This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the | README.md | for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

1.a

MDP $M = (S, A, R, T, \gamma)$

$\pi : S \rightarrow \Delta(A)^{1}$

M has a single, fixed, starting state $s_0 \in S$

Expression for $p^{\pi}(\tau)$

$\tau = (s_0, a_0, s_1, a_1, \ldots)$ by running π in M

$$V^{\pi}(s_0) = \mathbb{E}_{\tau\sim p^{\pi}}\left(\sum_{t=0}^{\infty} \gamma^{t} R(s_t, a_t) \mid s_0\right)$$

$$p^{\pi}(\tau) = \prod_{t=0}^{\infty} \pi(a_t|s_t)\, T(s_{t+1}|s_t, a_t)$$

1.b

Discounted, stationary state distribution of a policy π as

$$d^\pi(s) = \left(1-\gamma^t\right)\sum_{t=0}^{\infty}\gamma^t p(s_t = s)$$

$p(s_t = s)$ denotes the probability of being in state s at timestep t while following policy π

$f(s,a) = 1,\ \forall (s,a) \in S \times A$

$p(s_t = s)$

$f : S \times A \to \mathbb{R}$

$$\mathbb{E}_{\tau\sim p^\pi}\left[\sum_{t=0}^{\infty}\gamma^t f(s_t,a_t)\right] = \frac{1}{(1-\gamma)}\mathbb{E}_{s\sim d^\pi}\left[\mathbb{E}_{a\sim\pi(s)}\left[f(s,a)\right]\right]$$

$$\begin{aligned}
\mathbb{E}_{\tau\sim p^\pi}\left[\sum_{t=0}^{\infty}\gamma^t f(s_t,a_t)\right] &= \sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{\tau\sim p^\pi}\left[f(s_t,a_t)\right] \\
&= \mathbb{E}_{\tau\sim p^\pi}\left[f(s_0,a_0)\right] + \gamma\mathbb{E}_{\tau\sim p^\pi}\left[f(s_1,a_1)\right] + \gamma^2\mathbb{E}_{\tau\sim p^\pi}\left[f(s_2,a_2)\right] + \ \ldots \\
&= \sum_{a_0}\pi(a_0|s_0)f(s_0,a_0) + \gamma\sum_{a_0}\pi(a_0|s_0)\sum_{s_1}T(s_1|s_0,a_0)\sum_{a_1}\pi(a_1|s_1)f(s_1,a_1) + \ldots \\
&= \sum_s p(s_0 = s)\,\mathbb{E}_{a\sim\pi(s)}\left[f(s,a)\right] + \gamma\sum_s p(s_1 = s)\,\mathbb{E}_{a\sim\pi(s)}\left[f(s,a)\right] + \ldots \\
&= \sum_s\sum_{t=0}^{\infty}\gamma^t p(s_t = s)\,\mathbb{E}_{a\sim\pi(s)}\left[f(s,a)\right] \\
&= \frac{1}{(1-\gamma)}\sum_s d^\pi(s)\,\mathbb{E}_{a\sim\pi(s)}\left[f(s,a)\right] \\
&= \frac{1}{(1-\gamma)}\mathbb{E}_{s\sim d^\pi}\left[\mathbb{E}_{a\sim\pi(s)}\left[f(s,a)\right]\right]
\end{aligned}$$

1.c

$$
\begin{aligned}
V^{\pi}(s_0) - V^{\pi'}(s_0) &= \mathbb{E}_{\tau\sim\rho^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t \mathcal{R}(s_t,a_t)\right] - V^{\pi'}(s_0) \\
&= \mathbb{E}_{\tau\sim\rho^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t \left(\mathcal{R}(s_t,a_t) + V^{\pi'}(s_t) - V^{\pi'}(s_t)\right)\right] - V^{\pi'}(s_0) \\
&= \mathbb{E}_{\tau\sim\rho^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t \left(\mathcal{R}(s_t,a_t) + \gamma V^{\pi'}(s_{t+1}) - V^{\pi'}(s_t)\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}_{\tau\sim p^{\pi}}\left[\mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\left(R(s_t,a_t) + \gamma V^{\pi'}(s_{t+1}) - V^{\pi'}(s_t)\right)|s_t, a_t\right]\right] \\
&= \mathbb{E}_{\tau\sim p^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\left(R(s_t,a_t) + \gamma \mathbb{E}\left[V^{\pi'}(s_{t+1})|s_t,a_t\right] - V^{\pi'}(s_t)\right)\right] \\
&= \mathbb{E}_{\tau\sim p^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t\left(Q^{\pi'}(s_t,a_t) - V^{\pi'}(s_t)\right)\right] \\
&= \mathbb{E}_{\tau\sim p^{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi'}(s_t,a_t)\right] \\
&= \frac{1}{(1-\gamma)}\mathbb{E}_{s\sim d^{\pi}}\left[\mathbb{E}_{a\sim\pi(s)}\left[A^{\pi'}(s,a)\right]\right]
\end{aligned}
$$

