CS285: Deep Reinforcement Learning
Assignment 2
Written Report

Alan Sorani

May 13, 2025

1 Analysis

1. (a) Using policy gradients, we have

$$\nabla J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\left(\sum_{t=1}^{\infty} \nabla_\theta \log \pi_\theta (a_{i,t} \mid s_{i,t}) \right) R(\tau) \right]. \tag{1}$$

Due to the simplicity of our MDP, we can easily enumerate all the possible trajectories τ based on the first time-step t for which $s_t = S_F$. Let τ_i be the trajectory for which $s_{i,j} = s_1$ for all $j < i$ and for which $s_{i,i} = s_F$ (and therefore $s_{i,j} = S_F$ for all $j > i$). The probablity of the trajectory τ_i under the policy π_θ is then $\theta^{i-1}(1-\theta)$. We see that $R(\tau_i) = \sum_{j=1}^{i} r(s_{i,j} \mid a_{i,j}) = \sum_{j=1}^{i-1} 1 = i - 1$.
Writing the expectation of (1) explicitly, we get

$$\begin{aligned}
\nabla J(\theta) &= \sum_{i=1}^{\infty} p_\theta(\tau_i) \left[\left(\left(\sum_{t=1}^{i-1} \nabla_\theta \log(\theta) \right) + \nabla_\theta \log(1-\theta) \right) (i-1) \right] \\
&= \sum_{i=1}^{\infty} \theta^{i-1}(1-\theta) \left[\left(\frac{i-1}{\theta} - \frac{1}{1-\theta} \right) (i-1) \right] \\
&= \sum_{i=1}^{\infty} \theta^{i-1} \left[\left(\frac{(i-1)(1-\theta)}{\theta} - 1 \right) (i-1) \right] \\
&= \sum_{i=1}^{\infty} \theta^{i-1} \left[\left(\frac{(i-1)(1-\theta) - \theta}{\theta} \right) (i-1) \right] \\
&= \sum_{i=1}^{\infty} \theta^{i-2} \left[(i - \theta i - 1 + \theta - \theta)(i-1) \right] \\
&= \sum_{i=1}^{\infty} \theta^{i-2} (i - \theta i - 1)(i-1) \\
&= (1-\theta) \sum_{i=1}^{\infty} i(i-1)\theta^{i-2} - \sum_{i=1}^{\infty} (i-1)\theta^{i-2}.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
(1-\theta)\sum_{i=1}^{\infty} i\,(i-1)\,\theta^{i-2} &= (1-\theta)\sum_{i=1}^{\infty}\frac{\mathrm{d}^2}{\mathrm{d}^2\theta}\theta^i \\
&= (1-\theta)\,\frac{\mathrm{d}^2}{\mathrm{d}^2\theta}\sum_{i=1}^{\infty}\theta^i \\
&= (1-\theta)\,\frac{\mathrm{d}^2}{\mathrm{d}^2\theta}\frac{\theta}{1-\theta} \\
&= (1-\theta)\,\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{1}{(1-\theta)^2} \\
&= (1-\theta)\cdot\left(-\frac{2}{(1-\theta)^3}\right) \\
&= \frac{2}{(1-\theta)^2}
\end{aligned}$$

and

$$\begin{aligned}
\sum_{i=1}^{\infty}(i-1)\,\theta^{i-2} &= \sum_{i=1}^{\infty}\frac{\mathrm{d}}{\mathrm{d}\theta}\theta^{i-1} \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\sum_{i=1}^{\infty}\theta^{i-1} \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\frac{1}{1-\theta} \\
&= \frac{1}{(1-\theta)^2},
\end{aligned}$$

so we get

$$\nabla J(\theta) = \frac{2}{(1-\theta)^2} - \frac{1}{(1-\theta)^2} = \frac{1}{(1-\theta)^2}.$$

(b) We shall now compute $\mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)]$ directly and verify that

$$\nabla\mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)] = \frac{1}{(1-\theta)^2}.$$

