(Optional) Derivation

$$\nabla_\theta U(\theta) = \nabla_\theta \sum_\tau P(\tau;\theta) R(\tau) \quad (1)$$

$$= \sum_\tau \nabla_\theta P(\tau;\theta) R(\tau) \quad (2)$$

$$= \sum_\tau \frac{P(\tau;\theta)}{P(\tau;\theta)} \nabla_\theta P(\tau;\theta) R(\tau) \quad (3)$$

$$= \sum_\tau P(\tau;\theta) \frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta)} R(\tau) \quad (4)$$

$$= \sum_\tau P(\tau;\theta) \nabla_\theta \log P(\tau;\theta) R(\tau) \quad (5)$$

First, we note line (1) follows directly from $U(\theta) = \sum_\tau P(\tau;\theta) R(\tau)$, where we've only taken the gradient of both sides.

Then, we can get line (2) by just noticing that we can rewrite the gradient of the sum as the sum of the gradients.

In line (3), we only multiply every term in the sum by $\frac{P(\tau;\theta)}{P(\tau;\theta)}$, which is perfectly allowed because this fraction is equal to one!

Next, line (4) is just a simple rearrangement of the terms from the previous line. That is, $\frac{P(\tau;\theta)}{P(\tau;\theta)} \nabla_\theta P(\tau;\theta) = P(\tau;\theta) \frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta)}$.

Finally, line (5) follows from the chain rule, and the fact that the gradient of the log of a function is always equal to the gradient of the function, divided by the function. (*In case it helps to see this with simpler notation, recall that* $\nabla_x \log f(x) = \frac{\nabla_x f(x)}{f(x)}$.) Thus,