RL Theory[1]: Meeting 4 (Chapter 1/2)

Shivam Garg, RLAI@UAlberta

6th October 2020

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

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

$$V_\phi(s) = \max_a \phi(s,a)$$

$$\mathcal{T}\phi = r + \gamma P V_\phi$$

$\phi(s,a)$, $r(s,a)$: $|S||A| \times 1$

P: $p(s'|s,a)$

$p(s',a'|s,a)$ … $p(s_{|S|},a_{|A|}|s,a)$ ← p^π

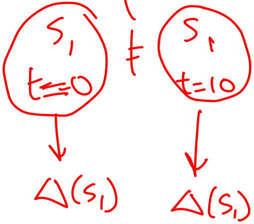
$|S \times A| \times |S \times A|$

$$\mathcal{T}^\pi V = r + \gamma P_\pi^{pred} V$$

$p(s'|s)$: $|S| \times |S|$

Complete Space of Policies

- We define the policy as $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$. This is a stationary markov policy.
- A deterministic stationary markov policy would be defined as $\pi : \mathcal{S} \rightarrow \mathcal{A}$.
- A general policy (possible non–deterministic, non–stationary, and non–markov) would be defined as $\pi : \mathcal{H} \rightarrow \Delta(\mathcal{A})$, 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 $s \in \mathcal{S}$,

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

Skip (didn't understand the proof yet)

- 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

$$Q = \mathcal{T}Q.$$

$$V^{\pi^*}(s) \geq V^{\pi}(s) \quad \forall s \in \mathcal{S}$$

$$\max_\pi \left(V^\pi(s_1) + V^\pi(s_2) + V^\pi(s_3) \right) = \max_\pi V^\pi(s_1) + \max_\pi V^\pi(s_2) + \cdots$$

$$Q^*(s,a) \doteq \max_\pi Q^\pi(s,a)$$

$$= \max_\pi \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^\pi(s') \right]$$

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

$$= r(s,a) + \gamma \sum_{s'} p(s'|s,a) V^*(s')$$

$$V^*(s') = \max_a Q^*(s',a) \quad \text{optimality eqn.}$$

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

$$\exists \pi^*$$

$$Q^* = \mathcal{T}Q^*$$

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

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

$$P V_Q = P^{\pi_Q} Q$$

(s,a)

$$\sum_{s'} p(s' \mid s,a) \max_{a'} Q(s',a')$$

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

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

for any π':

$$Q^{\pi'} - Q \leq 0$$

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

$$= (I - \gamma P^{\pi'})^{-1} \left[I - (I - \gamma P^{\pi'})(I - \gamma P^{\pi_Q})^{-1} \right] r$$

$$= (I - \gamma P^{\pi'})^{-1} \left[I - \gamma P^{\pi_Q} - (I - \gamma P^{\pi'}) \right] (I - \gamma P^{\pi_Q})^{-1} r$$

$$= (I - \gamma P^{\pi'})^{-1} \gamma \left[P^{\pi'} - P^{\pi_Q} \right] Q$$

$(I - \gamma P^{\pi'})^{-1} \geq 0$, $\left[P^{\pi'} - P^{\pi_Q} \right] Q \leq 0$

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)

Sample Complexity

- What is it? Why do we need it?
- Previously, we discussed DP algorithms like value iteration: $Q^{(k+1)} = \mathcal{T} Q^{(k)}$ for $k = 0, 1, 2, \ldots$. Let $\pi^{(k)} = \pi_{Q^{(k)}}$. Then for $k \geq \frac{1}{1-\gamma} \log\left(\frac{2}{\epsilon(1-\gamma)}\right)$,

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

- This assumes access to the true transition dynamics P, which is not available.
- So we address this question: How do these methods perform when we don't have the true P?

Sample Complexity (contd)

- We will begin by assuming a naïve model of the environment $\hat{P}$, defined as

$$\hat{P}(s'|s,a) = \frac{\text{count}(s', a, a)}{N}$$

- Define $\hat{M}, \hat{V}^\pi, \hat{Q}^\pi, \hat{Q}^*, \hat{\pi}^*$.
- Then we will see that given the inaccuracy in this model, how accurate our estimates of, say, $\hat{Q}^\pi$ can be?

Sample Complexity for a Naïve Model

- There exists a constant c. Let $\epsilon \in \left(0, \frac{1}{1-\gamma}\right)$. If,

$$\#\text{ of samples} \geq \frac{\gamma}{(1-\gamma)^4} \frac{|\mathcal{S}|^2 |\mathcal{A}| \log\left(\frac{c|\mathcal{S}||\mathcal{A}|}{\delta}\right)}{\epsilon^2}$$

then following hold with a probability of greater than $1-\delta$:

- (Model Accuracy) $\max_{s,a} \|P(\cdot|s,a) - \hat{P}(\cdot|s,a)\|_1 \leq (1-\gamma)^2 \frac{\epsilon}{2}$
- (Uniform Value Accuracy) $|Q^\pi - \hat{Q}^\pi\|_\infty \leq \frac{\epsilon}{2} \quad \forall \pi$
- (Near Optimal Planning) $\|Q^* - \hat{Q}^*\|_\infty \leq \epsilon$

Simulation Lemma

- $$Q^{\pi} - \hat{Q}^{\pi} = \gamma \left(I - \gamma \hat{P}^{\pi} \right)^{-1} \left(P - \hat{P} \right) V^{\pi}.$$

Another Useful Result

- For $x \in |\mathcal{S} \times \mathcal{A}|$

$$\|\left(I - \gamma \hat{P}^{\pi}\right)^{-1} x\|_{\infty} \leq \frac{\|x\|_{\infty}}{1 - \gamma}$$