Derivation of the Policy Gradient Formula

August 11, 2026

1 Setup

Consider a finite-horizon Markov decision process with horizon T. Let $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces. For simplicity, all probability distributions below are written using densities or probability mass functions; the same derivation can be formulated with probability kernels and Radon–Nikodym derivatives.

Let

$$\tau = (S_0, A_0, S_1, A_1, \ldots, S_{T-1}, A_{T-1}, S_T)$$

be the random trajectory. Its law is determined by

$$\begin{aligned} S_0 &\sim \rho_0, \\ A_t \mid S_t = s &\sim \pi_\theta(\cdot \mid s), \\ S_{t+1} \mid (S_t = s, A_t = a) &\sim P(\cdot \mid s, a), \end{aligned}$$

where π_θ is the parameterized policy. We assume that the initial-state distribution ρ_0 and the transition kernel P do not depend on θ.

Let the one-step reward be $R_t = r(S_t, A_t)$ and define the discounted return of a trajectory by

$$G(\tau) = \sum_{t=0}^{T-1} \gamma^t R_t, \qquad 0 \le \gamma \le 1.$$

The objective is

$$J(\theta) = \mathbb{E}_{\tau \sim p_\theta}[G(\tau)],$$

where p_θ is the probability law of the entire trajectory induced jointly by the policy and the environment. Thus, the expectation in $J(\theta)$ is over all random states and actions in the trajectory.

The trajectory density factors as

$$p_\theta(\tau) = \rho_0(s_0) \prod_{t=0}^{T-1} \pi_\theta(a_t \mid s_t) P(s_{t+1} \mid s_t, a_t). \tag{1}$$

2 Likelihood-Ratio Representation

Writing the expectation as an integral gives

$$J(\theta) = \int G(\tau) p_\theta(\tau) \, \mathrm{d}\tau.$$

Assume that differentiation may be interchanged with integration. Then

$$\begin{aligned}\nabla_\theta J(\theta) &= \int G(\tau)\nabla_\theta p_\theta(\tau)\,\mathrm{d}\tau \\ &= \int G(\tau)p_\theta(\tau)\nabla_\theta \log p_\theta(\tau)\,\mathrm{d}\tau \\ &= \mathbb{E}_{\tau\sim p_\theta}\left[G(\tau)\nabla_\theta \log p_\theta(\tau)\right].\end{aligned}$$

The second equality uses the score-function, or likelihood-ratio, identity

$$\nabla_\theta p_\theta(\tau) = p_\theta(\tau)\nabla_\theta \log p_\theta(\tau),$$

valid wherever $p_\theta(\tau) > 0$.

Taking logarithms in (1),

$$\begin{aligned}\log p_\theta(\tau) = {} & \log \rho_0(S_0) \\ & + \sum_{t=0}^{T-1} \left(\log \pi_\theta(A_t \mid S_t) + \log P(S_{t+1} \mid S_t, A_t)\right).\end{aligned}$$

Since neither ρ_0 nor P depends on θ,

$$\nabla_\theta \log p_\theta(\tau) = \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(A_t \mid S_t). \tag{2}$$

Consequently,

$$\begin{aligned}\nabla_\theta J(\theta) &= \mathbb{E}_{\tau\sim p_\theta}\left[\left(\sum_{k=0}^{T-1}\gamma^k R_k\right)\left(\sum_{t=0}^{T-1}\nabla_\theta \log \pi_\theta(A_t \mid S_t)\right)\right] \\ &= \sum_{t=0}^{T-1}\mathbb{E}_{\tau\sim p_\theta}\left[\nabla_\theta \log \pi_\theta(A_t \mid S_t)\sum_{k=0}^{T-1}\gamma^k R_k\right]. \end{aligned} \tag{3}$$

3 Removal of Past Rewards

For a fixed time t, define the information available immediately before A_t is sampled by

$$\mathcal{F}_t = \sigma(S_0, A_0, \ldots, A_{t-1}, S_t).$$

The random variable

$$\sum_{k=0}^{t-1}\gamma^k R_k$$

is $\mathcal{F}_t$-measurable. We now show that its contribution to (3) is zero.

