} - \gamma P V_{Q'} \|_\infty$$

$$= \gamma \| P (V_Q - V_{Q'}) \|_\infty$$

$$\le \gamma \| V_Q - V_{Q'} \|_\infty$$

$$= \gamma \max_s | V_Q(s) - V_{Q'}(s) |$$

$$= \gamma \max_s \max_a | Q(s,a) - Q'(s,a) |$$

$$= \gamma \| Q - Q' \|_\infty$$

$\max_{a'} Q'(s,a') \ge Q'(s,a^*)$

$-\max_{a'} Q'(s,a') \le -Q'(s,a^*)$

(Part 2): Bounds on $\|Q^{(k)} - Q^*\|_\infty$

$Q^{(k+1)} = TQ^{(k)}$

$\boxed{TQ^* = Q^*}$

$$\|Q^{(k)} - Q^*\|_\infty \le e^{-(1-\gamma)k}.$$

$|x| \le 1$

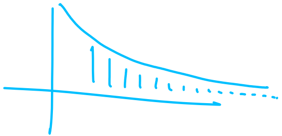
$(1+x)^k = 1 + kx + \frac{k(k-1)}{2!}x^2 + \frac{k(k-1)(k-2)}{3!}x^3 + \dots$

$\le$

$e^{xk} = 1 + kx + \frac{k^2}{2!}x^2 + \frac{k^3}{3!}x^3 + \dots$

$$
\begin{aligned}
\|Q^{(k)} - Q^*\|_\infty &= \|TQ^{(k-1)} - TQ^*\|_\infty \\
&\le \gamma \|Q^{(k-1)} - Q^*\|_\infty \\
&= \gamma \|TQ^{(k-2)} - TQ^*\|_\infty \\
&\le \gamma^2 \|Q^{(k-2)} - Q^*\|_\infty \\
&\;\vdots \\
&\le \gamma^k \|Q^{(0)} - Q^*\|_\infty = \gamma^k \underbrace{\|Q^*\|_\infty}_{\le 1}
\end{aligned}
$$

$\le \gamma^k = (1-(1-\gamma))^k \le e^{-(1-\gamma)k}$



(Part 3): Bounding the Suboptimality of π_Q (Singh & Yee, 1994)

Deterministic policies

$$V^{\pi_Q} \geq V^* - \frac{2}{1-\gamma}\|Q - Q^*\|_\infty \mathbb{1}.$$

$\|V^* - V^{\pi_Q}\|_\infty \leq \frac{2}{1-\gamma}\|Q - Q^*\|_\infty$

$V^*(s) - V^{\pi_Q}(s) \leq \frac{2}{1-\gamma}\|Q - Q^*\|_\infty$

≤ 3

$$V^*(s) - V^{\pi_Q}(s) = Q^*(s, \pi^*(s)) - Q(s, a)$$

$\max \cdot \leq 3$

$$= Q^*(s, \pi^*(s)) - Q^*(s,a) + Q^*(s,a) - Q(s,a)$$

$$= Q^*(s, \pi^*(s)) - Q^*(s,a) + \gamma \sum_{s'} p(s'|s,a)\left[V^*(s) - V^{\pi_Q}(s)\right]$$

$$\leq Q^*(s, \pi^*(s)) \boxed{- Q(s, \pi^*(s)) + Q(s,a)} - Q^*(s,a)$$

$$+ \gamma \sum_{s'} p(s'|s,a)\left[V^*(s) - V^{\pi_Q}(s)\right]$$

$\Rightarrow V^*(s) - V^{\pi_Q}(s) \leq 2\|Q^* - Q\|_\infty + \gamma\|V^* - V^{\pi_Q}\|_\infty$

$\|V^* - V^{\pi_Q}\|_\infty \leq 2\|Q^* - Q\|_\infty + \gamma\|V^* - V^{\pi_Q}\|_\infty$.