We have

$$\begin{aligned}
\mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)] &= \sum_{i=1}^{\infty} p_\theta(\tau_i)\,R(\tau_i) \\
&= \sum_{i=1}^{\infty}\theta^{i-1}(1-\theta)(i-1) \\
&= (1-\theta)\sum_{i=1}^{\infty}\theta^{i-1}(i-1) \\
&= (1-\theta)\sum_{i=1}^{\infty}\frac{\mathrm{d}}{\mathrm{d}\theta}\theta^i \\
&= (1-\theta)\,\frac{\mathrm{d}}{\mathrm{d}\theta}\sum_{i=1}^{\infty}\theta^i \\
&= (1-\theta)\,\frac{\mathrm{d}}{\mathrm{d}\theta}\frac{\theta}{1-\theta} \\
&= (1-\theta)\cdot\frac{1}{(1-\theta)^2} \\
&= \frac{1}{1-\theta}
\end{aligned}$$

and then indeed

$$\nabla \mathbb{E}_{\tau\sim\pi_\theta}\left[R\left(\tau\right)\right] = \nabla\frac{1}{1-\theta} = \frac{1}{\left(1-\theta\right)^2}.$$

2. We have

$$\mathrm{Var}\left[X\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2,$$

so

$$\begin{aligned}
\mathrm{Var}_{\tau\sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta\left(\tau\right) r\left(\tau\right)\right] &= \mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\left(\nabla_\theta \log p_\theta\left(\tau\right) r\left(\tau\right)\right)^2\right] - \mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta\left(\tau\right) r\left(\tau\right)\right]^2 \\
&= \mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\left(\nabla_\theta \log p_\theta\left(\tau\right) r\left(\tau\right)\right)^2\right] - \nabla J\left(\theta\right)^2 \\
&= \mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\left(\nabla_\theta \log p_\theta\left(\tau\right) r\left(\tau\right)\right)^2\right] - \frac{1}{\left(1-\theta\right)^4}.
\end{aligned}$$

Now, using our computations from the previous part, we get

$$\begin{aligned}
\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\left(\nabla_\theta \log p_\theta\left(\tau\right) r\left(\tau\right)\right)^2\right] &= \sum_{i=1}^{\infty} p_\theta\left(\tau_i\right)\left[\left(\left(\sum_{t=1}^{i-1}\nabla_\theta \log\left(\theta\right)\right) + \nabla_\theta \log\left(1-\theta\right)\right)\left(i-1\right)\right]^2 \\
&= \sum_{i=1}^{\infty} \theta^{i-1}\left(1-\theta\right)\left[\left(\frac{i-1}{\theta} - \frac{1}{1-\theta}\right)\left(i-1\right)\right]^2 \\
&= \sum_{i=1}^{\infty} \frac{\theta^{i-3}}{1-\theta}\left(i\left(1-\theta\right)-1\right)^2\left(i-1\right)^2 \\
&= \frac{1}{1-\theta}\sum_{i=1}^{\infty} \theta^{i-3}\left(i\left(1-\theta\right)-1\right)^2\left(i-1\right)^2 \\
&= \frac{1}{1-\theta}\sum_{i=1}^{\infty} \theta^{i-3}\left(i^2\left(1-\theta\right)^2 - 2i\left(1-\theta\right)+1\right)\left(i^2-2i+1\right) \\
&= \frac{1}{\theta^3\left(1-\theta\right)}\sum_{i=1}^{\infty} \theta^{i}\left(i^4\theta^2 - 2i^4\theta + i^4 - 2i^3\theta^2 + 6i^3\theta - 4i^3 + i^2\theta^2 - 6i^2\theta + 6i^2 + 2i\theta - 4i + 1\right).
\end{aligned}$$

We get that

$$\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\left(\nabla_\theta \log p_\theta\left(\tau\right) r\left(\tau\right)\right)^2\right] = \frac{1}{\theta^3\left(1-\theta\right)}\left(S_1 - S_2 + S_3 - S_4 + S_5 - S_6 + S_7 - S_8 + S_9 + S_10 - S_11 + S_12\right)$$

where

$$S_1 = \sum_{i=1}^{\infty} i^4\theta^{i+2}, \qquad S_7 = \sum_{i=1}^{\infty} i^2\theta^{i+2}$$

$$S_2 = 2\sum_{i=1}^{\infty} i^4\theta^{i+1}, \qquad S_8 = 6\sum_{i=1}^{\infty} i^2\theta^{i+1}$$

$$S_3 = \sum_{i=1}^{\infty} i^4\theta^{i}, \qquad S_9 = 6\sum_{i=1}^{\infty} i^2\theta^{i}$$

$$S_4 = 2\sum_{i=1}^{\infty} i^3\theta^{i+2}, \qquad S_{10} = 2\sum_{i=1}^{\infty} i\theta^{i+1}$$

$$S_5 = 6\sum_{i=1}^{\infty} i^3\theta^{i+1}, \qquad S_{11} = 4\sum_{i=1}^{\infty} i\theta^{i}$$

$$S_6 = 4\sum_{i=1}^{\infty} i^3\theta^{i}, \qquad S_{12} = \sum_{i=1}^{\infty} \theta^{i}.$$

