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.

Finishing the Calculation

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

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