MDP

PE

$|S| < \infty \quad |A| < \infty$

RL Theory[1]: Lecture 3 (Chapter 1)

π

$s_0 \; A_0 \; R_1 \; S_1 \; \ldots$

Shivam Garg, RLAI@UAlberta

$V^{\pi}(s)$

find V for π

23rd September 2020

PI

$\pi^{(0)}$

$Q^{\pi^{(0)}}$

$\pi^{(1)}(s) = \max_{a'} Q^{\pi^{(0)}}(s, a')$

$Q^{\pi^{(1)}} \ldots$

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

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

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

$\Rightarrow k \geq O\left(\log\left(1/\epsilon\right)\right)$

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

$k \geq \frac{1}{(1-\gamma)}\left[\log\left(\frac{2}{(1-\gamma)}\right) + \log\left(\frac{1}{\epsilon}\right)\right]$

- This rate is pretty good! Look at this:

Precision	ϵ	$\#(iter)$ $= k$
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

$\leq O\left(\log\left(\frac{1}{\epsilon}\right)\right)$

$1 = 0.1$

$\log(1/\epsilon) = \log(10) = 1$

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

$\log(1/\epsilon)$

- 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

s_1 — a_1, a_2

s_2 — a_1

- Example calculation from last time:

$$
\begin{aligned}
Q^\pi &= (1-\gamma)(I-\gamma P^\pi)^{-1} r \\
&= (1-0.9)\left(\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix} - 0.9\begin{bmatrix}1 & 0 & 0\\ 0.5 & 0 & 0.5\\ 1 & 0 & 0\end{bmatrix}\right)^{-1}\begin{bmatrix}0.1\\ 0.5\\ 0.2\end{bmatrix} \\
&= \begin{bmatrix}Q^\pi(s_1,a_1)\\ Q^\pi(s_1,a_2)\\ Q^\pi(s_2,a_1)\end{bmatrix} = \begin{bmatrix}0.1\\ 0.1445\\ 0.11\end{bmatrix}
\end{aligned}
$$

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

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

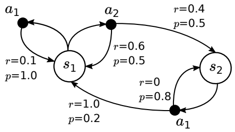


Figure: (**Rooftop**) MDP.

Playing with the Rooftop MDP: Value Iteration

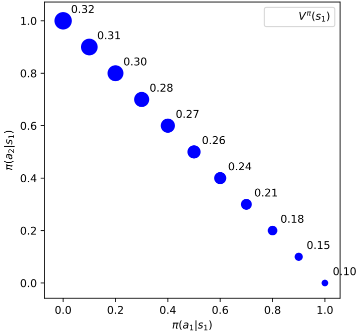


Figure: V^π

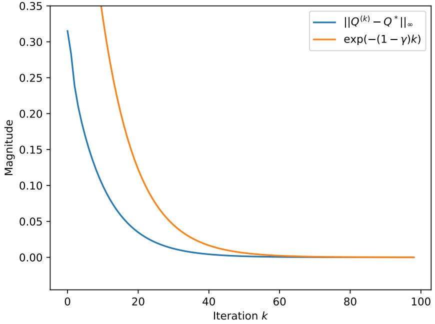


Figure: Q–value iteration.

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

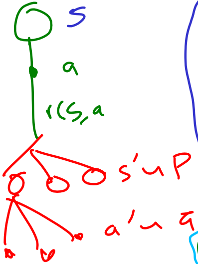
Note about Bellman Operator $\mathcal{T}^{\pi}$, Bellman Optimality Operator $\mathcal{T}$, and 3 types of P matrices

symbolic

$$Q^{\pi} = r + \gamma P^{\pi} Q^{\pi}$$

→ linear

Q^{π}



$$\begin{bmatrix} Q^{\pi}(s,a_1) \\ \\ \\ \end{bmatrix} = \begin{bmatrix} r(s_1,a_1) \\ \\ \\ \end{bmatrix} + \gamma \begin{bmatrix} p(s,a_1|s,a_1) & p(s_1 a_2|s a_1) & p(s_2 a_1|s_1 a_1) \\ & \pi \to P^{\pi} & \\ \\ \end{bmatrix} \begin{bmatrix} Q^{\pi}(s,a_1) \\ Q^{\pi}(s,a_2) \\ Q^{\pi}(s_2 a_1) \end{bmatrix}$$

matrix

$$Q^{\pi}(s,a) = r(s,a) + \gamma \sum_{s'} \sum_{a'} P_{\pi}(s',a'|s,a) Q^{\pi}(s',a')$$

P^{π} r

$$Q^{\pi} - \gamma P^{\pi} Q^{\pi} = r$$

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

Q^{π}

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

Policy Iteration

def. comp.

