ESE 6510 Note 2: Variance Reduction

Jefferson Ng

Feb 1st, 2026

1 Variance in Policy Gradient Estimation

Although the policy gradient

$$\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\sum_{t=0}^{T-1}\nabla_\theta \log \pi_\theta(a_t \mid s_t)\, R(\tau)\right]$$

is unbiased, it typically suffers from high variance. Since policy updates are performed using stochastic gradient ascent, noisy gradients can lead to unstable learning and slow convergence.

One major source of variance is that early policy gradient terms are multiplied by the total return of the entire trajectory. In long-horizon tasks, stochastic effects far in the future can significantly alter the total return, causing large fluctuations in the gradient contributions of early actions.

To make this dependence explicit, consider the gradient of the expected reward at a fixed timestep t:

$$\nabla_\theta\, \mathbb{E}_{\tau\sim p_\theta(\tau)}[r_t].$$

Applying the score function identity yields (using $T-1$ since no reward is received after the terminal transition)

$$\nabla_\theta \mathbb{E}[r_t] = \mathbb{E}\left[\sum_{j=0}^{T-1}\nabla_\theta \log \pi_\theta(a_j \mid s_j)\, r_t\right].$$

However, the reward r_t cannot depend on actions taken after time t. As a result, the corresponding terms contribute zero in expectation, and the expression above describes the contribution of a *single timestep* t:

$$\nabla_\theta \mathbb{E}[r_t] = \mathbb{E}\left[\sum_{j=0}^{t}\nabla_\theta \log \pi_\theta(a_j \mid s_j)\, r_t\right].$$

We now consider the entire trajectory by summing the contributions of all rewards $r_0, \ldots, r_{T-1}$:

$$\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}\left[\sum_{t=0}^{T-1} r_t\right] = \mathbb{E}\left[\sum_{t=0}^{T-1}\sum_{j=0}^{t}\nabla_\theta \log \pi_\theta(a_j \mid s_j)\, r_t\right].$$

Reordering the summations yields

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \underbrace{\sum_{j=t}^{T-1} r_j}_{\text{reward-to-go}}\right].$$

This form shows that each policy gradient term is weighted only by rewards that occur at or after the corresponding timestep. Removing dependence on unrelated future rewards preserves the expected gradient while reducing unnecessary variance.

The reward-to-go construction is one example of modifying the return signal to reduce variance. More generally, we may ask whether other transformations of the return can reduce variance without changing the expected policy gradient.

1.1 Why Baselines Do Not Change the Expected Policy Gradient

From the previous section, we ended with the reward-to-go policy gradient:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \underbrace{\sum_{j=t}^{T-1} r_j}_{\text{reward-to-go}}\right].$$

Now consider subtracting a state-dependent baseline inside the same bracket:

$$\nabla_\theta J_b(\theta) = \mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \left(\sum_{j=t}^{T-1} r_j - b(s_t)\right)\right]. \tag{1}$$

Expanding the product inside the sum gives

$$\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \sum_{j=t}^{T-1} r_j - \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\, b(s_t).$$

Using linearity of expectation,

$$\mathbb{E}[X - Y] = \mathbb{E}[X] - \mathbb{E}[Y], \qquad \mathbb{E}\left[\sum_t Z_t\right] = \sum_t \mathbb{E}[Z_t],$$

we obtain

$$\nabla_\theta J_b(\theta) = \nabla_\theta J(\theta) - \mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\, b(s_t)\right].$$

Thus, subtracting a baseline does not change the expected gradient provided that

$$\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\, b(s_t)\right] = 0.$$

It therefore suffices to show that for each timestep t,

$$\mathbb{E}_{\tau\sim p_\theta(\tau)}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\, b(s_t)\right] = 0.$$

We now split the trajectory into its prefix and suffix using the law of total expectation:

$$\mathbb{E}_\tau\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\, b(s_t)\right] = \mathbb{E}_{s_{0:t},\, a_{0:t-1}}\left[\mathbb{E}_{s_{t+1:T},\, a_{t:T}|s_{0:t},\, a_{0:t-1}}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\, b(s_t)\right]\right].$$

