Comp 767 Assignment 3

Jonathan Pearce 260672004

April 9, 2020

Problem 1.A We begin by expanding the expectation under the behaviour policy μ,

$$\mathbb{E}_\mu[\Delta\tilde{\theta}|s_0,a_0] = \mathbb{E}_\mu\Big[\sum_{t=0}^{\infty}\alpha(\tilde{R}_t^\lambda - \theta^T\varphi_t)\varphi_t\rho_1\rho_2...\rho_t\Big|s_0,a_0\Big]$$

From class (and the textbook) we have the equation for the lambda return, $\tilde{R}_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\tilde{R}_{t:t+n}$. With this we get,

$$= \mathbb{E}_\mu\Big[\sum_{t=0}^{\infty}\alpha\Big((1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\tilde{R}_{t:t+n} - \theta^T\varphi_t\Big)\varphi_t\rho_1\rho_2...\rho_t\Big|s_0,a_0\Big]$$

Using properties of the geometric series and the fact that $\lambda \in [0,1)$, we have $(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1} = 1$. Therefore $(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\theta^T\varphi_t = \theta^T\varphi_t$. It follows,

$$\begin{aligned}
&= \mathbb{E}_\mu\Big[\sum_{t=0}^{\infty}\alpha\Big((1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\tilde{R}_{t:t+n} - (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\theta^T\varphi_t\Big)\varphi_t\rho_1\rho_2...\rho_t\Big|s_0,a_0\Big]\\
&= \mathbb{E}_\mu\Big[\sum_{t=0}^{\infty}\alpha(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\Big(\tilde{R}_{t:t+n} - \theta^T\varphi_t\Big)\varphi_t\rho_1\rho_2...\rho_t\Big|s_0,a_0\Big]\\
&= \mathbb{E}_\mu\Big[\sum_{t=0}^{\infty}\sum_{n=1}^{\infty}\alpha(1-\lambda)\lambda^{n-1}\Big(\tilde{R}_{t:t+n} - \theta^T\varphi_t\Big)\varphi_t\rho_1\rho_2...\rho_t\Big|s_0,a_0\Big] \qquad (1)
\end{aligned}$$

Next, we expand the expectation under the target policy π,

$$\mathbb{E}_\pi[\Delta\theta|s_0,a_0] = \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty}\alpha(R_t^\lambda - \theta^T\varphi_t)\varphi_t\Big|s_0,a_0\Big]$$

From class (and the textbook) we have the equation for the lambda return, $R_t^\lambda = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}R_{t:t+n}$. With this we get,

$$= \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty}\alpha\Big((1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}R_{t:t+n} - \theta^T\varphi_t\Big)\varphi_t\Big|s_0,a_0\Big]$$

Using $(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\theta^T\varphi_t = \theta^T\varphi_t$ from above, it follows,

$$\begin{aligned}
&= \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty}\alpha\Big((1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}R_{t:t+n} - (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\theta^T\varphi_t\Big)\varphi_t\Big|s_0,a_0\Big]\\
&= \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty}\alpha(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\Big(R_{t:t+n} - \theta^T\varphi_t\Big)\varphi_t\Big|s_0,a_0\Big]\\
&= \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty}\sum_{n=1}^{\infty}\alpha(1-\lambda)\lambda^{n-1}\Big(R_{t:t+n} - \theta^T\varphi_t\Big)\varphi_t\Big|s_0,a_0\Big] \qquad (2)
\end{aligned}$$

Our goal is to equate equations (1) and (2). Noting that α and λ are constants, it is therefore sufficient to prove that for any fixed n,

$$\mathbb{E}_\mu\Big[\sum_{t=0}^{\infty}(\tilde{R}_{t:t+n} - \theta^T\varphi_t)\varphi_t\rho_1\rho_2...\rho_t\Big|s_0,a_0\Big] = \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty}(R_{t:t+n} - \theta^T\varphi_t)\varphi_t\Big|s_0,a_0\Big]$$