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

$\pi^{(0)} \to Q^{\pi^{(0)}} \to \pi^{(1)} \to Q^{\pi^{(1)}} \to \cdots \cdots \pi^*$ → prediction

? $Q^{\pi(k)} \to Q^*$

$\pi^{(k+1)}(s) = \arg\max_a Q^{\pi^{(k)}}(s,a)$

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

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

$\to Q^* - Q^{\pi_k} \leq \epsilon$

$k \geq \frac{1}{1-\gamma} \log\left(\frac{1}{\epsilon(1-\gamma)}\right)$

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

(Part 1) Policy Iteration: Convergence Proof

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

$$\boxed{\mathcal{T} Q = r + \gamma P V_Q}$$

$$V_Q = \begin{bmatrix} V_Q(s_1) \\ V_Q(s_2) \\ \vdots \end{bmatrix}$$

$$\boxed{V_Q(s) = \max_a Q(s,a)}$$

$$\mathcal{T} Q^{\pi_k}(s,a) = r(s,a) + \sum_{s'} p(s'|s,a) \left[\max_{a'} Q^{\pi_k}(s',a') \right]$$

$$\geq r(s,a) + \sum_{s'} p(s'|s,a) \left[\sum_{a'} \pi_k(a'|s') \, Q^{\pi_k}(s',a') \right]$$

$$= Q^{\pi_k}(s,a)$$

(Part 2) Policy Iteration: Convergence Proof

$\boxed{Q^{\pi_{k+1}} \geq Q^{\pi_k}}$

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

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

$$Q^{\pi_{k+1}}(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) \sum_{a'} \pi_{k+1}(a'|s') Q^{\pi_{k+1}}(s',a')$$

$$\geq r(s,a) + \gamma \sum_{s'} p(s'|s,a) \underbrace{\sum_{a'} \pi_{k+1}(a'|s') Q^{\pi_k}(s',a')}_{\max_{a'} Q^{\pi_k}(s',a')}$$

$\pi_{k+1}(s) = \arg\max_a Q^{\pi_k}(s,a)$

$$= r(s,a) + \gamma \sum_{s'} p(s'|s,a) \max_{a'} Q^{\pi_k}(s',a')$$

$$= \mathcal{T} Q^{\pi_k}(s,a)$$

(Part 3) Policy Iteration: Convergence Proof

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

$$-Q^{\pi_{k+1}} \le -TQ^{\pi_k} \qquad \Big\| \qquad Q^* = TQ^*$$

$$\begin{aligned}
\|Q^* - Q^{\pi_{k+1}}\|_\infty &\le \|Q^* - TQ^{\pi_k}\|_\infty \\
&= \|TQ^* - TQ^{\pi_k}\|_\infty \\
&\le \gamma \underbrace{\|Q^* - Q^{\pi_k}\|_\infty}
\end{aligned}$$

(contraction of T)

$$\begin{aligned}
&\le \gamma \left(\gamma \|Q^* - Q^{\pi_{k-1}}\|_\infty \right) \\
&= \gamma^2 \|Q^* - Q^{\pi_{k-1}}\|_\infty \\
&\;\vdots \\
&\le \gamma^{k+1} \|Q^* - Q^{\pi(0)}\|_\infty
\end{aligned}$$

$$\begin{aligned}
&\|Q^* - Q^{\pi_{k+1}}\|_\infty \\
&\le \gamma^{k+1} \underbrace{\|Q^* - Q^{\pi(0)}\|_\infty}_{\text{fixed}}
\end{aligned}$$

(Final Part) Policy Iteration: Convergence Proof

- $\|Q^* - Q^{\pi(k)}\|_\infty \leq \gamma^k \|Q^* - Q^{\pi(0)}\|_\infty$

$$= (1-(1-\gamma))^k \|Q^* - Q^{\pi(0)}\|_\infty$$

for any k

$$\leq e^{-(1-\gamma)k} \|Q^* - Q^{\pi(0)}\|_\infty \leq \epsilon$$

$$\boxed{k \geq \frac{1}{1-\gamma} \log\left(\frac{\|Q^* - Q^{\pi(0)}\|_\infty}{\epsilon}\right)}$$

→ atmost this many

$O(\log(1/\epsilon))$

$\|Q^* - Q^{\pi(k)}\|_\infty \leq \epsilon$

- 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)