2.a

The maximum sum of rewards that can be achieved in a single trajectory in the test environment assuming $\gamma = 1$ is 6.2. This value is attainable in a single trajectory within the path $0 \rightarrow 2 \rightarrow 3 \rightarrow 2 \rightarrow 3 \rightarrow 0$.

No other trajectory can achieve a greater cumulative reward because first, the maximum reward achieved is 3 when the path goes from $2 \rightarrow 3$, wait for a step, execute it again.

There are 5 steps and 2 optimal moves–going less than 2 will have a smaller result. Going to 2 twice gives 0 reward on the 2 steps, which means that 4 steps give a maximum of 6.

The best reward that is achieved that is not starting from state 1 is 0.2. This yields an upper bound of 6.2

3.b

Maintain a table containing the value of $Q(s,a)$, an estimate of $Q^*(s,a)$ for every (s,a) pair
Update rule:

$Q(s,a) \leftarrow Q(s,a) + \alpha \left(r + \gamma \max_{a' \in A} Q(s',a') - Q(s,a) \right)$

Where $\alpha > 0$ is the learning rate, $\gamma \in [0,1)$ the discount factor

Q function is an unbiased estimator of Q^*, meaning that $\mathbb{E}[Q(s,a)] = Q^*(s,a)$ for all states s and actions a

$\forall s,\ \mathbb{E}[\max_a Q(s,a)] \geq \max_a Q^*(s,a)$

The expectation $\mathbb{E}[Q(s,a)]$ is over the randomness in Q resulting from the stochasticity of the exploration process

The expectation of max is $\geq$ max of the expectation:

$\mathbb{E}[\max_a Q(s,a)] \geq \max_a \mathbb{E}[\max_a Q(s,a)]$

$\mathbb{E}[Q(s,a)] = Q^*(s,a)$

For all actions, we have $max_{a'} Q(s,a') \geq Q(s,a)$ with probability 1:

$\forall a,\ \mathbb{E}\left[\max_{a'} Q(s,a')\right] \geq \mathbb{E}[Q(s,a)]$

If $X \geq Y$ with probability 1, then $\mathbb{E}[X] \geq \mathbb{E}[Y]$. If $\forall x$, $c \geq f(x)$ then $c \geq max_x f(x)$.

5.a

Represent the Q function as $Q_\theta(s,a,) = \theta^\top \hat{\delta}(s,a)$, where $\theta \in \mathbb{R}^{|S||A|}$ and $\hat{\delta} : S \times A \to \mathbb{R}^{|S||A|}$ with $[\hat{\delta}(s,a)]_{s',a'} = \begin{cases} 1 & \textit{if } s'=s,\ a'=a \\ 0 & \textit{otherwise} \end{cases}$

The equation for the tabular q-learning update rule:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left(r + \gamma \max_{a' \in A} Q(s',a') - Q(s,a)\right)$$

$$\theta \leftarrow \theta + \alpha\left(r + \gamma \max_{a' \in A} Q_\theta(s',a') - Q_\theta(s,a)\right) \nabla_\theta Q_\theta(s,a)$$

$$\nabla_\theta Q_\theta(s,a) = \nabla_\theta\left(\theta^\top \hat{\delta}(s,a)\right) = \hat{\delta}(s,a)$$

Update rule:

$$\theta \leftarrow \theta + \alpha\left(r + \gamma \max_{a' \in A} \theta^\top \hat{\delta}(s',a') - \theta^\top \hat{\delta}(s,a)\right) \hat{\delta}(s,a)$$

$$= \theta + \alpha\left(r + \gamma \max_{a' \in A} \theta_{s',a'} - \theta_{s,a}\right) \hat{\delta}(s,a)$$

$$\theta_{\bar{s},\bar{a}} \leftarrow \begin{cases} \theta_{\bar{s},a} + \alpha\left(r + \gamma\, max_{a' \in A} \theta_{\bar{s}',a'} - \theta_{\bar{s},a}\right) & \textit{if } (\bar{s},\bar{a}) = (s,a) \\ \theta_{\bar{s},\bar{a}} & \textit{otherwise} \end{cases}$$