Maria Vittoria Vestini
1795724

Value Iteration, Policy Iteration and iLQR

1. Derive the formula to get the minimum number of iterations of Value Iteration that are needed if we want an error on the quality of the policy that is at most ϵ

Value Iteration is a method used to compute the Optimal Value V^* of a Markov Decision Process, the main idea behind this method is to used mathematical properties to find a way to optain V^* starting from an itial value V_0.

The reason why we need alternative ways to compute the optimal Value is that the direct computation is highly expensive: given the definition of $V^\pi(s) := r + \gamma \mathbb{E}_{p(\cdot|s,\pi)}[V^\pi(s')]$ it's possible to extract the matricial form of the equation $V^\pi = R + \gamma P^\pi V^\pi$ and as a consequence the value of V^π

$$V^\pi = (I - \gamma P^\pi)^{-1} R$$

But this calculation – even if it's correct – is not feasible since $I - \gamma P^\pi \in \mathbb{R}^{SxS}$, where S is the number of states of the MDP, and just inverting the matrix will be a $\mathcal{O}(S^3)$ and this evaluation should be done for each possible policy π in order to determine V^*.

In Value Iteration, using two mathematical concepts:

- **Fixed Point**, given a map f we say that x is a fixed point for f if $x = f(x)$.
- **Contraction**, a map $f : X \to X$ is called a contraction if $\exists\ k \in (0,1) \mid \forall\ (x,\ x') \in X \times X \to |f(x) - f(x')| \leq k|x - x'|$.

an iterative algorithm is defined to compute V^*. In fact, those concepts taken together guarantee the following:

> Given $f : X \to X$ a contraction mapping, with a fixed point $\bar{x}$ and an initial value $x_0 \in X$.
>
> Then the iterative algorithm $x_i = f(x_{i-1})$ converges: $x_i \xrightarrow{i \to \infty} \bar{x}$.

So, we can define Value Iteration by taking $f = BV(s) = \max_a(r(s,a) + \gamma \mathbb{E}_{p(\cdot|s,a)}[V(s')])$ – the Bellman Operator – that, as shown in class, is indeed a contraction and its fixed value is V^*.

As for the number of iteration needed to obtain an error of the policy's quality to be less then ϵ, we want $\|V^{\pi_i} - V^*\| \leq \epsilon$, where π_i is computed using the Bellman Operator for Q:

$$\text{Operator } TQ(r,a) = r(s,a) + \gamma \mathbb{E}_{p(\cdot|s,a)} \max_{a'}[Q(s',a')]$$
$$\text{The policy } \pi_i(s) = \operatorname*{argmax}_a Q_i(s,a) \text{ and } Q_i = TQ_{i-1}$$

It can be shown, proof will be at the end of the section, that $\|V^{\pi_i} - V^*\| \leq \frac{2\gamma^i}{1-\gamma}\|Q_0 - Q^*\|$, so we can rewrite the problem as $\frac{2\gamma^i}{1-\gamma}\|Q_0 - Q^*\| \leq \epsilon$.

$$\frac{2\gamma^i}{1-\gamma}\|Q_0 - Q^*\| = \textit{Assume initialization } Q_0 = 0 \Rightarrow = \frac{2(1-(1-\gamma))^i}{1-\gamma}\|Q^*\| \leq$$

$$\leq \textit{since } 1 + x \leq e^x \textit{and } Q^* \in [0, \frac{1}{1-\gamma}] \leq \frac{2e^{-(1-\gamma)i}}{(1-\gamma)^2}$$

So to have that the error on the quality of the policy that is at most ϵ we have to find the value of i such that $\frac{2e^{-(1-\gamma)i}}{(1-\gamma)^2} \leq \epsilon$, since this would imply $\frac{2\gamma^i}{1-\gamma}\|Q_0 - Q^*\| \leq \epsilon$.

$$\frac{2e^{-(1-\gamma)i}}{(1-\gamma)^2} \leq \epsilon \Leftrightarrow e^{-(1-\gamma)i} \leq \frac{\epsilon(1-\gamma)^2}{2} \Leftrightarrow \log(e^{-(1-\gamma)i}) \leq \log\Big(\frac{\epsilon(1-\gamma)^2}{2}\Big) \Leftrightarrow$$

$$\Leftrightarrow -i(1-\gamma) \leq -log\Big(\frac{2}{\epsilon(1-\gamma)^2}\Big) \Leftrightarrow i \geq \frac{\log\Big(\frac{2}{\epsilon(1-\gamma)^2}\Big)}{1-\gamma}$$

Proof.