We begin by moving the summation outside the expectation:

$$\mathbb{E}_\mu\Big[\sum_{t=0}^{\infty}(\tilde{R}_{t:t+n}-\theta^T\varphi_t)\varphi_t\rho_1\rho_2...\rho_t\Big|s_0,a_0\Big]=\sum_{t=0}^{\infty}\mathbb{E}_\mu\Big[(\tilde{R}_{t:t+n}-\theta^T\varphi_t)\varphi_t\rho_1\rho_2...\rho_t\Big|s_0,a_0\Big]$$

We now expand the expectation over the behaviour policy μ by enumerating over all possible trajectories of length t. Let $\text{traj}(t)$ be the set of all trajectories of length t. For a fixed trajectory $\tau \in \text{traj}(t)$, let $p_\mu(\tau)$ be the probability of this trajectory.

$$=\sum_{t=0}^{\infty}\sum_{\tau\in\text{traj}(t)}p_\mu(\tau)\mathbb{E}_\mu\Big[(\tilde{R}_{t:t+n}-\theta^T\varphi_t)\varphi_t\rho_1\rho_2...\rho_t\Big|s_0,a_0,s_1,a_1,...,s_t,a_t\Big]$$

Given $\tau=\{s_0,a_0,s_1,a_1,...,s_t,a_t\}$, $\theta^T,\varphi_t,\rho_1,\rho_2...\rho_t$ are all fixed. By the linearity of expectation we get,

$$=\sum_{t=0}^{\infty}\sum_{\tau\in\text{traj}(t)}p_\mu(\tau)\Big(\mathbb{E}_\mu\Big[\tilde{R}_{t:t+n}\Big|s_0,a_0,s_1,a_1,...,s_t,a_t\Big]-\theta^T\varphi_t\Big)\varphi_t\rho_1\rho_2...\rho_t$$

By the Markov property, $\mathbb{E}_\mu\Big[\tilde{R}_{t:t+n}\Big|s_0,a_0,s_1,a_1,...,s_t,a_t\Big]=\mathbb{E}_\mu\Big[\tilde{R}_{t:t+n}\Big|s_t,a_t\Big]$

$$=\sum_{t=0}^{\infty}\sum_{\tau\in\text{traj}(t)}p_\mu(\tau)\Big(\mathbb{E}_\mu\Big[\tilde{R}_{t:t+n}\Big|s_t,a_t\Big]-\theta^T\varphi_t\Big)\varphi_t\rho_1\rho_2...\rho_t$$

Substituting in $\rho_t=\frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}$

$$=\sum_{t=0}^{\infty}\sum_{\tau\in\text{traj}(t)}p_\mu(\tau)\Big(\mathbb{E}_\mu\Big[\tilde{R}_{t:t+n}\Big|s_t,a_t\Big]-\theta^T\varphi_t\Big)\varphi_t\prod_{i=1}^{t}\frac{\pi(a_i|s_i)}{\mu(a_i|s_i)}$$

From Sutton and Barto (page 85) we can expand $p_\mu(\tau)$. We get $p_\mu(\tau)=\prod_{j=1}^{t}\mu(a_j|s_j)T(s_j|s_{j-1},a_{j-1})$. Where $T(s_j|s_{j-1},a_{j-1})$ is the probability of transitioning to state s_j given we were at s_{j-1} and took action a_{j-1}

$$=\sum_{t=0}^{\infty}\sum_{\tau\in\text{traj}(t)}\Big(\mathbb{E}_\mu\Big[\tilde{R}_{t:t+n}\Big|s_t,a_t\Big]-\theta^T\varphi_t\Big)\varphi_t\prod_{i=1}^{t}\frac{\pi(a_i|s_i)}{\mu(a_i|s_i)}\prod_{j=1}^{t}\mu(a_j|s_j)T(s_j|s_{j-1},a_{j-1})$$

$$=\sum_{t=0}^{\infty}\sum_{\tau\in\text{traj}(t)}\Big(\mathbb{E}_\mu\Big[\tilde{R}_{t:t+n}\Big|s_t,a_t\Big]-\theta^T\varphi_t\Big)\varphi_t\prod_{i=1}^{t}\frac{\pi(a_i|s_i)}{\mu(a_i|s_i)}\mu(a_i|s_i)T(s_i|s_{i-1},a_{i-1})$$

$$=\sum_{t=0}^{\infty}\sum_{\tau\in\text{traj}(t)}\Big(\mathbb{E}_\mu\Big[\tilde{R}_{t:t+n}\Big|s_t,a_t\Big]-\theta^T\varphi_t\Big)\varphi_t\prod_{i=1}^{t}\pi(a_i|s_i)T(s_i|s_{i-1},a_{i-1})$$

