1.a.

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

b. $$E_{\tau\sim\rho^\pi}\left[\sum_{t=0}^{\infty} \gamma^t f(s_t, a_t)\right] = \sum_{t=0}^{\infty} \gamma^t E_{\tau\sim\rho^\pi}[f(s_t, a_t)]$$

$$= E_{\tau\sim\rho^\pi}[f(s_0, a_0)] + \gamma E_{\tau\sim\rho^\pi}[f(s_1, a_1)] + \gamma^2 E_{\tau\sim\rho^\pi}[f(s_2, a_2)] + \cdots$$

$$= \sum_s P(s_0 = s) E_{a\sim\pi(s)}[f(s, a)] + \gamma \sum_s P(s_1 = s) E_{a\sim\pi(s)}[f(s, a)] + \cdots$$

$$= \sum_s \sum_{t=0}^{\infty} \gamma^t P(s_t = s) E_{a\sim\pi(s)}[f(s, a)]$$

$$= \frac{1}{1-\gamma} \sum_s d^\pi(s) E_{a\sim\pi(s)}[f(s, a)]$$

$$= \frac{1}{1-\gamma} E_{s\sim d^\pi}\left[E_{a\sim\pi(s)}[f(s, a)]\right]$$

c. $$V^\pi(s_0) - V^{\pi'}(s_0) = E_{\tau\sim\rho^\pi}\left[\sum_{t=0}^{\infty} \gamma^t (R(s_t, a_t) + \gamma V^{\pi'}(s_{t+1}) - V^{\pi'}(s_t))\right]$$

$$= E_{\tau\sim\rho^\pi}\left[E\left[\sum_{t=0}^{\infty} \gamma^t (R(s_t, a_t) + \gamma V^{\pi'}(s_{t+1}) - V^{\pi'}(s_t)) \middle| s_t, a_t\right]\right]$$

$$= E_{\tau\sim\rho^\pi}\left[\sum_{t=0}^{\infty} \gamma^t (R(s_t, a_t) + \gamma E[V^{\pi'}(s_{t+1})|s_t, a_t] - V^{\pi'}(s_t))\right]$$

$$= E_{\tau\sim\rho^\pi}\left[\sum_{t=0}^{\infty} \gamma^t (Q^{\pi'}(s_t, a_t) - V^{\pi'}(s_t))\right]$$

$$= E_{\tau\sim\rho^\pi}\left[\sum_{t=0}^{\infty} \gamma^t A^{\pi'}(s_t, a_t)\right]$$

$$= \frac{1}{1-\gamma} E_{s\sim d^\pi}\left[E_{a\sim\pi(s)}[A^{\pi'}(s_t, a_t)]\right]$$

2.a.

The maximum sum of rewards that can be achieved in a single trajectory is 6.2. To get this optimal reward, we need to make the following moves: 0 -> 2 ->3 ->2 ->3 ->0.

The max reward attainable from a single move is 3, when we move from state 2 to 3 taking the action 3. As we have 5 steps, we can only repeat this move twice. Doing so takes 4 steps and gives us a max reward of 6. Then at the final time step the highest reward we can achieve starting from state 3 is 0.2, taking an action of 0 to go to state 0.

3.b.

Based on Jensen's inequality, we can show that the expectation of the max is at least the max of the expectation. Given that Q is an unbiased estimator of Q* the inequality can then be written w.r.t Q*.

$$E\left[\max_a Q(s,a)\right] \geq \max_a E\left[Q(s,a)\right]$$
$$= \max_a Q^*(s,a)$$

5.a. Deriving the gradient w.r.t. θ we have:

$$\nabla_\theta Q_\theta(s,a) = \nabla_\theta\left(\theta^T \delta(s,a)\right) = \delta(s,a)$$

So the update rule for θ becomes:

$$\theta \leftarrow \theta + \alpha\left(r + \gamma \max_{a' \in A} \theta^T \delta(s',a') - \theta^T \delta(s,a)\right)\delta(s,a)$$
$$= \theta + \alpha\left(r + \gamma \max_{a' \in A} \theta_{s',a'} - \theta_{s,a}\right)\delta(s,a)$$

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

This is equivalent to Q function update as $Q_\theta(s,a) = \theta_{s,a}$