

CME 241: Assignment 3

Matteo Santamaria (msantama@stanford.edu)

January 22, 2021

1. Deterministic Policies

From the **RLForFinanceBook**, we have four primary MDP Bellman Policy Equations:

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(s, a) \cdot \left[\mathcal{R}(s, a) + \gamma \cdot \sum_{s' \in \mathcal{S}} \mathcal{P}(s, a, s') \cdot V^\pi(s') \right] : s \in \mathcal{N} \quad (1)$$

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(s, a) \cdot Q^\pi(s, a) : s \in \mathcal{N} \quad (2)$$

$$Q^\pi(s, a) = \mathcal{R}(s, a) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s, a, s') \cdot V^\pi(s') : s \in \mathcal{N}, a \in \mathcal{A} \quad (3)$$

$$Q^\pi(s, a) = \mathcal{R}(s, a) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s, a, s') \sum_{a' \in \mathcal{A}} \pi(s', a') \cdot Q^\pi(s', a') : s \in \mathcal{N}, a \in \mathcal{A} \quad (4)$$

By definition, a *deterministic policy* is a policy π_D such that $\pi(s, \pi_D(s)) = 1$ and $\pi(s, a) = 0$ for all $a \in \mathcal{A} : a \neq \pi_D(s)$. As such, under a deterministic policy our Bellman Policy Equations can be simplified:

$$V^{\pi_D}(s) = \mathcal{R}(s, \pi_D(s)) + \gamma \cdot \sum_{s' \in \mathcal{S}} \mathcal{P}(s, \pi_D(s), s') \cdot V^{\pi_D}(s') : s \in \mathcal{N} \quad (5)$$

$$V^{\pi_D}(s) = Q^{\pi_D}(s, \pi_D(s)) : s \in \mathcal{N} \quad (6)$$

$$Q^{\pi_D}(s, \pi_D(s)) = \mathcal{R}(s, \pi_D(s)) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s, \pi_D(s), s') \cdot V^{\pi_D}(s') : s \in \mathcal{N} \quad (7)$$

$$Q^{\pi_D}(s, \pi_D(s)) = \mathcal{R}(s, \pi_D(s)) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s, \pi_D(s), s') \cdot Q^{\pi_D}(s', \pi_D(s')) : s \in \mathcal{N} \quad (8)$$

2. Analytic MDP

The general MDP Bellman Optimality equation can be expressed as follows:

$$V^*(s) = \mathcal{R}(s) + \gamma \cdot \sum_{s' \in \mathcal{S}} \mathcal{P}(s, s') \cdot V^*(s') : \forall s \in \mathcal{S}$$

We can then decompose the terms in the following way way:

$$\mathcal{R}(s) = \mathbb{E}[R_{t+1} \mid S_t = s] = \sum_{a \in \mathcal{A}} \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$$

$$\mathcal{R}(s) = \sum_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} r \cdot \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a]$$

$$\mathcal{R}(s) = \int_{a=0}^1 a(1-a) + (1-a)(1+a) da = \int_{a=0}^1 1 + a - 2a^2 da = \frac{5}{6}$$

$$\mathcal{P}(s, s') = \int_{a=0}^1 \mathcal{P}(s, a, s')$$

$$\mathcal{P}(s, s+1) = \int_{a=0}^1 a da = \frac{1}{2}$$

$$\mathcal{P}(s, s) = \int_{a=0}^1 1 - a da = \frac{1}{2}$$

Substituting these reductions into our original equation, we get:

$$V^*(s) = \frac{5}{6} + \frac{1}{2} \left[\frac{1}{2} V^*(s) + \frac{1}{2} V^*(s+1) \right] \implies V^*(s) = \frac{10}{9} + \frac{1}{3} V^*(s+1)$$

Which can be expressed as an infinite sum:

$$V^*(s) = \sum_{i=0}^{\infty} \frac{10}{9} \left(\frac{1}{3} \right)^i = \frac{5}{3}$$

Notice that $V^*(s)$ does not depend on s . We can find π^* by computing

$$\begin{aligned} \pi^*(s) &= \arg \max_{a \in \mathcal{A}} \{Q^*(s, a)\} = \arg \max_{a \in \mathcal{A}} \left\{ \mathcal{R}(s, a) + \gamma \cdot \sum_{s' \in \mathcal{S}} \mathcal{P}(s, a, s') \cdot V^*(s') \right\} \\ \pi^*(s) &= \arg \max_{a \in \mathcal{A}} \left\{ (1-a)a + (1+a)(a-1) + \frac{1}{2} \cdot \frac{5}{3} \right\} = \arg \max_{a \in \mathcal{A}} \left\{ \frac{11}{6} + a - 2a^2 \right\} \end{aligned}$$

Setting the derivative with respect to a equal to zero and solving for a yields

$$\pi^*(s) = \frac{1}{4}$$

3. Question 4

To find the optimal action for this MDP, we begin with a general expression for V^* .

$$V^*(s) = \min_{a \in \mathcal{A}} \left\{ \mathcal{R}(s, a) + \gamma \cdot \sum_{s' \in \mathcal{N}} \mathcal{P}(s, a, s') \cdot V^*(s') \right\} : s \in \mathcal{N}$$

For the myopic case where $\gamma = 0$, this can be reduced to

$$V^*(s) = \min_{a \in \mathcal{A}} [\mathcal{R}(s, a)] : s \in \mathcal{N}$$

And we can write

$$\mathcal{R}(s, a) = \mathbb{E}[R_{s'} | S_{t+1} = s', A_t = a] = \mathbb{E}[e^{as'}]$$

If $s' \sim \mathcal{N}(s, \sigma^2)$, then $as' \sim \mathcal{N}(as, a^2\sigma^2)$ so $\mathcal{R}(s, a) = e^{as + a^2\sigma^2/2}$. Now we want to minimize this cost with respect to a .

$$\frac{d\mathcal{R}}{da} = 0 \implies a^* = \frac{-s}{\sigma^2}$$

Which gives us our optimal action a^* for any state. To find the corresponding optimal cost, we evaluate

$$\mathcal{R}(s, a^*) = \exp\left[\frac{-s}{\sigma^2} \left(1 - \frac{s}{2}\right)\right]$$