Jordan Rowley

CME 241: Assignment 3

3.1

(a) For a deterministic policy π_D, $\pi_D(s,a)$ is an indicator function. Hence, $V^{\pi_D}(s) = Q^{\pi_D}(s, \pi_D(s))$

(b) Now $Q^{\pi_D}(s,a) = \mathbb{E}_{\pi_D}[G_t|S_t = s, A_t = a]$, therefore

$$Q^{\pi_D}(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}[s,a,s']V^{\pi_D}(s')$$

(c) From part (a), we have that $Q^{\pi_D}(s, \pi_D(s)) = V^{\pi_D}(s)$

(d) From part (c), we have that $Q^{\pi_D}(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}[s,a,s']Q^{\pi_D}(s, \pi_D(s'))$. Note that it need not be true that $\pi_D(s) = a$!

3.2

From the outset, the initial state does not affect any value function, hence $V^*(s) = V^*(s')$. The MDP Bellman Optimality Equation therefore reduces to

$$V^*(s) = \max_{a \in \mathcal{A}}(\mathcal{R}(s,a) + \gamma V^*(s))$$

Likewise, $\mathcal{R}(s,a)$ is not a function of the state s, because $\mathbb{E}[r_{t+1}|s_t = s, a_t = a] = (1-a)a - (1+a)(1-a) = -2a^2 + a + 1$. Now

$$V^*(s)(1-\gamma) = \max_{a \in [0,1]}(-2a^2 + a + 1)$$

This quadratic function in a is maximized when $a^* = 1/4$. This implies that the optimal value function is given by $a^*/(1-\gamma) = V^*(s) = 1/2$.

Likewise, the optimal policy is given by $\pi^*(s) = a^* = 1/4$.

3.4

We seek to minimize $\mathbb{E}[e^{as'}]$ where $s' \sim \mathcal{N}(s, \sigma^2)$. Now

$$\mathbb{E}[e^{as'}] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{ax} e^{-\frac{(x-s)^2}{2\sigma^2}} dx = e^{a + \frac{\sigma^2}{2}a^2} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{(x-(s+\sigma^2 a))^2}{2\sigma^2}} dx = e^{sa + \frac{\sigma^2}{2}a^2}$$

Minimizing this function amounts to simply minimizing $f(a) = sa + \frac{\sigma^2}{2}a^2$, which is a convex quadratic with vertex at $(s - \sigma^{-2}, -2s^2\sigma^{-2})$.

Therefore, the optimal action at any state is $a^* = -s\sigma^{-2}$, and the associated optimal cost is equal in expectation to $e^{-2s^2\sigma^{-2}}$.