Similarly to above we have $p_\pi(\tau)=\prod_{i=1}^{t}\pi(a_i|s_i)T(s_i|s_{i-1},a_{i-1})$

$$=\sum_{t=0}^{\infty}\sum_{\tau\in\text{traj}(t)}\Big(\mathbb{E}_\mu\Big[\tilde{R}_{t:t+n}\Big|s_t,a_t\Big]-\theta^T\varphi_t\Big)\varphi_t p_\pi(\tau) \tag{3}$$

Now using the fact that $\mathbb{E}_\mu\Big[\tilde{R}_{t:t+n}\Big|s_t,a_t\Big]=\mathbb{E}_\pi\Big[R_{t:t+n}\Big|s_t,a_t\Big]$ (proof on the next page) we obtain:

$$=\sum_{t=0}^{\infty}\sum_{\tau\in\text{traj}(t)}\Big(\mathbb{E}_\pi\Big[R_{t:t+n}\Big|s_t,a_t\Big]-\theta^T\varphi_t\Big)\varphi_t p_\pi(\tau) \tag{4}$$

By the Markov property, $\mathbb{E}_\pi\Big[R_{t:t+n}\Big|s_0,a_0,s_1,a_1,...,s_t,a_t\Big]=\mathbb{E}_\pi\Big[R_{t:t+n}\Big|s_t,a_t\Big]$

$$=\sum_{t=0}^{\infty}\sum_{\tau\in\text{traj}(t)}\Big(\mathbb{E}_\pi\Big[R_{t:t+n}\Big|s_0,a_0,s_1,a_1,...,s_t,a_t\Big]-\theta^T\varphi_t\Big)\varphi_t p_\pi(\tau)$$

$$=\sum_{t=0}^{\infty}\sum_{\tau\in\text{traj}(t)}p_\pi(\tau)\mathbb{E}_\pi\Big[(R_{t:t+n}-\theta^T\varphi_t)\varphi_t\Big|s_0,a_0,s_1,a_1,...,s_t,a_t\Big]$$

Putting the expectation back together,

$$= \sum_{t=0}^{\infty} \mathbb{E}_\pi \Big[(R_{t:t+n} - \theta^T \varphi_t) \varphi_t \Big| s_0, a_0 \Big]$$
$$= \mathbb{E}_\pi \Big[\sum_{t=0}^{\infty} (R_{t:t+n} - \theta^T \varphi_t) \varphi_t \Big| s_0, a_0 \Big]$$

Below is the proof of $\mathbb{E}_\mu \Big[\tilde{R}_{t:t+n} \Big| s_t, a_t \Big] = \mathbb{E}_\pi \Big[R_{t:t+n} \Big| s_t, a_t \Big]$ used in our transition from equation (3) to (4) above.

