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, $\nabla_\theta \log P(\tau;\theta) = \frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta)}$.

The final "trick" that yields line (5) (i.e., $\nabla_\theta \log P(\tau;\theta) = \frac{\nabla_\theta P(\tau;\theta)}{P(\tau;\theta)}$) is referred to as the **likelihood ratio trick** or **REINFORCE trick**.

Likewise, it is common to refer to the gradient as the **likelihood ratio policy gradient**:

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

Once we've written the gradient as an expected value in this way, it becomes much easier to estimate.

Sample-Based Estimate

In the video on the previous page, you learned that we can approximate the likelihood ratio policy gradient with a sample-based average, as shown below:

$$\nabla_\theta U(\theta) \approx \frac{1}{m}\sum_{i=1}^{m} \nabla_\theta \log \mathbb{P}(\tau^{(i)};\theta)R(\tau^{(i)})$$

where each $\tau^{(i)}$ is a sampled trajectory.