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.b

Based on Jensen's inequality, we have:

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

for random variable X and a convex function f.

The maximum function is a convex function and we consider Q-values for all actions in a given state as a random variable. Applying Jensen's inequality, we get:

$$\mathbb{E}[\max_{a \in \mathcal{A}} Q(s,a)] \geq \max_{a \in \mathcal{A}} (\mathbb{E}[Q(s,a)])$$

We are given that $\mathbb{E}[Q(s,a)] = Q^*(s,a)$, so

$$\mathbb{E}[\max_{a \in \mathcal{A}} Q(s,a)] \geq \max_{a \in \mathcal{A}} Q^*(s,a)$$

2.a

The partial derivative of a dot product is simply the other vector in the dot product (i.e. $\partial(u \cdot v)/\partial u = v$), so

$$\nabla_\theta Q_\theta(s,a) = \frac{\partial Q_\theta(s,a)}{\partial \theta_{s,a}} = \frac{\partial \theta_{s,a} \cdot \delta(s,a)}{\partial \theta_{s,a}} = \delta(s,a)$$

Substitute this and into the linear update equation:

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

We get:

$$\theta \leftarrow \theta + \alpha(r + \gamma \max_{a'} Q_\theta(s',a') - Q_\theta(s,a))\delta(s,a)$$

Shows that only the single θ component correspond to (s,a) will be updated and thus identical to:

$$\theta \cdot \delta(s,a) \leftarrow \theta \cdot \delta(s,a) + \alpha(r + \gamma \max_{a'} Q_\theta(s',a') - Q_\theta(s,a))\delta(s,a)$$

And therefore

$$Q_\theta(s,a) \leftarrow Q_\theta(s,a) + \alpha(r + \gamma \max_{a'} Q_\theta(s',a') - Q_\theta(s,a))$$

3.c

The linear model is obviously learn quicker and give you more stable results while DQN can sometimes fail to reach the optimal policy with the epoch we choose. DQN might need even more training steps to reach the optimal policy.

In conclusion, for the simple test environment, the linear model is the better choice due to its efficiency and appropriate complexity. The DQN is unnecessarily powerful for this task.

4.a

4.c

4.d