(Optional) Derivation

Before calculating the expression above, we will need to further simplify $\nabla_\theta \log \mathbb{P}(\tau^{(i)};\theta)$. The derivation proceeds as follows:

$$\begin{aligned}
\nabla_\theta \log \mathbb{P}(\tau^{(i)};\theta) &= \nabla_\theta \log \left[\prod_{t=0}^{H} \mathbb{P}(s_{t+1}^{(i)}|s_t^{(i)}, a_t^{(i)}) \pi_\theta(a_t^{(i)}|s_t^{(i)}) \right] \\
&= \nabla_\theta \left[\sum_{t=0}^{H} \log \mathbb{P}(s_{t+1}^{(i)}|s_t^{(i)}, a_t^{(i)}) + \sum_{t=0}^{H} \log \pi_\theta(a_t^{(i)}|s_t^{(i)}) \right] \\
&= \nabla_\theta \sum_{t=0}^{H} \log \mathbb{P}(s_{t+1}^{(i)}|s_t^{(i)}, a_t^{(i)}) + \nabla_\theta \sum_{t=0}^{H} \log \pi_\theta(a_t^{(i)}|s_t^{(i} \\
&= \nabla_\theta \sum_{t=0}^{H} \log \pi_\theta(a_t^{(i)}|s_t^{(i)}) \\
&= \sum_{t=0}^{H} \nabla_\theta \log \pi_\theta(a_t^{(i)}|s_t^{(i)})
\end{aligned}$$

First, line (1) shows how to calculate the probability of an arbitrary trajectory $\tau^{(i)}$. Namely, $\mathbb{P}(\tau^{(i)};\theta) = \prod_{t=0}^{H} \mathbb{P}(s_{t+1}^{(i)}|s_t^{(i)}, a_t^{(i)}) \pi_\theta(a_t^{(i)}|s_t^{(i)})$, where we have to take into account the action-selection probabilities from the policy and the state transition dynamics of the MDP.

Then, line (2) follows from the fact that the log of a product is equal to the sum of the logs.

Then, line (3) follows because the gradient of the sum can be written as the sum of gradients.

Next, line (4) holds, because $\sum_{t=0}^{H} \log \mathbb{P}(s_{t+1}^{(i)}|s_t^{(i)}, a_t^{(i)})$ has no dependence on θ, so $\nabla_\theta \sum_{t=0}^{H} \log \mathbb{P}(s_{t+1}^{(i)}|s_t^{(i)}, a_t^{(i)}) = 0$.

Finally, line (5) holds, because we can rewrite the gradient of the sum as the sum of gradients.