Now,

$$\sum_{i=1}^{\infty} \theta^i = \frac{\theta}{1-\theta}$$

so by differentiating

$$\sum_{i=1}^{\infty} i\theta^{i-1} = \frac{1}{(1-\theta)^2}$$

and by multiplying by θ

$$\sum_{i=1}^{\infty} i\theta^{i} = \frac{\theta}{(1-\theta)^2}.$$

By differentiating this and then multiplying by θ we get

$$\sum_{i=1}^{\infty} i^2\theta^{i} = \frac{\theta^2+\theta}{(1-\theta)^3}.$$

By repeating this we get

$$\sum_{i=1}^{\infty} i^3\theta^{i} = \frac{\theta^3+4\theta^2+\theta}{(1-\theta)^4}$$

and finally

$$\sum_{i=1}^{\infty} i^4\theta^{i} = \frac{\theta^4+11\theta^3+11\theta^2+\theta}{(1-\theta)^5}.$$

Using these equations we get

$$\begin{aligned}
S_1 &= \theta^2 \cdot \frac{\theta^4+11\theta^3+11\theta^2+\theta}{(1-\theta)^5}, & S_7 &= \theta^2 \cdot \frac{\theta^2+\theta}{(1-\theta)^3} \\
S_2 &= 2\theta \cdot \frac{\theta^4+11\theta^3+11\theta^2+\theta}{(1-\theta)^5}, & S_8 &= 6\theta \cdot \frac{\theta^2+\theta}{(1-\theta)^3} \\
S_3 &= \frac{\theta^4+11\theta^3+11\theta^2+\theta}{(1-\theta)^5}, & S_9 &= 6 \cdot \frac{\theta^2+\theta}{(1-\theta)^3} \\
S_4 &= 2\theta^2 \cdot \frac{\theta^3+4\theta^2+\theta}{(1-\theta)^4}, & S_{10} &= 2\theta \cdot \frac{\theta}{(1-\theta)^2} \\
S_5 &= 6\theta \cdot \frac{\theta^3+4\theta^2+\theta}{(1-\theta)^4}, & S_{11} &= 4 \cdot \frac{\theta}{(1-\theta)^2} \\
S_6 &= 4 \cdot \frac{\theta^3+4\theta^2+\theta}{(1-\theta)^4}, & S_{12} &= \frac{\theta}{1-\theta}.
\end{aligned}$$

Going back to the calculation of $\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\left(\nabla_\theta \log p_\theta(\tau)\, r(\tau)\right)^2\right]$ we get

$$\begin{aligned}
\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\left(\nabla_\theta \log p_\theta(\tau)\, r(\tau)\right)^2\right] &= \frac{1}{\theta^3(1-\theta)}\left(S_1 - S_2 + S_3 - S_4 + S_5 - S_6 + S_7 - S_8 + S_9 + S_10 - S_11 + S_12\right) \\
&= \frac{1}{\theta^3(1-\theta)} \cdot \frac{\theta^2\left(4\theta^2+9\theta+1\right)}{(1-\theta)^3} \\
&= \frac{4\theta^2+9\theta+1}{\theta(1-\theta)^4}
\end{aligned}$$

and so

$$\begin{aligned}\operatorname{Var}_{\tau\sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta(\tau)\, r(\tau)\right] &= \frac{4\theta^2+9\theta+1}{\theta(1-\theta)^4} - \frac{1}{(1-\theta)^4} \\ &= \frac{4\theta^2+8\theta+1}{\theta(1-\theta)^4}.\end{aligned}$$

To find the values of θ which minimize or maximize the variance, we compare

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\operatorname{Var}_{\tau\sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta(\tau)\, r(\tau)\right] = \frac{12\theta^3+36\theta^2+5\theta-1}{\theta^2(1-\theta)^5}$$

to 0. We get that the variance is minimal for $\theta \approx 0.10988$ and goes to infinity as θ approaches 0 or 1.