$A(s) - B(s)$
$\leq \|A - B\|_\infty$

(Final Part): Proof of Q–Value Iteration Convergence

$-\log(x) = \log\left(\frac{1}{x}\right)$

$$V^{\pi_Q} \geq V^* + \left(\frac{2}{1-\gamma}\right)\left(-\|Q - Q^*\|_\infty\right)$$

$$-\|Q^{(k)} - Q^*\|_\infty \geq -e^{-(1-\gamma)k}$$

$$V^{\pi^{(k)}} \geq V^* - \frac{2}{1-\gamma} e^{-(1-\gamma)k}$$

$$\boxed{V^* - V^{\pi^{(k)}} \leq \frac{2}{1-\gamma} e^{-(1-\gamma)k} \leq \epsilon}$$

$$\frac{2}{1-\gamma} e^{-(1-\gamma)k} \leq \epsilon \Rightarrow e^{-(1-\gamma)k} \leq \frac{(1-\gamma)\epsilon}{2} \Rightarrow k \geq \frac{1}{1-\gamma}\left(-\log\left(\frac{(1-\gamma)\epsilon}{2}\right)\right)$$

$$\Leftrightarrow k \geq \frac{1}{1-\gamma} \log \frac{2}{(1-\gamma)\epsilon}$$

Q–Value Iteration Convergence: But what the @#$% does it mean?

- $V^{\pi(k)} \geq V^* - \epsilon \mathbb{1}$ for $k \geq \frac{1}{1-\gamma} \log\left(\frac{2}{\epsilon(1-\gamma)}\right)$.

$\gamma = 0.99$

$\log(ab) = \log a + \log b$

$\Rightarrow B\left(A + \log\left(\frac{1}{\epsilon}\right)\right)$

- This rate is pretty good! Look at this:

Precision	ϵ	$\#(iter)$
1	0.1	1.5×10^6
2	0.01	1.7×10^6
3	0.001	1.9×10^6
4	0.0001	2.1×10^6
5		2.3×10^6

0.7777

0.7

- This rate is effectively **linear** in the precision of the value function!
- Think of precision as being $\mathbf{log}\,(\mathbf{1}/\epsilon)$.

GD: $O\left(\frac{1}{\epsilon}\right) \equiv O(e^P)$ (exp)

r $\quad$ P

- But this is NOT the sample based update. In practice, we know neither of r and P. The rate for the sample based update would be much worse.

Playing with the Rooftop MDP

π

- Example calculation from last time:

$$Q^\pi = (1-\gamma)(I - \gamma P^\pi)^{-1} r$$

$$= (1-0.9)\left(\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix} - 0.9\begin{bmatrix}1 & 0 & 0\\ 0.5 & 0 & 0.5\\ 1 & 0 & 0\end{bmatrix}\right)^{-1}\begin{bmatrix}0.1\\ 0.5\\ 0.2\end{bmatrix}$$

$$= \begin{bmatrix}Q^\pi(s_1, a_1)\\ Q^\pi(s_1, a_2)\\ Q^\pi(s_2, a_1)\end{bmatrix} = \begin{bmatrix}0.1\\ 0.1445\\ 0.11\end{bmatrix}$$

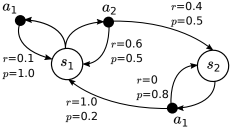
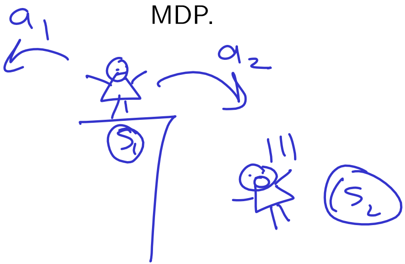


Figure: **(Rooftop)** MDP.

$\pi(a_1 | s_1)$

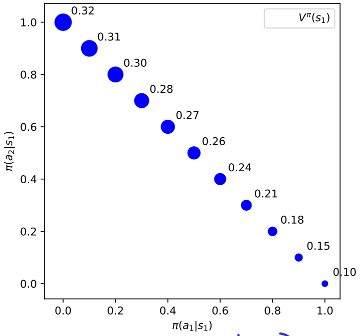
$\pi(a_2 | s_1)$



- Define $\pi_Q(s) := \arg\max_{a\in\mathcal{A}} Q(s,a)$.
- Define $V_Q(s) := \max_{a\in\mathcal{A}} Q(s,a)$.

