I would like to use my 2-day extension.

RL HW 3

Q2.2 $d_\pi(s) = \sum_{t=0}^{\infty} \gamma^t \Pr(s|s_0, \pi)$

↓ rewrite

$$V_\pi(s) = \sum_a \pi(a|s_0) \Pr(s|a) R(s,a)$$

$$\therefore \quad V_\pi(s) = d_\pi(s) R(s) \qquad (1)$$

↖ discount · probability · reward

Policy Gradient Thm:

$$\nabla_\theta J(\theta) = \nabla V_\pi(s_0) \propto \sum_s \Pr(s) \sum_a q_\pi(s,a) \nabla_\pi(a|s,\theta)$$

$$= \nabla d_\pi(s_0) R(s_0) \quad \text{(constant)}$$

$$\propto \sum_{t=0}^{\infty} \gamma^t \nabla_\theta V_\pi(s_0) \quad \text{(deriv. of (1))}$$

$$\propto \sum_{t=0}^{\infty} \gamma^t \sum_s \Pr(s) \sum_a q_\pi(s,a) \nabla_\pi(a|s,\theta)$$

$$\nabla_\pi E_s[V_\pi(s)] = \nabla_\pi V_\pi(s)$$

where $V_\pi(s) = \sum_a q_\pi(s,a) \pi_\theta(a|s)$ from Bellman Eqn.

Answers to other problems (incl. 2.1 & discussions) in colab notebook