3. (a) An advantage estimator is a function A, possibly dependent on time, which gives the following estimate:

$$\nabla J(\theta) \approx \mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\sum_{t=1}^{\infty} \nabla_\theta \log \pi_\theta(a_{i,t} \mid s_{i,t})\, A_t(\tau)\right].$$

We consider

$$A_t(\tau) = \sum_{t'=t}^{\infty} r(s_{i,t'}, a_{i,t'})$$

which is the *return-to-go* advantage estimator.
We see that for any $t \in \mathbb{N}_+$,

$$A_t(\tau_i) = \sum_{t'=t}^{i} r(s_{i,t'} \mid a_{i,t'}) = \sum_{t'=t}^{i-1} 1 = i-1-(t-1) = i-t.$$

Hence

$$\begin{aligned}\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\sum_{t=1}^{\infty} \nabla_\theta \log \pi_\theta(a_{i,t} \mid s_{i,t})\, A_t(\tau)\right] &= \sum_{i=1}^{\infty} p_\theta(\tau_i)\left[\sum_{t=1}^{i-1} \nabla_\theta \log(\theta)\cdot(i-t)\right] \\ &= \sum_{i=1}^{\infty} \theta^{i-1}(1-\theta)\left(\sum_{t=1}^{i-1}\frac{i-t}{\theta}\right) \\ &= \sum_{i=1}^{\infty} \theta^{i-2}(1-\theta)\left(i(i-1) - \sum_{t=1}^{i-1} t\right) \\ &= \sum_{i=1}^{\infty} \theta^{i-2}(1-\theta)\left(i(i-1) - \frac{i(i-1)}{2}\right) \\ &= \frac{1}{2}\sum_{i=1}^{\infty} \theta^{i-2}(1-\theta)\, i(i-1) \\ &= \frac{1}{2}\left(\sum_{i=1}^{\infty}\theta^{i-2}i^2 - \sum_{i=1}^{\infty}\theta^{i-2}i - \sum_{i=1}^{\infty}\theta^{i-1}i^2 + \sum_{i=1}^{\infty}\theta^{i-1}i\right).\end{aligned}$$

Write

$$\begin{aligned}S_1' &\coloneqq \sum_{i=1}^{\infty}\theta^{i-2}i^2 \\ S_2' &\coloneqq \sum_{i=1}^{\infty}\theta^{i-2}i \\ S_3' &\coloneqq \sum_{i=1}^{\infty}\theta^{i-1}i^2 \\ S_4' &\coloneqq \sum_{i=1}^{\infty}\theta^{i-1}i.\end{aligned}$$

We calculated in the previous part that

$$S_4' = \frac{1}{(1-\theta)^2}.$$

Then

$$S_2' = \sum_{i=1}^{\infty} \theta^{i-2} i = \theta^{-1} \sum_{i=1}^{\infty} \theta^{i-1} i = \theta^{-1} S_4' = \frac{1}{\theta (1-\theta)^2}.$$

We also showed that

$$\sum_{i=1}^{\infty} i^2 \theta^i = \frac{\theta^2 + \theta}{(1-\theta)^3},$$

so similarly we get

$$S_3' = \frac{\theta + 1}{(1-\theta)^3}$$

and

$$S_1' = \frac{\theta + 1}{\theta (1-\theta)^3}.$$

Finally, we get

$$\begin{aligned}
\mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=1}^{\infty} \nabla_\theta \log \pi_\theta \left(a_{i,t} \mid s_{i,t} \right) \right] &= \frac{1}{2} \left(S_1' - S_2' - S_3' + S_4' \right) \\
&= \frac{1}{2} \cdot \frac{(\theta+1) - (1-\theta) - \left(\theta^2 + \theta\right) + \theta (1-\theta)}{\theta (1-\theta)^3} \frac{1}{2} \cdot \frac{-2\theta^2 + 2\theta}{\theta (1-\theta)^3} \\
&= \frac{\theta (1-\theta)}{\theta (1-\theta)^3} \\
&= \frac{1}{(1-\theta)^2} \\
&= \nabla J(\theta).
\end{aligned}$$

Hence our advantage estimator is unbiased in the sense that in expectation it has the same value

$$\mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=1}^{\infty} \nabla_\theta \log \pi_\theta \left(a_{i,t} \mid s_{i,t} \right) A_t(\tau) \right]$$

as the true value of the gradient, $\nabla J(\theta)$.