The score term depends only on (s_t, a_t), and the baseline depends only on s_t. Conditioned on the prefix $(s_{0:t}, a_{0:t-1})$, the only remaining randomness inside the inner expectation is the action $a_t \sim \pi_\theta(\cdot \mid s_t)$. Hence the inner expectation reduces to

$$= \mathbb{E}_{s_{0:t},\, a_{0:t-1}}\left[b(s_t)\, \mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t)}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right]\right].$$

Finally,

$$\mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t)}[\nabla_\theta \log \pi_\theta(a_t \mid s_t)] = \sum_{a_t} \nabla_\theta \pi_\theta(a_t \mid s_t) = \nabla_\theta \sum_{a_t} \pi_\theta(a_t \mid s_t) = \nabla_\theta 1 = 0.$$

Substituting this back yields

$$\mathbb{E}_\tau\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\, b(s_t)\right] = 0,$$

so subtracting a state-dependent baseline does not change the expected policy gradient.

Why $b(s_t) = \mathbb{E}_{a_t}[f(s_t, a_t)]$ is still a valid baseline

A common question is whether a baseline of the form

$$b(s_t) = \mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t)}[f(s_t, a_t)]$$

is allowed, since f depends on the action.
Once the expectation over a_t is taken, $b(s_t)$ becomes a deterministic function of s_t only, and therefore can be treated as constant inside the inner expectation over a_t in the unbiasedness proof.
In particular,

$$V^\pi(s_t) = \mathbb{E}_{a_t\sim\pi_\theta(\cdot|s_t)}[Q^\pi(s_t, a_t)]\,,$$

so $V^\pi(s_t)$ depends only on s_t and is therefore a valid baseline.

1.2 Why Subtracting a Baseline Reduces Variance

Having established that subtracting a state-dependent baseline preserves the expected policy gradient, we now examine how the choice of baseline affects the *variance* of the estimator.

Let's consider a single summand of the policy gradient Eq. 1

$$\nabla_\theta J_b^t(\theta) = \mathbb{E}_{(s_t,a_t)}\left[\nabla_\theta \log \pi_\theta(a_t \mid s_t)\left(\sum_{j=t}^{T-1} r_j - b(s_t)\right)\right], \tag{2}$$

and study the dependence between its variance and the baseline. To simplify the notation, we define

$$\ell_t \triangleq \nabla_\theta \log \pi_\theta(a_t \mid s_t), \qquad R_t \triangleq \sum_{j=t}^{T-1} r_j, \qquad b_t \triangleq b(s_t).$$

The policy gradient at time t is then

$$g_t \triangleq \mathbb{E}_{(s_t,a_t)}\left[l_t(R_t - b_t)\right]$$

The variance of this single-step gradient is

$$\mathrm{Var}[g_t] = \mathbb{E}\big[\ell_t^2(R_t - b)^2\big] - \big(\mathbb{E}[\ell_t R_t]\big)^2,$$

where we used $\mathbb{E}[\ell_t b] = 0$ from the previous section.

Expanding the quadratic term gives

$$\mathbb{E}[\ell_t^2 R_t^2] - 2b\,\mathbb{E}[\ell_t^2 R_t] + b^2\mathbb{E}[\ell_t^2].$$

Differentiating with respect to b and setting the derivative to zero yields the variance-minimizing baseline at this timestep:

$$b^* = \frac{\mathbb{E}[\ell_t^2 R_t]}{\mathbb{E}[\ell_t^2]} \approx \mathbb{E}[R_t] = V^\pi(s_t) \tag{3}$$

Eq. 3 shows that the optimal baseline is the weighted sum of expected returns. That is not too far from the expected return from the current state to the end of the episode, i.e., the value function. This gives us:

$$\boxed{\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t)\left(\sum_{k=t}^{T-1} r_k - V^\pi(s_t)\right)\right]} \tag{4}$$

We generally call the quantity inside the parentheses *the advantage function*:

$$A^\pi(s_t, a_t) = \sum_{k=t}^{T-1} r_k - V^\pi(s_t).$$

Note that the derivation above provides intuition for variance reduction, but is not formal; a more formal treatment can be found in Greensmith et al. (2004).