$$\mathbb{E}_\mu \Big[\tilde{R}_{t:t+n} \Big| s_t, a_t \Big] = \mathbb{E}_\mu \Big[r_{t+1} + \gamma r_{t+2} \rho_{t+1} + ... + \gamma^{n-1} r_{t+n} \rho_{t+1} ... \rho_{t+n-1} + \gamma^n \theta^T \varphi_{t+n} \rho_{t+1} ... \rho_{t+n} \Big| s_t, a_t \Big]$$
$$= \mathbb{E}_\mu \Big[r_{t+1} + \gamma r_{t+2} \rho_{t+1} + ... + \gamma^{n-1} r_{t+n} \prod_{i=t+1}^{t+n-1} \rho_i + \gamma^n \theta^T \varphi_{t+n} \prod_{i=t+1}^{t+n} \rho_i \Big| s_t, a_t \Big]$$
$$= \mathbb{E}_\mu \Big[r_{t+1} \Big| s_t, a_t \Big] + \mathbb{E}_\mu \Big[\gamma r_{t+2} \rho_{t+1} \Big| s_t, a_t \Big] + ... + \mathbb{E}_\mu \Big[\gamma^{n-1} r_{t+n} \prod_{i=t+1}^{t+n-1} \rho_i \Big| s_t, a_t \Big] + \mathbb{E}_\mu \Big[\gamma^n \theta^T \varphi_{t+n} \prod_{i=t+1}^{t+n} \rho_i \Big| s_t, a_t \Big]$$
$$= \mathbb{E}_\mu \Big[r_{t+1} \Big| s_t, a_t \Big] + \gamma \mathbb{E}_\mu \Big[r_{t+2} \rho_{t+1} \Big| s_t, a_t \Big] + ... + \gamma^{n-1} \mathbb{E}_\mu \Big[r_{t+n} \prod_{i=t+1}^{t+n-1} \rho_i \Big| s_t, a_t \Big] + \gamma^n \mathbb{E}_\mu \Big[\theta^T \varphi_{t+n} \prod_{i=t+1}^{t+n} \rho_i \Big| s_t, a_t \Big]$$

Examining $\mathbb{E}_\mu \Big[r_{t+n} \prod_{i=t+1}^{t+n-1} \rho_i \Big| s_t, a_t \Big]$. We take a similar approach to as above and expand the expectation by enumerating over all trajectories of length n.

$$\mathbb{E}_\mu \Big[r_{t+n} \prod_{i=t+1}^{t+n-1} \rho_i \Big| s_t, a_t \Big] = \sum_{\tau \in \text{traj}(n)} p_\mu(\tau) \mathbb{E}_\mu \Big[r_{t+n} \prod_{i=t+1}^{t+n-1} \rho_i \Big| s_t, a_t, s_{t+1}, a_{t+1}, ..., s_{t+n}, a_{t+n} \Big]$$
$$= \sum_{\tau \in \text{traj}(n)} p_\mu(\tau) r_{t+n} \prod_{i=t+1}^{t+n-1} \rho_i$$
$$= \sum_{\tau \in \text{traj}(n)} r_{t+n} \prod_{j=t+1}^{t+n-1} \mu(a_j|s_j) T(s_j|s_{j-1}, a_{j-1}) \prod_{i=t+1}^{t+n-1} \rho_i$$
$$= \sum_{\tau \in \text{traj}(n)} r_{t+n} \prod_{j=t+1}^{t+n-1} \mu(a_j|s_j) T(s_j|s_{j-1}, a_{j-1}) \prod_{i=t+1}^{t+n-1} \frac{\pi(a_i|s_i)}{\mu(a_i|s_i)}$$
$$= \sum_{\tau \in \text{traj}(n)} r_{t+n} \prod_{i=t+1}^{t+n-1} \pi(a_i|s_i) T(s_i|s_{i-1}, a_{i-1})$$
$$= \sum_{\tau \in \text{traj}(n)} p_\pi(\tau) r_{t+n}$$
$$= \sum_{\tau \in \text{traj}(n)} p_\pi(\tau) \mathbb{E}_\pi \Big[r_{t+n} \Big| s_t, a_t, s_{t+1}, a_{t+1}, ..., s_{t+n}, a_{t+n} \Big]$$
$$= \mathbb{E}_\pi \Big[r_{t+n} \Big| s_t, a_t \Big]$$

Therefore we get,

$$\mathbb{E}_\mu \Big[\tilde{R}_{t:t+n} \Big| s_t, a_t \Big] = \mathbb{E}_\pi \Big[r_{t+1} \Big| s_t, a_t \Big] + \gamma \mathbb{E}_\pi \Big[r_{t+2} \Big| s_t, a_t \Big] + ... + \gamma^{n-1} \mathbb{E}_\pi \Big[r_{t+n} \Big| s_t, a_t \Big] + \gamma^n \mathbb{E}_\pi \Big[\theta^T \varphi_{t+n} \Big| s_t, a_t \Big]$$
$$= \mathbb{E}_\pi \Big[r_{t+1} \Big| s_t, a_t \Big] + \mathbb{E}_\pi \Big[\gamma r_{t+2} \Big| s_t, a_t \Big] + ... + \mathbb{E}_\pi \Big[\gamma^{n-1} r_{t+n} \Big| s_t, a_t \Big] + \mathbb{E}_\pi \Big[\gamma^n \theta^T \varphi_{t+n} \Big| s_t, a_t \Big]$$
$$= \mathbb{E}_\pi \Big[r_{t+1} \gamma r_{t+2} + ... + \gamma^{n-1} r_{t+n} + \gamma^n \theta^T \varphi_{t+n} \Big| s_t, a_t \Big]$$
$$= \mathbb{E}_\pi \Big[R_{t:t+n} \Big| s_t, a_t \Big]$$