(b) We now compute the variance of the return-to-go policy variant. From what we've learnt in the lecture, return-to-go is used to reduce variance, so we expect a lower variance than the one we got in part 2.
We have

$$\begin{aligned}
\mathrm{Var}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=1}^{\infty} \nabla_\theta \log \pi_\theta \left(a_{i,t} \mid s_{i,t} \right) A_t(\tau) \right] &= \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\left(\sum_{t=1}^{\infty} \nabla_\theta \log \pi_\theta \left(a_{i,t} \mid s_{i,t} \right) A_t(\tau) \right)^2 \right] \\
&\quad - \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\sum_{t=1}^{\infty} \nabla_\theta \log \pi_\theta \left(a_{i,t} \mid s_{i,t} \right) A_t(\tau) \right]^2 \\
&= \mathbb{E}_{\tau \sim p_\theta(\tau)} \left[\left(\sum_{t=1}^{\infty} \nabla_\theta \log \pi_\theta \left(a_{i,t} \mid s_{i,t} \right) A_t(\tau) \right)^2 \right] - \frac{1}{(1-\theta)^4}.
\end{aligned}$$

Now, using our computations from part 3(a) we get

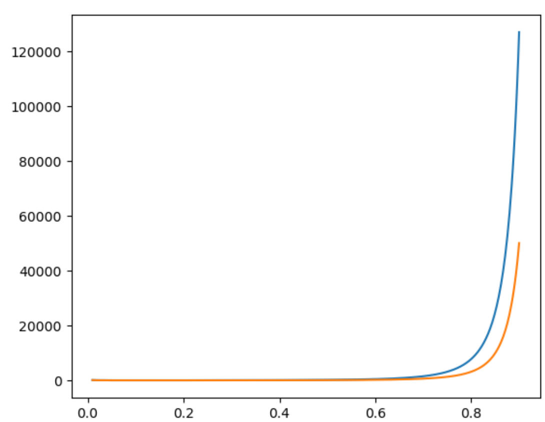
$$\begin{aligned}
\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\left(\sum_{t=1}^{\infty}\nabla_\theta\log\pi_\theta\left(a_{i,t}\mid s_{i,t}\right)A_t\left(\tau\right)\right)^2\right] &= \sum_{i=1}^{\infty}p_\theta\left(\tau_i\right)\left(\sum_{t=1}^{i-1}\nabla_\theta\log\left(\theta\right)\cdot\left(i-t\right)\right)^2 \\
&= \frac{1}{4}\sum_{i=1}^{\infty}\theta^{i-3}\left(1-\theta\right)i^2\left(i-1\right)^2 \\
&= \frac{1-\theta}{4\theta^3}\sum_{i=1}^{\infty}\theta^i i^2\left(i^2-2i+1\right) \\
&= \frac{1-\theta}{4\theta^3}\sum_{i=1}^{\infty}\theta^i\left(i^4-2i^3+i^2\right) \\
&= \frac{1-\theta}{4\theta^3}\left(\sum_{i=1}^{\infty}\theta^i i^4-2\sum_{i=1}^{\infty}\theta^i i^3+\sum_{i=1}^{\infty}\theta^i i^2\right) \\
&= \frac{1-\theta}{4\theta^3}\left(\frac{\theta^4+11\theta^3+11\theta^2+\theta}{\left(1-\theta\right)^5}-2\cdot\frac{\theta^3+4\theta^2+\theta}{\left(1-\theta\right)^4}+\frac{\theta^2+\theta}{\left(1-\theta\right)^3}\right) \\
&= \frac{1-\theta}{4\theta^3}\cdot\frac{4\theta^2\left(\theta^2+4\theta+1\right)}{\left(1-\theta\right)^5} \\
&= \frac{\theta^2+4\theta+1}{\theta\left(1-\theta\right)^4}
\end{aligned}$$

where the second-to-last equation is from calculations in part 2. Hence

$$\mathrm{Var}_{\tau\sim p_\theta(\tau)}\left[\sum_{t=1}^{\infty}\nabla_\theta\log\pi_\theta\left(a_{i,t}\mid s_{i,t}\right)A_t\left(\tau\right)\right] = \frac{\theta^2+4\theta+1}{\theta\left(1-\theta\right)^4}-\frac{1}{\left(1-\theta\right)^4} = \frac{\theta^2+3\theta+1}{\theta\left(1-\theta\right)^4}.$$

In Figure 1 we plot the variance before and after using return-to-go, and in Figure 2 we plot the reduction in variance gained by using return-to-go, which we notice is non-negative for all θ.