Conditionally on $\mathcal{F}_t$, the state S_t is known and the only new randomness in the score term comes from

$$A_t \sim \pi_\theta(\,\cdot \mid S_t).$$

Hence

$$\begin{aligned}&\mathbb{E}_{\tau\sim p_\theta}\left[\nabla_\theta \log \pi_\theta(A_t \mid S_t) \mid \mathcal{F}_t\right] \\ &\quad = \int \pi_\theta(a \mid S_t)\nabla_\theta \log \pi_\theta(a \mid S_t)\,\mathrm{d}a \\ &\quad = \int \nabla_\theta \pi_\theta(a \mid S_t)\,\mathrm{d}a \\ &\quad = \nabla_\theta \int \pi_\theta(a \mid S_t)\,\mathrm{d}a = 0.\end{aligned}$$

Therefore, by the tower property of conditional expectation,

$$\mathbb{E}_{\tau\sim p_\theta}\left[\nabla_\theta \log \pi_\theta(A_t \mid S_t)\sum_{k=0}^{t-1}\gamma^k R_k\right]$$
$$= \mathbb{E}_{\tau\sim p_\theta}\left[\left(\sum_{k=0}^{t-1}\gamma^k R_k\right)\mathbb{E}_{\tau\sim p_\theta}\left[\nabla_\theta \log \pi_\theta(A_t \mid S_t) \mid \mathcal{F}_t\right]\right] = 0.$$

Thus only rewards at time t and later remain:

$$\nabla_\theta J(\theta) = \sum_{t=0}^{T-1}\mathbb{E}_{\tau\sim p_\theta}\left[\nabla_\theta \log \pi_\theta(A_t \mid S_t)\sum_{k=t}^{T-1}\gamma^k R_k\right]. \tag{4}$$

Define the return measured from time t by

$$G_t = \sum_{\ell=0}^{T-1-t}\gamma^\ell R_{t+\ell}.$$

Since

$$\sum_{k=t}^{T-1}\gamma^k R_k = \gamma^t G_t,$$

we obtain

$$\boxed{\nabla_\theta J(\theta) = \sum_{t=0}^{T-1}\gamma^t \mathbb{E}_{\tau\sim p_\theta}\left[\nabla_\theta \log \pi_\theta(A_t \mid S_t) G_t\right].} \tag{5}$$

This is the REINFORCE form of the policy gradient.

4 Conditional Expectation and the Action-Value Function

Define the finite-horizon action-value function at time t by

$$Q_t^{\pi_\theta}(s,a) = \mathbb{E}_{\pi_\theta}\left[G_t \mid S_t = s, A_t = a\right]. \tag{6}$$

The conditional expectation in (6) is taken over the future randomness after the state-action pair $(S_t, A_t) = (s, a)$ is fixed. In particular, it averages over future state transitions and over future actions sampled from π_θ.

Since

$$\nabla_\theta \log \pi_\theta(A_t \mid S_t)$$

is measurable with respect to $\sigma(S_t, A_t)$, the tower property gives

$$\mathbb{E}_{\tau\sim p_\theta}\left[\nabla_\theta \log \pi_\theta(A_t \mid S_t) G_t\right]$$
$$= \mathbb{E}_{\tau\sim p_\theta}\left[\nabla_\theta \log \pi_\theta(A_t \mid S_t)\mathbb{E}_{\tau\sim p_\theta}[G_t \mid S_t, A_t]\right]$$
$$= \mathbb{E}_{(S_t,A_t)\sim p_{\theta,t}}\left[\nabla_\theta \log \pi_\theta(A_t \mid S_t) Q_t^{\pi_\theta}(S_t, A_t)\right],$$

where $p_{\theta,t}$ denotes the marginal law of (S_t, A_t) under the trajectory distribution p_θ.

Let $d_t^{\pi_\theta}$ denote the marginal law of S_t. Then $p_{\theta,t}$ has the factorization

$$p_{\theta,t}(\mathrm{d}s, \mathrm{d}a) = d_t^{\pi_\theta}(\mathrm{d}s)\,\pi_\theta(\mathrm{d}a \mid s),$$

and, for every integrable function f,

$$\mathbb{E}_{(S_t,A_t)\sim p_{\theta,t}}[f(S_t, A_t)]$$
$$= \int_{\mathcal{S}} d_t^{\pi_\theta}(\mathrm{d}s)\int_{\mathcal{A}}\pi_\theta(\mathrm{d}a \mid s) f(s,a).$$

Thus the policy gradient becomes

$$\boxed{\nabla_\theta J(\theta) = \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{(S_t,A_t)\sim p_{\theta,t}} \left[\nabla_\theta \log \pi_\theta(A_t \mid S_t) Q_t^{\pi_\theta}(S_t, A_t)\right].} \tag{7}$$

The expectation in (7) is over two levels of randomness: first S_t is distributed according to the state distribution generated by the policy and environment up to time t; then, conditional on $S_t = s$, the action is sampled from $\pi_\theta(\cdot \mid s)$.

