Policy Gradient Methods: REINFORCE

Brandyn Tucknott

30 September 2025

REINFORCE (sometimes called Monte-Carlo Policy Gradient) is a policy gradient method based on the identity for a policy gradient

$$\nabla_\theta J(\theta) = \mathbf{E}_{\pi_\theta}\left(\sum_{t\in 0:T} \nabla_\theta \ln \pi_\theta\left(A_t|s_t\right) \sum_{t\in 0:T}\left(\gamma^t R_t | S_0 = s_0\right)\right).$$

The **unbiased estimator** of the policy gradient can be written as

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{n=1}^{N}\left[\sum_{t\in 0:T} \nabla_\theta \ln \pi_\theta\left(A_{t,n}|S_{t,n}\right) \sum_{\tau\in t:T}\left(\gamma^{\tau-t} R_{\tau,n}\right)\right].$$

The **score function** $\nabla_\theta \ln \pi_\theta\left(A_t|S_t\right)$ as the direction in parameter space which increases the probability of taking action A_t in state S_t. The policy gradient is the weighted average of all possible directions with all possible actions at any state, weighted by reward signals. This means that state-action pairs with a high reward are reinforced.

Algorithm 1 REINFORCE

Input: differentiable policy parameterization $\pi(a|s,\theta)$
Hyperparameters:

- Learning rate $\alpha > 0$

Initialize the policy parameter θ at random

for each episode: **do**
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$ following $\pi(\cdot|\cdot,\theta)$.
  for each step of the episode $t = 0, 1, 2, \ldots, T-1$: **do**
    $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ ($G =$ discounted reward sum)
    $\theta \leftarrow \theta + \alpha\gamma^t G \nabla \ln \pi\left(A_T|S_t,\theta\right)$ (modify policy parameters θ)