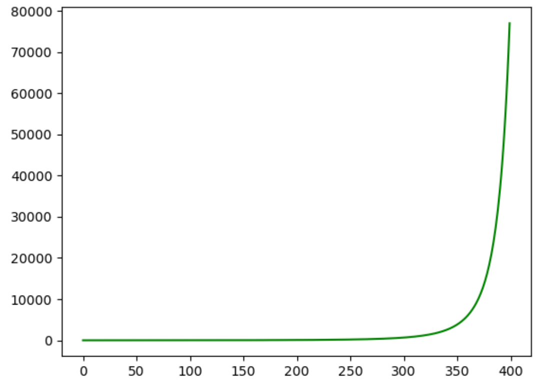
Figure 1: The variance, depending on θ, of the policy gradient without return-to-go (in blue) and with return to go (in orange).



4. (a) We have

$$\nabla_\theta J\left(\theta\right) = \mathbb{E}_{\tau\sim p_{\theta'}(\tau)}\left[\frac{\prod_{t=1}^{H}\pi_\theta\left(a_t\mid s_t\right)}{\prod_{t=1}^{H}\pi_{\theta'}\left(a_t\mid s_t\right)}\nabla_\theta\log p_\theta\left(\tau\right)R\left(\tau\right)\right].$$

Figure 2: The reduction in variance gained by using return-to-go, as a function of θ.



Let $\tau_0 = (s_1, \dots, s_H)$. Because $R = \chi_{\tau_0}$, we get

$$\begin{aligned}
\nabla_\theta J(\theta) &= p_{\theta'}(\tau_0) \frac{p_\theta(\tau_0)}{p_{\theta'}(\tau_0)} \nabla_\theta \log p_\theta(\tau_0) \\
&= p_\theta(\tau_0) \nabla_\theta \log p_\theta(\tau_0) \\
&= \nabla_\theta p_\theta(\tau_0) \\
&= \nabla_\theta \prod_{t=1}^{H-1} p_\theta(a_1 \mid s_t).
\end{aligned}$$

(b) We have

$$\begin{aligned}
\mathrm{Var}_{\tau\sim p_{\theta'}(\tau)}\left[\frac{\prod_{t=1}^H \pi_\theta(a_t \mid s_t)}{\prod_{t=1}^H \pi_{\theta'}(a_t \mid s_t)} \nabla_\theta \log p_\theta(\tau) R(\tau)\right] &= \mathbb{E}_{\tau\sim p_{\theta'}(\tau)}\left[\left(\frac{\prod_{t=1}^H \pi_\theta(a_t \mid s_t)}{\prod_{t=1}^H \pi_{\theta'}(a_t \mid s_t)} \nabla_\theta \log p_\theta(\tau) R(\tau)\right)^2\right] \\
&\quad - \mathbb{E}_{\tau\sim p_{\theta'}(\tau)}\left[\frac{\prod_{t=1}^H \pi_\theta(a_t \mid s_t)}{\prod_{t=1}^H \pi_{\theta'}(a_t \mid s_t)} \nabla_\theta \log p_\theta(\tau) R(\tau)\right]^2 \\
&= p_{\theta'}\left(\frac{p_\theta(\tau_0)}{p_{\theta'}(\tau_0)} \nabla_\theta \log p_\theta(\tau_0)\right)^2 - \left(\nabla_\theta \prod_{t=1}^{H-1} p_\theta(a_1 \mid s_t)\right)^2 \\
&= \frac{p_\theta(\tau_0)^2}{p_{\theta'}(\tau_0)} (\nabla_\theta \log p_\theta(\tau_0))^2 - \left(\nabla_\theta \prod_{t=1}^{H-1} p_\theta(a_1 \mid s_t)\right)^2
\end{aligned}$$

Now, $p_\theta(\tau_0)$ grows exponentially in H and therefore so does its derivative. Then $\log p_\theta$ is a linear term in H so its derivative is constant. We get that both of the summands approach zero, as the base of the exponent is smaller than 1, and so the variance goes to 0 as $H \to \infty$.