Problem 1.B In order to derive an equivalent online algorithm we must first evaluate the term $(\tilde{R}_t^\lambda - \theta^T\varphi_t)$. We start by noting the recursive definition of $\tilde{R}_{t:t+n}$:

$$\begin{aligned}\tilde{R}_{t:t+n} &= r_{t+1} + \gamma r_{t+2}\rho_{t+1} + ... + \gamma^{n-1} r_{t+n}\rho_{t+1}...\rho_{t+n-1} + \gamma^n \theta^T\varphi_{t+n}\rho_{t+1}...\rho_{t+n} \\ &= r_{t+1} + \gamma\Big(r_{t+2}\rho_{t+1} + ... + \gamma^{n-2} r_{t+n}\rho_{t+1}...\rho_{t+n-1} + \gamma^{n-1}\theta^T\varphi_{t+n}\rho_{t+1}...\rho_{t+n}\Big) \\ &= r_{t+1} + \gamma\tilde{R}_{t+1:t+n}\end{aligned}$$

Now, let $\delta_{t+k} = r_{t+k+1} + \gamma\rho_{t+k+1}\theta^T\varphi_{t+k+1} - \theta^T\varphi_{t+k}$ be the TD error at time $t+k$. We begin by expanding $\tilde{R}_{t:t+n} - \theta^T\varphi_t$ with $n = 1, 2, 3...$ in order to identify a recursive form:

$$\begin{aligned}\tilde{R}_{t:t+1} - \theta^T\varphi_t &= r_{t+1} + \gamma\rho_{t+1}\theta^T\varphi_{t+1} - \theta^T\varphi_t \\ &= \delta_t \\ \tilde{R}_{t:t+2} - \theta^T\varphi_t &= r_{t+1} + \gamma\rho_{t+1}\tilde{R}_{t+1:t+2} - \theta^T\varphi_t \\ &= r_{t+1} + \gamma\rho_{t+1}\tilde{R}_{t+1:t+2} - \theta^T\varphi_t + \gamma\rho_{t+1}\theta^T\varphi_{t+1} - \gamma\rho_{t+1}\theta^T\varphi_{t+1} \\ &= \delta_t + \gamma\rho_{t+1}(\tilde{R}_{t+1:t+2} - \theta^T\varphi_{t+1}) \\ &= \delta_t + \gamma\rho_{t+1}\delta_{t+1} \\ \tilde{R}_{t:t+3} - \theta^T\varphi_t &= r_{t+1} + \gamma\rho_{t+1}r_{t+2} + \gamma^2\rho_{t+1}\rho_{t+2}\tilde{R}_{t+2:t+3} - \theta^T\varphi_t \\ &= r_{t+1} + \gamma\rho_{t+1}r_{t+2} + \gamma^2\rho_{t+1}\rho_{t+2}\tilde{R}_{t+2:t+3} - \theta^T\varphi_t + \gamma\rho_{t+1}\theta^T\varphi_{t+1} - \gamma\rho_{t+1}\theta^T\varphi_{t+1} \\ &= \delta_t + \gamma\rho_{t+1}r_{t+2} + \gamma^2\rho_{t+1}\rho_{t+2}\tilde{R}_{t+2:t+3} - \gamma\rho_{t+1}\theta^T\varphi_{t+1} \\ &= \delta_t + \gamma\rho_{t+1}r_{t+2} + \gamma^2\rho_{t+1}\rho_{t+2}\tilde{R}_{t+2:t+3} - \gamma\rho_{t+1}\theta^T\varphi_{t+1} + \gamma^2\rho_{t+1}\rho_{t+2}\theta^T\varphi_{t+2} - \gamma^2\rho_{t+1}\rho_{t+2}\theta^T\varphi_{t+2} \\ &= \delta_t + \gamma\rho_{t+1}(r_{t+2} + \gamma\rho_{t+2}\theta^T\varphi_{t+2} - \theta^T\varphi_{t+1}) + \gamma^2\rho_{t+1}\rho_{t+2}\tilde{R}_{t+2:t+3} - \gamma^2\rho_{t+1}\rho_{t+2}\theta^T\varphi_{t+2} \\ &= \delta_t + \gamma\rho_{t+1}\delta_{t+1} + \gamma^2\rho_{t+1}\rho_{t+2}\tilde{R}_{t+2:t+3} - \gamma^2\rho_{t+1}\rho_{t+2}\theta^T\varphi_{t+2} \\ &= \delta_t + \gamma\rho_{t+1}\delta_{t+1} + \gamma^2\rho_{t+1}\rho_{t+2}(\tilde{R}_{t+2:t+3} - \theta^T\varphi_{t+2}) \\ &= \delta_t + \gamma\rho_{t+1}\delta_{t+1} + \gamma^2\rho_{t+1}\rho_{t+2}\delta_{t+2}\end{aligned}$$