Playing with the Rooftop MDP: Value Iteration

$\pi(a_2|s_1)$ $\to a_2$ $\|Q^{(k)} - Q^*\|_\infty \le e^{-(1-\gamma)k}$



$\pi(a_1|s_1)$

Figure: V^π

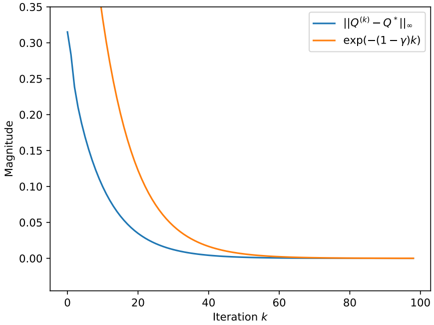


Figure: Q–value iteration.

$$Q^* = \begin{bmatrix} 0.29 & 0.32 & 0.27 \end{bmatrix}^\top.$$

Break for Questions (and Bonus: A joke!)

- **Me:** Jeez. I'm pretty bad at all this theory! It's hard.

Break for Questions (and Bonus: A joke!)

- **Me:** Jeez. I'm pretty bad at all this theory! It's hard.
- **Fancy Anime Girl:**

Break for Questions (and Bonus: A joke!)

- **Me:** Jeez. I'm pretty bad at all this theory! It's hard.
- **Fancy Anime Girl:**

- **Me:** Oooh! Thanks. You really think so?

Break for Questions (and Bonus: A joke!)

- **Me:** Jeez. I'm pretty bad at all this theory! It's hard.
- **Fancy Anime Girl:**

- **Me:** Oooh! Thanks. You really think so?
- **Fancy Anime Girl:**

- **Me:** -_-

Note about Bellman Operator $\mathcal{T}^{\pi}$, Bellman Optimality Operator $\mathcal{T}$, and 3 types of P matrices

Policy Iteration

- Begin with a policy $\pi^{(0)}$
- Policy Evaluation: Find $Q^{\pi^{(k)}}$; say, by using $Q^{\pi} = (1-\gamma)(I - \gamma P^{\pi^{(k)}})^{-1} r$.
- Policy Improvement: Calculate $\pi^{(k+1)} = \pi_{Q^{\pi^{(k)}}}$.

- **Convergence:** For $k \geq \frac{1}{1-\gamma} \log \left(1/\epsilon\right)$

$$Q^{\pi_k} \geq Q^* - \epsilon.$$

- I think we need to combine this with the (Singh & Yee, 1994) equation.

(Part 1) Policy Iteration: Convergence Proof

- $Q^{\pi_{k+1}} \geq \mathcal{T} Q^{\pi_k} \geq Q^{\pi_k}$.

(Part 2) Policy Iteration: Convergence Proof

- $Q^{\pi_{k+1}} \geq \mathcal{T} Q^{\pi_k} \geq Q^{\pi_k}$.

(Part 3) Policy Iteration: Convergence Proof

- $\|Q^{\pi_{k+1}} - Q^*\|_\infty \leq \gamma \|Q^{\pi_k} - Q^*\|_\infty$.

(Final Part) Policy Iteration: Convergence Proof

-

- How would we show that Generalized Policy Iteration converges?

Complete Space of Policies

- We define the policy as $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. This is a stationary markov policy.
- A deterministic stationary markov policy would be defined as $\pi : \mathcal{S} \to \mathcal{A}$.
- A general policy (possible non–deterministic, non–stationary, and non–markov) would be defined as $\pi : \mathcal{H} \to \Delta(\mathcal{A})$, where $\mathcal{H}$ is the set of all trajectories.
- Is is even possible to assume the optimal policy to be stationary and deterministic?

(Part 1) Existence of a Stationary and Deterministic Optimal Policy

- Let Π be the set of all non–stationary and randomized policies. There exists a stationary and deterministic policy π such that for all states $s \in \mathcal{S}$,

$$V^{\pi}(s) = \max_{\pi' \in \Pi} V^{\pi'}(s).$$

- Note: I did not understand what randomized means.

(Part 2) Existence of a Stationary and Deterministic Optimal Policy

-

(Part 1) Fixed Point of $Q = \mathcal{T} Q$

- Define optimal Q^* by $Q^*(s, a) = \max_{\pi \in \Pi} Q^\pi(s, a)$. Then $Q = Q^*$ if and only if it satisfies
$$Q = \mathcal{T} Q.$$

(Part 2) Fixed Point of $Q = \mathcal{T}Q$

Summary

- Value iteration
- Policy iteration
- Bellman Operator and Bellman Optimality Operator
- 3 types of transition matrices
- Existence of a stationary and deterministic optimal policy
- Fixed point of $Q = \mathcal{T}Q$
- Did NOT cover the LP formulation (next time)