CSCI 145 Problem 24: Where You Put Zero

July 25, 2026

Submission Instructions

Please upload your work to Gradescope by **11:59pm on DATE**.

- Turn in a single PDF of *handwritten* solutions. Typed (including LaTeX) solutions will not be accepted.
- Problems are grouped into three-week **cycles**. Once per cycle, you will meet with the grader and present the solution to *one* problem from that cycle, selected at random when you arrive — so bring your work and be ready to explain any of them.

Problem 24: Where You Put Zero

CartPole paid +1 for every step the pole stayed up. Suppose the simulator's author had felt generous and tossed a flat +100 bonus into every episode's return. Nothing about the task changes: the same pushes balance the pole, every policy keeps its rank, and the gradient we should follow points exactly where it pointed before. And yet, hand the demo those rewards and REINFORCE barely improves on a random policy. This problem catches the culprit on the smallest reinforcement learning problem in existence, a slot machine with two arms, where every expectation, gradient, and variance can be computed exactly by enumerating the only two things that can happen. In class we derived the log-derivative trick and the policy gradient

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim\pi_{\boldsymbol{\theta}}}\Big[G(\tau)\sum_{t\geq 0}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a_t|s_t)\Big];$$

here you will run that machinery by hand and measure the noise it produces, picking up the variance-reduction arc from Problem 2 one lecture before it pays off.

Setup. There is one state and two actions: pull arm 1 and receive reward r_1, or pull arm 2 and receive reward r_2 (both deterministic, with $r_1 \neq r_2$). An episode is a single pull, so the return is just the reward received, written R, and the discount never enters. The policy is a softmax over the two arms with a single scalar parameter θ, logits $(\theta, 0)$:

$$\pi_\theta(1) = \frac{e^\theta}{e^\theta + 1} =: p, \qquad \pi_\theta(2) = \frac{1}{e^\theta + 1} = 1 - p.$$

(This is the two-action softmax, which you may recognize as the sigmoid from Logistic Regression.) The objective is the expected reward, $J(\theta) = \mathbb{E}[R]$.

Part A: The Exact Gradient, by Enumeration

Only two outcomes exist, so the expectation is a two-term sum:

$$J(\theta) = p\, r_1 + (1 - p)\, r_2.$$

First show that $\frac{dp}{d\theta} = p(1-p)$, either by differentiating $p = e^\theta/(e^\theta + 1)$ directly with the quotient rule or by recalling the sigmoid derivative from the Logistic Regression lecture. Then differentiate J to obtain the *exact* policy gradient

$$\nabla_\theta J(\theta) = p(1-p)(r_1 - r_2).$$

Sanity-check the formula in words: why is the sign of the gradient the sign of $r_1 - r_2$, and why does the gradient vanish as $p \to 0$ or $p \to 1$ even when one arm is strictly better? Finally, evaluate it at $\theta = 0$ (so $p = \frac{1}{2}$) with rewards $r_1 = 1$, $r_2 = 0$: you should get $\nabla_\theta J = \frac{1}{4}$. Keep this number; Parts B and C measure how well REINFORCE estimates it.

Part B: The Single-Sample Estimator Is Unbiased

A real agent cannot enumerate; it pulls one arm and updates. The single-sample REINFORCE estimator from lecture is

$$\widehat{g} = R\,\nabla_\theta \log\pi_\theta(a), \qquad a \sim \pi_\theta.$$

First compute the two possible score terms, $\nabla_\theta \log \pi_\theta(1) = 1 - p$ and $\nabla_\theta \log \pi_\theta(2) = -p$. (These are the in-class "softmax minus one-hot" result in miniature: only the first logit carries θ, so the gradient is the first entry of one-hot minus softmax.) Conclude that $\widehat{g}$ takes exactly two values:

$$\widehat{g} = \begin{cases} r_1(1-p) & \text{with probability } p, \\ -r_2\, p & \text{with probability } 1-p. \end{cases}$$

Now verify unbiasedness by enumerating both outcomes: show

$$\mathbb{E}[\widehat{g}] = p \cdot r_1(1-p) + (1-p) \cdot (-r_2\, p) = p(1-p)(r_1 - r_2) = \nabla_\theta J(\theta),$$

the exact gradient from Part A, for *every* θ, r_1, r_2. Then look at the two values themselves at $\theta = 0$, $r_1 = 1$, $r_2 = 0$: the estimator says $\frac{1}{2}$ half the time and 0 half the time. It is right *on average*, and never right on any individual pull. Unbiasedness, as in Lecture 2, is only half the story; the other half is variance.

Part C: The Scandal

Work at $\theta = 0$ (so $p = \frac{1}{2}$) with rewards $r_1 = 1$, $r_2 = 0$ throughout this part.

First compute the estimator's variance from Part B's two outcomes:

$$\mathrm{Var}(\widehat{g}) = \mathbb{E}[\widehat{g}^2] - \left(\mathbb{E}[\widehat{g}]\right)^2 = \frac{1}{16}.$$

Now the generous simulator author strikes: add the same constant c to both rewards, so the arms pay $1 + c$ and c. Show three things.

1. **The problem is unchanged:** by Part A, the exact gradient is $p(1-p)\big((1+c) - c\big) = \frac{1}{4}$, the same for every c, so shifting both rewards moves neither the ranking of the arms nor the direction to climb.
2. **The estimator is still unbiased:** its two outcomes become $\frac{1+c}{2}$ and $-\frac{c}{2}$, each with probability $\frac{1}{2}$; enumerate to check the mean is still $\frac{1}{4}$.
3. **The noise explodes:** show

   $$\mathrm{Var}(\widehat{g}) = \frac{(2c+1)^2}{16} = \frac{c^2}{4} + \frac{c}{4} + \frac{1}{16},$$

   growing like c^2. Evaluate at $c = 100$ (the generous author's 101-versus-100 bandit): the variance is $\frac{201^2}{16} \approx 2525$, a standard deviation of about 50, which is two hundred times the gradient of $\frac{1}{4}$ it is trying to estimate. A single pull now reports "move +50.5" or "move −50" when the truth is +0.25.

Conclude in one sentence: what did adding c change about the bandit, and what did it change about our ability to estimate its gradient?

Where this goes next. Look at your variance formula once more: it is a parabola in c with a single minimum, so some shift is *best*, better even than $c = 0$. An estimator whose quality depends on an arbitrary convention (where the reward scale's zero sits) is one we can improve by choosing that convention well, and you already own the tool: Problem 2's control variate, which subtracted a well-chosen quantity to shrink variance without touching the mean. The final lecture subtracts a *baseline* from the return before it multiplies the score, proves the estimator stays unbiased, and derives the variance-minimizing choice. On Problem 25 you will apply it to this very bandit and watch the scandal resolve.