Let $\rho_{t:t+k} = \prod_{i=0}^{k}\rho_{t+i}$ and $\rho_t = 1$. Then we apply this process recursively and get:

$$\tilde{R}_{t:t+n} - \theta^T\varphi_t = \sum_{k=0}^{n-1}\gamma^k\rho_{t:t+k}\delta_{t+k} \tag{5}$$

From the Sutton and Barto textbook (page 290) we get an alternate definition of $\tilde{R}_t^\lambda$ which considers the termination time T of an episode:

$$\tilde{R}_t^\lambda = (1-\lambda)\sum_{n=1}^{T-t-1}\lambda^{n-1}\tilde{R}_{t:t+n} + \lambda^{T-t-1}\tilde{R}_{t:T} \tag{6}$$

Using equation (6) to expand $\tilde{R}_t^\lambda - \theta^T\varphi_t$.

$$\tilde{R}_t^\lambda - \theta^T\varphi_t = (1-\lambda)\sum_{n=1}^{T-t-1}\lambda^{n-1}\tilde{R}_{t:t+n} + \lambda^{T-t-1}\tilde{R}_{t:T} - \theta^T\varphi_t$$

Using the property of the finite geometric series we have $(1-\lambda)\sum_{n=1}^{T-t-1}\lambda^{n-1} + \lambda^{T-t-1} = 1$. Putting into our equation we get:

$$\begin{aligned}&= (1-\lambda)\sum_{n=1}^{T-t-1}\lambda^{n-1}\tilde{R}_{t:t+n} + \lambda^{T-t-1}\tilde{R}_{t:T} - \Big((1-\lambda)\sum_{n=1}^{T-t-1}\lambda^{n-1} + \lambda^{T-t-1}\Big)\theta^T\varphi_t \\ &= (1-\lambda)\sum_{n=1}^{T-t-1}\lambda^{n-1}(\tilde{R}_{t:t+n} - \theta^T\varphi_t) + \lambda^{T-t-1}(\tilde{R}_{t:T} - \theta^T\varphi_t)\end{aligned}$$

Substituting in equation (5)

$$= (1-\lambda)\sum_{n=1}^{T-t-1}\lambda^{n-1}\sum_{k=0}^{n-1}\gamma^k\rho_{t:t+k}\delta_{t+k} + \lambda^{T-t-1}\sum_{k=0}^{T-t-1}\gamma^k\rho_{t:t+k}\delta_{t+k}$$

