Policy Gradient Derivation

Alexander Schiendorfer

January 21, 2026

Here we derive the policy gradient (Equation 2.5) from the gradient of the objective shown in Equation 2.6, a first reading. Stuff we want to express the gradient in terms of:

$$\begin{aligned}
\pi_\theta(a \mid s) &\overset{\text{def}}{=} \mathrm{P}_{\pi_\theta}(A_t = a \mid S_t = s) \\
J(\pi_\theta) &\overset{\text{def}}{=} \mathbb{E}_{x \sim \pi_\theta}[R(\tau)] \\
\tau &\overset{\text{def}}{=} \text{trajectory } (s_0, a_0, s_1, a_1, \ldots, s_T, a_T)
\end{aligned}$$

Note that:

- τ is a trajectory, a sequence of states and actions
- $R(\tau) = \sum_{t=0}^{T} \gamma^t r_t$ is the (discounted) return for trajectory τ
- $p(\tau \mid \theta)$ is the probability of trajectory τ under policy π_θ
- The reward function $R(\tau)$ does not depend on the policy parameters θ

Equation 2.6 presents a problem because we cannot differentiate $R(\tau)$ with respect to θ: The reward, in turn, is changed by an action which, in turn, changes the unknown transition function $\mathcal{R}(s_t, a_t, s_{t+1})$ which cannot be differentiated. The only way for the policy variables θ to influence $R(\tau)$ is by changing $p(\tau \mid \theta)$ and action distributions which have been received by an agent.

We therefore want to transform Equation 2.6 into a form where we can take a gradient with respect to θ. To do so, we'll use some handy identities.

Given a function $f(x)$ and a parametric probability distribution $p(x \mid \theta)$, and its expectation $\mathbb{E}_{x \sim p(\cdot \mid \theta)}[f(x)]$, the gradient of the expectation can be rewritten as follows:

Definition of expectation as an integral (or sum for discrete x):

$$\nabla_\theta \mathbb{E}_{x \sim p(\cdot \mid \theta)}[f(x)] = \nabla_\theta \int dx\, f(x) p(x \mid \theta) \tag{1}$$

Bring ∇_θ inside the integral (derivative and integral can be exchanged under mild regularity conditions):

$$= \int dx\, \nabla_\theta (p(x \mid \theta) f(x)) \tag{2}$$

Apply the product rule (chain rule) to the gradient of the product:

$$= \int dx\, (f(x) \nabla_\theta p(x \mid \theta) + p(x \mid \theta) \nabla_\theta f(x)) \tag{3}$$

The reward function $f(x)$ does not depend on θ, so $\nabla_\theta f(x) = 0$:

$$= \int dx\, f(x) \nabla_\theta p(x \mid \theta) \tag{4}$$

Apply the log derivative trick by multiplying and dividing by $p(x \mid \theta)$:[1]

$$= \int dx\, f(x) p(x \mid \theta) \frac{\nabla_\theta p(x \mid \theta)}{p(x \mid \theta)} \tag{5}$$

[1] This is sometimes called the REINFORCE trick or the likelihood ratio trick. The key insight is that $\frac{\nabla_\theta p(x|\theta)}{p(x|\theta)} = \nabla_\theta \log p(x \mid \theta)$ since $\nabla_\theta \log p(x \mid \theta) = \frac{1}{p(x|\theta)} \nabla_\theta p(x \mid \theta)$.