- $\|V^{\pi_i} - V^*\| \leq \frac{2\gamma^i}{1-\gamma}\|Q_0 - Q^*\|$:

$$\|V^{\pi_i} - V^*\| = \max_s -(V^{\pi_i}(s) - V^*(s)) \textit{ [this quantity is } \geq 0 \textit{ since } V^* \geq V^{\pi_i}]$$

$$V^{\pi_i}(s) - V^*(s) = Q^{\pi_i}(s, \pi^i(s)) - Q^*(s, \pi^i(s)) + Q^*(s, \pi^i(s)) - Q^*(s, \pi^*(s)) =$$

$$= \gamma \mathbf{E}_{s' \sim P(s,\pi^t(s))}(V^{\pi_i}(s') - V^*(s')) + Q^*(s, \pi^i(s)) - Q^*(s, \pi^*(s)) \geq$$

$$\textit{[Add the negative quantity: } - Q^{\pi_i}(s, \pi^i(s)) + Q^{\pi_i}(s, \pi^*(s))]$$

$$\geq \gamma \mathbf{E}_{s' \sim P(s,\pi^t(s))}(V^{\pi_i} - V^*) + Q^*(s, \pi^i(s)) - Q^{\pi_i}(s, \pi^i(s)) + Q^{\pi_i}(s, \pi^*(s)) - Q^*(s, \pi^*(s)) \geq$$

$$\textit{[Since } |Q^i(s) - Q^*(s)| \leq \|Q^i - Q^*\|_\infty \leq \gamma^i \|Q^0 - Q^*\|_\infty]$$

$$\geq \gamma \mathbf{E}_{s' \sim P(s,\pi^t(s))}(V^{\pi_i} - V^*) - 2\gamma^i \|Q^0 - Q^*\|_\infty \geq ... \geq -\frac{2\gamma^i}{1-\gamma}\|Q^0 - Q^*\|_\infty$$

2. Compute $V_{k+1}(s_6)$ following the Value Iteration algorithm.

- States $\{s_i : i \in \{1, .., 7\}\}$
- Rewards $r(s,a) = \begin{cases} 0.5 & \text{if } s = s_1 \\ 5 & \text{if } s = s_7 \\ 0 & \text{otherwise} \end{cases}$
- Dynamics $p(s_6|s_6, a_1) = 0.3 \;\; p(s_7|s_6, a_1) = 0.7$
- Policy $\pi(s) = a_1, \forall s$
- $v_k = [0.5, 0, 0, 0, 0, 0, 5]$
- $\gamma = 0.9$

We want to compute one iteration of the iterative algorithm used to comput the Value Function that, as state before, has a matricial form:

$$V_t = R + \gamma P V_{t-1}$$

In this specific case we have to compute $V_{k+1}(s_6)$ so the components needed to find this quantity are $r(s_6, a_1), v_k$ and s_{k+1}.

- $R = [0.5, 0, 0, 0, 0, 0, 5] \Rightarrow r(s_k = s_6, a_1) = 0$
- $V_{t-1} = v_k = [0.5, 0, 0, 0, 0, 0, 5]$
- $s_{k+1} \sim p(\cdot|s_k = s_6, a_1) = [0, 0, 0, 0, 0, 0.3, 0.7]$

So we can easily compute $v_{k+1}(s_6)$ as

$$r(s_k = s_6, a_1) + \gamma < v_k, p(\cdot|s_k = s_6, a_1) >$$
$$\Downarrow$$
$$v_{k+1}(s_6) = 0 + 0.9(0 * 0.3 + 5 * 0.7) = 3.15$$