proof of (2):

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\Bigg[\sum_{t=0}^{H} \gamma^t \, \nabla_\theta \log \pi_\theta(a_t \mid s_t) \big(\sum_{t'=t}^{H} \gamma^{\,t'-t} r(s_{t'}, a_{t'}) - b(s_t)\big)\Bigg]$$

This form shows the policy gradient as a sum over actions, each scaled by an advantage estimate: a discounted future reward sum minus a baseline.

Maximize

We want to maximize expected return:

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\big[R(\tau)\big], \quad R(\tau) = \sum_{t=0}^{H} \gamma^t \, r(s_t, a_t), \quad \tau = (s_0, a_0, \dots, s_H, a_H).$$

Use the log-derivative trick

$$\nabla_\theta J(\theta) = \nabla_\theta \, \mathbb{E}_{\tau\sim\pi_\theta}\big[R(\tau)\big] = \mathbb{E}_{\tau\sim\pi_\theta}\big[\nabla_\theta \log p_{\pi_\theta}(\tau) \, R(\tau)\big].$$

proof of the trick:

$$\nabla_\theta p(\tau;\theta) = p(\tau;\theta) \, \nabla_\theta \log p(\tau;\theta) \tag{1}$$

$$\begin{aligned}
\nabla_\theta \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)] &= \nabla_\theta \int p_{\pi_\theta}(\tau) \, R(\tau) \, \mathrm{d}\tau && (2)\\
&= \int \nabla_\theta p_{\pi_\theta}(\tau) \, R(\tau) \, \mathrm{d}\tau && (3)\\
&= \int p_{\pi_\theta}(\tau) \, \nabla_\theta \log p_{\pi_\theta}(\tau) \, R(\tau) \, \mathrm{d}\tau && (4)\\
&= \mathbb{E}_{\tau\sim\pi_\theta}\big[\nabla_\theta \log p_{\pi_\theta}(\tau) \, R(\tau)\big] && (5)\\
&= \nabla_\theta J(\theta) \, . && (6)
\end{aligned}$$

Decompose $\log p_{\pi_\theta}(\tau)$**:**

$$p(\tau) = \rho_0(s_0) \prod_{t=0}^{H} \pi_\theta(a_t \mid s_t) \, P(s_{t+1} \mid s_t, a_t),$$

so

$$\nabla_\theta \log p_{\pi_\theta}(\tau) = \sum_{t=0}^{H} \nabla_\theta \log \pi_\theta(a_t \mid s_t).$$

Plug in:

$$\nabla_\theta J(\theta) = \mathbb{E}_\tau\Big[\Big(\sum_{t=0}^{H} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\Big)\, R(\tau)\Big] = \sum_{t=0}^{H} \mathbb{E}_\tau\big[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\, R(\tau)\big].$$

Reward-to-Go (Causality)

Only future rewards depend on a_t. Dropping past rewards: (mathematically if you write out the integral terms that do not depend on the integrator factor out and you are left with the gradient of integral of a probability which is the constant 1 and hence zero)

$$\nabla_\theta J(\theta) = \sum_{t=0}^{H} \mathbb{E}_\tau\Big[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\, \underbrace{\sum_{t'=t}^{H} \gamma^{t'} r(s_{t'}, a_{t'})}_{\text{reward-to-go}}\Big].$$

Factor out γ^t:

$$= \sum_{t=0}^{H} \mathbb{E}_\tau\Big[\gamma^t\, \nabla_\theta \log \pi_\theta(a_t \mid s_t)\, \sum_{t'=t}^{H} \gamma^{t'-t} r(s_{t'}, a_{t'})\Big].$$

Add a Baseline (As to why you can do this and what a baseline is see pp 329 of Sutton)

Subtract a baseline $b(s_t)$:

$$\nabla_\theta J(\theta) = \sum_{t=0}^{H} \mathbb{E}_\tau\Big[\gamma^t\, \nabla_\theta \log \pi_\theta(a_t \mid s_t)\, \Big(\sum_{t'=t}^{H} \gamma^{t'-t} r(s_{t'}, a_{t'}) - b(s_t)\Big)\Big].$$

In the finite-horizon derivation we write

$$\nabla_\theta = \sum_{t=0}^{H} \mathbb{E}_{s_t \sim d_t^\pi,\, a_t \sim \pi_\theta}\big[\gamma^t\, \nabla_\theta \log \pi_\theta(a_t \mid s_t)\, \hat{A}(s_t, a_t)\big].$$

Here both the state-distribution d_t^π and the discount weight γ^t are explicitly indexed by t.

When passing to the infinite-horizon form, those two pieces are folded into a single "discounted occupancy" measure

$$d^\pi(s) \;=\; (1-\gamma) \sum_{t=0}^{\infty} \gamma^t\, d_t^\pi(s),$$

so that the gradient can be written as

$$\nabla_\theta J = \frac{1}{1-\gamma}\, \mathbb{E}_{s \sim d^\pi,\, a \sim \pi_\theta}\big[\nabla_\theta \log \pi_\theta(a \mid s)\, \hat{A}(s, a)\big].$$

The stray subscript "$_t$" on the state distribution in the infinite-horizon equation was simply a leftover from the finite-horizon version. The correct infinite-horizon line should read

$$\nabla_\theta J = \frac{1}{1-\gamma}\, \mathbb{E}_{s\sim d^\pi(s),\, a\sim\pi_\theta(a|s)}\big[\nabla_\theta \log \pi_\theta(a \mid s)\, \hat{A}(s,a)\big],$$

with no time index on d^π.

While a simple time-dependent baseline $b_t = \frac{1}{N}\sum_{i=1}^{N} R_{i,t}$ is often used to reduce variance in Monte Carlo policy gradient estimators, it does not represent a true value function $V(s_t) = \mathbb{E}[R_t \mid s_t]$. This is because b_t marginalizes over all states s_t encountered at time t rather than conditioning on a specific state. Thus, it captures only the average return under the state visitation distribution $d_t(s)$, not the expected return from a particular state $s_t = s$. Although useful for variance reduction, this approach lacks the precision and generalization capabilities of a learned, state-dependent baseline.