$$
\begin{aligned}
&= \sum_{k=0}^{T-t-1} \Big(\sum_{n=k+1}^{T-t-1} (1-\lambda)\lambda^{n-1}\gamma^k \rho_{t:t+k}\delta_{t+k} + \lambda^{T-t-1}\gamma^k \rho_{t:t+k}\delta_{t+k} \Big) \\
&= \sum_{k=0}^{T-t-1} \gamma^k \rho_{t:t+k}\delta_{t+k} \Big(\sum_{n=k+1}^{T-t-1} (1-\lambda)\lambda^{n-1} + \lambda^{T-t-1} \Big) \\
&= \sum_{k=0}^{T-t-1} \gamma^k \rho_{t:t+k}\delta_{t+k} \Big((1-\lambda)\frac{\lambda^k(1-\lambda^{T-t-1-k})}{1-\lambda} + \lambda^{T-t-1} \Big) \\
&= \sum_{k=0}^{T-t-1} \gamma^k \lambda^k \rho_{t:t+k}\delta_{t+k}
\end{aligned}
$$

Therefore the update rule can be re-writen as follows,

$$\Delta\tilde{\theta}_t = \alpha \Big(\sum_{k=0}^{T-t-1} \gamma^k \lambda^k \rho_{t:t+k}\delta_{t+k} \Big) \varphi_t \rho_1 \rho_2 ... \rho_t \tag{7}$$

It is important to have the update rule for standard on-policy TD(λ) in order to identify the modifications required to create the desired online algorithm.

$$\Delta\theta_t = \alpha \Big(\sum_{k=0}^{T-t-1} \gamma^k \lambda^k \delta_{t+k} \Big) \varphi_t \tag{8}$$

The most obvious new feature of our algorithm is that for every step of an episode we must compute the importance sampling ratio ρ_t. Further, we need to keep a running product of all previously computed importance sampling ratios $\rho_1\rho_2...\rho_t$. The next important observation is that the TD error δ has a new form that we defined above, $\delta_t = R_t + \gamma\rho_t\theta^T\varphi_{t+1} + \theta^T\varphi_t$. The weight update $\Delta\tilde{\theta}_t$ is calculated with the same formula as in the TD(λ) algorithm, therefore $\Delta\tilde{\theta}_t = \alpha\delta_t z_t$, where z is the trace (we will use an accumulating trace). Next we must update the trace z at each time step in an episode. This is done by modifying the trace update from TD(λ) : $z_{t+1} = \gamma\lambda z_t + \varphi_{t+1}$ and identifying the differences between equations (7) and (8) above. First, we now multiply the term $\gamma\lambda z_t$ by ρ_{t+1}. Second, we now multiply the term φ_{t+1} by our running product of all previously computed importance sampling ratios $\rho_1\rho_2...\rho_t$. Combining these rules/equations, making sure we update state and actions at the end of every step in an episode and ensuring that the weight updates are done at the end of the episode as specified in the question we obtain the following online algorithm

Algorithm 1: Online Importance Sampled TD(λ)

Initialize $r = 1$
Initialize $\mathbf{z} = \mathbf{0}$
Repeat for each episode:
  Initialize S
  Choose $A \sim \mu(\cdot|S)$
  Repeat for each step t of episode:
    Take action A, observe R, S'
    Choose $A' \sim \mu(\cdot|S')$
    $\rho = \frac{\pi(S',A')}{\mu(S',A')}$
    $r = \rho r$
    $\delta = R + \gamma\rho\boldsymbol{\theta}^T\boldsymbol{\varphi}_{t+1} + \boldsymbol{\theta}^T\boldsymbol{\varphi}_t$
    $\Delta\boldsymbol{\theta}_t = \alpha\delta\mathbf{z}$
    $\mathbf{z} = \gamma\lambda\rho\mathbf{z} + r\boldsymbol{\varphi}_{t+1}$
    $S = S'$
    $A = A'$
  end
  $\boldsymbol{\theta} = \boldsymbol{\theta} + \sum_t \Delta\boldsymbol{\theta}_t$
end

