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.

2.b

Prove $Var(R_{t+1}) \geq Var(R_t)$
r_{t+1} is correlated with previous rewards

$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
$Cov(X+Y,Z) = Cov(X,Z) + Cov(Y,Z)$

$Var(R_{t+1}) = Var(R_t + r_{t+1}) = Var(R_t) + Var(r_{t+1}) + 2Cov(R_t, r_{t+1})$
Since $Var(r_{t+1}) \geq 0$, $Cov(R_t,\ r_{t+1}) \geq 0$
Assumes $Cov(R_t,\ r_{t+1}) = Cov\left(\sum_{i=0}^{t} r_i, r_{t+1}\right) = \sum_{i=0}^{t} Cov(r_i, r_{t+1}) > 0$
Strict inequality $Var(R_{t+1}) > Var(R_t)$

If relax to $\frac{1}{t+1}\sum_{i=0}^{t} Cov(r_i, r_{t+1}) \geq 0$, then $Var(R_{t+1}) \geq Var(R_t)$

2.c

$$\mathbb{E}_{\substack{s_{0:\infty}\\a_{0:\infty}}}[\sum_{t=0}^{\infty}\hat{A}_t(s_{0:\infty},a_{0:\infty})\nabla_\theta \log \pi_\theta(a_t,s_t)]$$
$$= \mathbb{E}_{\substack{s_{0:\infty}\\a_{0:\infty}}}[\sum_{t=0}^{\infty}(\hat{Q}_t(s_{t:\infty},a_{t:\infty}) - b_t(s_{0:t},a_{0:t-1}))\nabla_\theta \log \pi_\theta(a_t,s_t)]$$
$$= \mathbb{E}_{\substack{s_{0:\infty}\\a_{0:\infty}}}[\sum_{t=0}^{\infty}(\hat{Q}_t(s_{t:\infty},a_{t:\infty}))\nabla_\theta \log \pi_\theta(a_t,s_t)] - \mathbb{E}_{\substack{s_{0:\infty}\\a_{0:\infty}}}[\sum_{t=0}^{\infty}(b_t(s_{0:t},a_{0:t-1}))\nabla_\theta \log \pi_\theta(a_t,s_t)]$$

True function: $A^\pi(s_t,a_t)$

The general form of an estimator $\widehat{A}_t(s_{0:\infty},a_{0:\infty})$

$\widehat{A}_t(s_{0:\infty},a_{0:\infty}) = \widehat{Q}_t(s_{t:\infty},a_{t:\infty}) - b_t(s_{0:t},a_{0:t-1})$

$\mathbb{E}_{\substack{s_{t+1:\infty}\\a_{t+1:\infty}}}\left[\widehat{Q}_t(s_{t:\infty},a_{t:\infty})\right] = Q^\pi(s_t,a_t)$

Prove $\mathbb{E}_{\substack{s_{0:\infty}\\a_{0:\infty}}}\left[\sum_{t=0}^{\infty}\widehat{A}_t(s_{0:\infty},a_{0:\infty})\nabla_\theta log\pi_\theta(a_t,s_t)\right] = g$

$\mathbb{E}_\tau\left[(b_t(s_{0:t},a_{0:t-1}))\nabla_\theta log\pi_\theta(a_t,s_t) = 0\right]$

Second term:

$$\mathbb{E}_{\substack{s_{0:\infty}\\a_{0:\infty}}}\left[b_t(s_{0:t},a_0)\nabla_\theta log\pi_\theta(a_t,s_t)\right]$$
$$\mathbb{E}_{\substack{s_{0:t}\\a_{0:t}}}\left[b_t(s_{0:t},a_0)\nabla_\theta log\pi_\theta(a_t,s_t)\right]$$
$$= \mathbb{E}_{\substack{s_{0:t}\\a_{0:t-1}}}\left[\mathbb{E}_{a_t}\left[b_t(s_{0:t},a_{0:t-1})\nabla_\theta log\pi_\theta(a_t,s_t)\right]\right]$$
$$= \mathbb{E}_{\substack{s_{0:t}\\a_{0:t-1}}}\left[b_t(s_{0:t},a_{0:t-1})\,\mathbb{E}_{a_t}\left[\nabla_\theta log\pi_\theta(a_t,s_t)\right]\right]$$
$$= \mathbb{E}_{\substack{s_{0:t}\\a_{0:t-1}}}\left[b_t(s_{0:t},a_{0:t-1})\cdot 0\right]$$
$$= 0$$

First term:

$$\mathbb{E}_{\substack{s_{0:\infty}\\a_{0:\infty}}}\left[\left(\widehat{Q}_t(s_{t:\infty},a_{t:\infty})\right)\nabla_\theta log\pi_\theta(a_t,s_t)\right]$$
$$= \mathbb{E}_{\substack{s_{0:t}\\a_{0:t}}}\left[\mathbb{E}_{\substack{s_{t+1:\infty}\\a_{t+1:\infty}}}\left[\left(\widehat{Q}_t(s_{t:\infty},a_{t:\infty})\right)\nabla_\theta log\pi_\theta(a_t,s_t)\right]\right]$$
$$= \mathbb{E}_{\substack{s_{0:t}\\a_{0:t}}}\left[\nabla_\theta log\pi_\theta(a_t,s_t)\,\mathbb{E}_{\substack{s_{t+1:\infty}\\a_{t+1:\infty}}}\left[\left(\widehat{Q}_t(s_{t:\infty},a_{t:\infty})\right)\right]\right]$$
$$= \mathbb{E}_{\substack{s_{0:t}\\a_{0:t}}}\left[\nabla_\theta log\pi_\theta(a_t,s_t)\,Q^\pi(s_t,a_t)\right]$$