Substitute $\frac{\nabla_\theta p(x|\theta)}{p(x|\theta)} = \nabla_\theta \log p(x \mid \theta)$ (Equation 2.14 in the original):

$$= \int dx\, f(x) p(x \mid \theta) \nabla_\theta \log p(x \mid \theta) \tag{6}$$

Recognize this as an expectation (definition of expectation):

$$= \mathbb{E}_{x \sim p(\cdot|\theta)}[f(x) \nabla_\theta \log p(x \mid \theta)] \tag{7}$$

Now we apply this result to our specific case where $f(x) = R(\tau)$ and $p(x \mid \theta) = p(\tau \mid \theta)$.
Substituting Equation 2.14 into 2.11 gives Equation 2.12. This can be written as an expectation to give Equation 2.13. Finally, we simply rewrite the expression as an expectation.
Now, it should be apparent that this identity can be applied to our objective. By substitution $x = \tau, f(\tau) = R(\tau)$ and $p(x \mid \theta) = p(\tau \mid \theta)$, Equation 2.6 can be written as Equation 2.15:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau) \nabla_\theta \log p(\tau \mid \theta)] \tag{8}$$

However, the term $p(\tau \mid \theta)$ in Equation 2.15 needs to relate to the policy π_θ which we have control over. Therefore, it needs to be expanded further.
Observe that the trajectory τ is a particular sequence of state-action pairs, which are sampled, respectively, from the agent's action probability $\pi_\theta(a_t \mid s_t)$ and the environment's transition probability $p(s_{t+1} \mid s_t, a_t)$. Since the probabilities are conditionally independent, the probability of the trajectory is the product of these individual probabilities, as shown in Equation 2.16.

$$p(\tau \mid \theta) = \prod_{t \geq 0} p(s_{t+1} \mid s_t, a_t) \pi_\theta(a_t \mid s_t) \tag{9}$$

Apply logarithms[2] to both sides to match Equation 2.16 with Equation 2.15:

$$\log p(\tau \mid \theta) = \log \prod_{t \geq 0} p(s_{t+1} \mid s_t, a_t) \pi_\theta(a_t \mid s_t) \tag{10}$$

$$\log p(\tau \mid \theta) = \sum_{t \geq 0} \Big(\log p(s_{t+1} \mid s_t, a_t) + \log \pi_\theta(a_t \mid s_t) \Big) \tag{11}$$

Now, taking the gradient of Equation (11) with respect to θ, we notice that the transition probabilities $p(s_{t+1} \mid s_t, a_t)$ do not depend on θ, so their gradient is zero:

$$\nabla_\theta \log p(\tau \mid \theta) = \nabla_\theta \sum_{t \geq 0} \Big(\log p(s_{t+1} \mid s_t, a_t) + \log \pi_\theta(a_t \mid s_t) \Big) \tag{12}$$

$$\nabla_\theta \log p(\tau \mid \theta) = \sum_{t \geq 0} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \tag{13}$$

Substituting this back into Equation (8):

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[R(\tau) \sum_{t \geq 0} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \right] \tag{14}$$

Since $R(\tau) = \sum_{t=0}^{T} r_t$ (or $\sum_{t=0}^{T} \gamma^t r_t$ for the discounted case) is the total return of the trajectory, we can write this more explicitly:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{T} R(\tau) \nabla_\theta \log \pi_\theta(a_t \mid s_t) \right] \tag{15}$$

This is the **simple policy gradient formula**. This is the key result: we've transformed the gradient of an expectation into an expectation of a gradient times the log probability of actions weighted by the total return.
The beauty of this formulation is that:

- We don't need to know the environment dynamics (transition probabilities $p(s_{t+1} \mid s_t, a_t)$)—they vanish when taking the gradient!

[2] Recall that $\log(ab) = \log a + \log b$ and $\log \prod_i x_i = \sum_i \log x_i$.

- We can estimate the gradient just by sampling trajectories from the current policy
- The gradient can be computed using only the policy π_θ and observed returns $R(\tau)$
- This can be implemented as the REINFORCE algorithm: sample trajectories, compute returns, and update θ in the direction of the weighted log probabilities

Important note: This simple version uses the total trajectory return $R(\tau)$ for all time steps. A more sophisticated variant (using the causality principle) would use the "reward-to-go" $\sum_{t'=t}^{T} r_{t'}$ instead, which has lower variance because actions at time t cannot influence rewards received before time t.