Problem 2a There are three significant advantages to using Q-learning with experience replay; increase in data efficiency, reducing the variance of updates and avoiding oscillations or divergence in the network parameters. Experience replay stores tuples of agent experience $e_t = (s_t, a_t, r_t, s_{t+1})$ into a dataset $D_t = \{e_1, ..., e_t\}$. Q-updates are then subsequently made with samples of experience drawn randomly from the dataset, (s,a,r,s') $\sim U(D_t)$. This sampling procedure enables experiences to be used in weight updates more than once, and therefore Q-learning with experience replay is more data efficient than standard online Q-learning. This sampling procedure also prevents learning from (highly correlated) consecutive samples, which can be very inefficient. Sampling experience randomly from D_t ensures these correlations are broken up and subsequently reduces the variance of the updates. Finally, when learning on policy, the current parameters determine the next data sample that is used to update the parameter values. This can create unwanted feedback loops and the parameters can end up in local minima or diverge completely. Experience replay makes sure the behaviour distribution is averaged over many of its previous samples, this leads to more smoothed out learning updates and avoids oscillations and/or divergence in the parameters all together. One final note is that learning with experience replay must be off-policy because the current parameters are different than the parameters used to generate samples. This is why experience replay can be paired with Q-learning.

Problem 2b From the DQN paper the authors write "In the future, it will be important to explore the potential use of biasing the content of experience replay towards salient events...and relates to the notion of 'prioritized sweeping' in reinforcement learning.". In standard experience replay we sample from the dataset D_t uniformly, thus giving equal priority to all transitions currently stored in memory. The authors acknowledge that utilizing a more mathematically motivated sampling strategy could enable algorithms to focus on transitions from which they can learn the most, this is the idea behind prioritized experience replay. In prioritized experience replay, experiences $e_t = (s_t, a_t, r_t, s_{t+1})$ are given a priority according to their absolute TD error. The idea being to give more attention to transitions that are 'surprising' (i.e. have a large TD error). In theory prioritized experience replay would simply replace the uniform sampling of experience replay with a standard priority queue in order to select the next piece of experience to learn from. In practice, we require slightly more sophisticated methods to get prioritized experience replay to work well, these concepts are discussed below in part 2c.

Problem 2c On a high level the trade off between using prioritized experience replay or standard experience replay is a complexity vs performance trade off. An immediate downside of using prioritized experience replay over original experience replay is that it is more complicated in theory and in implementation. From reading the prioritized experience replay paper (Schaul et al., 2015) it becomes clear that getting prioritized experience replay to work effectively in practice involves much more than just replacing the random sampling from database D_t that is used in original experience replay, with a standard priority queue where the priority of a sample is according to its absolute TD error (greedy prioritized experience replay). In fact, there are at least three potential issues with greedy prioritized experience replay. First, it is not uncommon for a visited transition to never be replayed or to replayed for the first time long after being saved into memory. Greedy prioritized experience replay is also sensitive to noise spikes. Finally, it only focuses on a small subset of the experience which can lead to over-fitting. These three issues can be overcome by changing from greedy prioritization to stochastic prioritization, where we now sample transitions with probability p_t that is proportional to the last encountered absolute TD error:

$$p_t \propto \left| R_{t+1} - \gamma_{t+1} \max_{a'} q_{\bar{\theta}}(S_{t+1}, a') - q_{\theta}(S_{t+1}, A_t) \right|$$

Stochastic prioritization guarantees that the "probability of being sampled is monotonic in a transition's priority" and that all transitions no matter how low priority have non zero probability of being sampled. In the prioritized experience replay paper (Schaul et al., 2015) they introduce and analyze two variants of stochastic prioritization and demonstrate that they can be implemented relatively efficiently with respect to runtime and memory. However, regardless of implementation, stochastic based sampling will always be more complex with respect to time and space complexity then the uniform sampling used in original experience replay. Stochastic prioritization is still not perfect, it creates a bias because the updates no longer correspond to the same distribution as their expectation. In the paper (Schaul et al., 2015), they explain how to correct this bias. Their solution is to introduce importance sampling. This is another example of added complexity that comes with a successful implementation of prioritized experience replay.

Withstanding the added complexity of implementing prioritized experience replay there are still significant (experimental) advantages to using it over original experience replay. In the paper, prioritized experience replay improves peek performance as well as training efficiency (with respect to number of training steps) for both DQN and Double DQN that originally used standard experience replay. Another important point about prioritized experience replay is that it preserves the three advantages of using experience replay that were discussed in part 2a.