Note: $\mathbb{E}_{\substack{s_{0:t}\\a_{0:t}}}\left[\nabla_\theta log\pi_\theta(a_t,s_t)\,V^\pi(s_t)\right] = 0$

$$= \mathbb{E}_{\substack{s_{0:t}\\a_{0:t}}}\left[\nabla_\theta log\pi_\theta(a_t,s_t)(Q^\pi(s_t,a_t) - V^\pi(s_t))\right]$$
$$= \mathbb{E}_{\substack{s_{0:t}\\a_{0:t}}}\left[\nabla_\theta log\pi_\theta(a_t,s_t)\,A^\pi(s_t,a_t)\right] = 0$$

$$\mathbb{E}_{\substack{s_{0:\infty}\\a_{0:\infty}}}\left[\sum_{t=0}^{\infty}\widehat{A}_t(s_{0:\infty},a_{0:\infty})\nabla_\theta log\pi_\theta(a_t,s_t)\right] = g$$

2.d

TD error $\delta_t^{\widehat{V}}(s_t, a_t) = r_t + \gamma \widehat{V}(s_{t+1}) - \widehat{V}(s_t)$

If $\widehat{V} = V^{\pi}$, prove $\delta_t^{\widehat{V}}$ is an unbiased estimate of A^{π}

$$
\begin{aligned}
\mathbb{E}\left[\delta_t^{\widehat{V}}(s_t, a_t) \,|s_t, a_t\right] &= \mathbb{E}\left[\delta_t^{V^{\pi}}(s_t, a_t) \,|s_t, a_t\right] \\
&= \mathbb{E}\left[r_t + \gamma V^{\pi}(s_{t+1}) - V^{\pi}(s_t) \,|s_t, a_t\right] \\
&= \mathbb{E}\left[r_t + \gamma V^{\pi}(s_{t+1}) |s_t, a_t\right] - V^{\pi}(s_t) \\
&= Q^{\pi}(s_t, a_t) - V^{\pi}(s_t) \\
&= A^{\pi}(s_t, a_t)
\end{aligned}
$$

2.e

Define $\widehat{A}_t^{(k)} = \Sigma_{i=0}^{k-1} \gamma^i \delta_{t+i}^{\widehat{V}}$

Show $\widehat{A}_t^{(k)} = \ -\widehat{V}\ (s_t) + \gamma^k \widehat{V}\ (s_{t+k}) + \Sigma_{i=0}^{k-1} \gamma^i r_{t+i}$

$$
\begin{aligned}
\widehat{A}_t^{(k)} &= \sum_{i=0}^{k-1} \gamma^i \delta_{t+i}^{\widehat{V}} \\
&= \sum_{i=0}^{k-1} \gamma^i \left[r_{t+i} + \gamma \widehat{V}\ (s_{t+i+1}) - \widehat{V}\ (s_{t+i}) \right] \\
&= \ -\widehat{V}\ (s_t) + r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \ldots + \gamma^{k-1} r_{t+k-1} + \gamma^k \widehat{V}\ (s_{t+k}) \\
&= \widehat{V}\ (s_t) + \gamma^k \widehat{V}\ (s_{t+k}) + \sum_{i=0}^{k-1} \gamma^i r_{t+i}
\end{aligned}
$$

As k increases, reward term in sum increases, variance increases. The bias $\gamma^k \widehat{V}\ (s_{t+k})$ decreases as k increases.

2.f

Show $\widehat{A}_t^{(\infty)} = \Sigma_{i=0}^{\infty} \gamma^i r_{t+i} - \widehat{V}(s_t)$
$0 \leq \gamma \leq 1$

$$
\begin{aligned}
\widehat{A}_t^{(\infty)} &= \lim_{k\to\infty} \widehat{A}_t^{(k)} \\
&= \lim_{k\to\infty} \left(-\widehat{V}(s_t) + \gamma^k \widehat{V}(s_{t+k}) + \sum_{i=0}^{k-1} \gamma^i r_{t+i} \right) \\
&= -\widehat{V}(s_t) + \left(\lim_{k\to\infty} \gamma^k \widehat{V}(s_{t+k}) \right) + \left(\lim_{k\to\infty} \sum_{i=0}^{k-1} \gamma^i r_{t+i} \right) \\
&= -\widehat{V}(s_t) + 0 + \left(\lim_{k\to\infty} \sum_{i=0}^{k-1} \gamma^i r_{t+i} \right) \\
&= -\widehat{V}(s_t) + \sum_{i=0}^{\infty} \gamma^i r_{t+i}
\end{aligned}
$$