5 State-Dependent Baselines and Advantage

Let $b_t : \mathcal{S} \to \mathbb{R}$ be any integrable function that does not depend on the sampled action. We have

$$\begin{aligned}&\mathbb{E}_{(S_t,A_t)\sim p_{\theta,t}} \left[\nabla_\theta \log \pi_\theta(A_t \mid S_t) b_t(S_t)\right] \\ &\quad = \mathbb{E}_{S_t \sim d_t^{\pi_\theta}} \left[b_t(S_t) \mathbb{E}_{A_t \sim \pi_\theta(\cdot|S_t)} \left[\nabla_\theta \log \pi_\theta(A_t \mid S_t)\right]\right] = 0.\end{aligned}$$

Therefore subtracting a state-dependent baseline does not change the expected policy gradient:

$$\begin{aligned}&\mathbb{E}_{(S_t,A_t)\sim p_{\theta,t}} \left[\nabla_\theta \log \pi_\theta(A_t \mid S_t) Q_t^{\pi_\theta}(S_t, A_t)\right] \\ &\quad = \mathbb{E}_{(S_t,A_t)\sim p_{\theta,t}} \left[\nabla_\theta \log \pi_\theta(A_t \mid S_t) \left(Q_t^{\pi_\theta}(S_t, A_t) - b_t(S_t)\right)\right].\end{aligned}$$

Define the state-value function by

$$V_t^{\pi_\theta}(s) = \mathbb{E}_{A\sim\pi_\theta(\cdot|s)} \left[Q_t^{\pi_\theta}(s, A)\right]$$

and the advantage function by

$$A_t^{\pi_\theta}(s, a) = Q_t^{\pi_\theta}(s, a) - V_t^{\pi_\theta}(s).$$

Choosing $b_t = V_t^{\pi_\theta}$ in (7) yields

$$\boxed{\nabla_\theta J(\theta) = \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{(S_t,A_t)\sim p_{\theta,t}} \left[\nabla_\theta \log \pi_\theta(A_t \mid S_t) A_t^{\pi_\theta}(S_t, A_t)\right].} \tag{8}$$

The baseline changes the variance of a Monte Carlo gradient estimator but not its expectation. The value function is a particularly natural baseline because the advantage measures the value of the sampled action relative to the policy's average action value at the same state.

6 Infinite-Horizon Discounted Form

Now suppose $T = \infty$ and $0 \le \gamma < 1$. Under standard integrability and regularity assumptions allowing the preceding exchanges of infinite sums, expectations, and derivatives, define the normalized discounted state-occupancy measure

$$d_{\pi_\theta}^{\gamma} = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t \operatorname{Law}_{\pi_\theta}(S_t). \tag{9}$$

This is a probability measure because each $\mathrm{Law}_{\pi_\theta}(S_t)$ has total mass one and

$$(1-\gamma)\sum_{t=0}^{\infty}\gamma^t = 1.$$

If $S \sim d_{\pi_\theta}^{\gamma}$ and, conditionally on $S = s$, $A \sim \pi_\theta(\,\cdot \mid s)$, then

$$\boxed{\nabla_\theta J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{\substack{S\sim d_{\pi_\theta}^{\gamma} \\ A\sim\pi_\theta(\cdot|S)}}\left[\nabla_\theta \log \pi_\theta(A \mid S) A^{\pi_\theta}(S,A)\right].} \tag{10}$$

Equation (10) makes precise the common shorthand

$$\nabla_\theta J(\theta) \propto \mathbb{E}_{s,a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a \mid s) A^{\pi_\theta}(s,a)\right].$$

The notation $s, a \sim \pi_\theta$ suppresses two distinct distributions. The state is distributed according to the discounted occupancy measure induced by repeated interaction of π_θ with the environment, while the action is then sampled conditionally from $\pi_\theta(\,\cdot \mid s)$.