CSCI 145 Problem 25: The Best Baseline, State by State

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 25: The Best Baseline, State by State

Problem 24 ended in scandal: adding the same constant c to both of a bandit's rewards changed nothing about the problem and blew the REINFORCE estimator's variance up like c^2. In class we found the cure. Subtracting a baseline b from the return leaves the estimator unbiased, and the variance-minimizing constant choice is

$$b^\star = \frac{\mathbb{E}[G\,u^2]}{\mathbb{E}[u^2]} = \frac{\mathrm{Cov}(Gu, u)}{\mathrm{Var}(u)}, \qquad u = \nabla_\theta \log \pi_\theta(a) \text{ the score},$$

which is exactly Problem 2's control-variate coefficient $c^\star = \mathrm{Cov}(f, g)/\mathrm{Var}(g)$ with $f = Gu$ and surrogate $g = u$. This is the last problem of the course, and it pushes one step past the lecture. First apply the class formula to Problem 24's bandit and watch the scandal resolve (Part A), then let the baseline depend on the state and derive the true per-state optimum (Part B), and finally recognize the practical version of that optimum as an old friend from Problem 23 (Part C), closing the variance-reduction arc that Problem 2 opened.

Setup (from Problem 24). One state, two arms with deterministic rewards $r_1 \neq r_2$; an episode is a single pull, so the return R is just the reward received. The policy is a softmax with one scalar parameter θ and logits $(\theta, 0)$, so $\pi_\theta(1) = e^\theta/(e^\theta + 1) =: p$ and $\pi_\theta(2) = 1 - p$, with scores $\nabla_\theta \log \pi_\theta(1) = 1 - p$ and $\nabla_\theta \log \pi_\theta(2) = -p$. The baselined single-sample estimator is

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

and the class proof guarantees $\mathbb{E}[\widehat{g}_b] = \nabla_\theta J(\theta)$ for every b.

Part A: The Scandal, Resolved

Work at $\theta = 0$ (so $p = \frac{1}{2}$) with the shifted rewards $r_1 = 1 + c$, $r_2 = c$ from Problem 24, where the estimator's variance was $\frac{(2c+1)^2}{16}$, about 2525 at $c = 100$.

First observe that the two scores are $\frac{1}{2}$ and $-\frac{1}{2}$, so the squared score is $u^2 = \frac{1}{4}$ *whichever* arm is pulled. Use this to evaluate the class formula: show

$$b^\star = \frac{\mathbb{E}[R\,u^2]}{\mathbb{E}[u^2]} = \mathbb{E}[R] = \tfrac{1}{2} + c,$$

so the optimal baseline tracks the generous author's shift exactly.

Now enumerate the two possible values of $\widehat{g}_{b^\star} = (R - b^\star)\,u$ and show that *both* equal $\frac{1}{4}$, the exact gradient from Problem 24 Part A. Conclude that $\mathrm{Var}(\widehat{g}_{b^\star}) = 0$ for every c: the baseline removes the scandal entirely, saving all $\frac{(2c+1)^2}{16}$ of the variance (all ≈ 2525 of it at $c = 100$). Problem 24 ended with an estimator that was right on average and never right on any individual pull; finish Part A with one sentence on what is true now.

Part B: Let the Baseline Depend on the State

A constant worked because the bandit has one state; a real environment has many, and a good baseline in one state is a terrible one in another. So let the baseline be a *function* $b(s)$: condition on the state, let $a \sim \pi_\theta(\cdot|s)$ with score $u(s, a) = \nabla_\theta \log \pi_\theta(a|s)$ and (possibly random) return G, and start by justifying the decomposition

$$\mathrm{Var}(\widehat{g}) = \mathbb{E}_s\Big[\,\mathbb{E}\big[(G - b(s))^2\,u^2 \,\big|\, s\big]\Big] - \big(\nabla_\theta J\big)^2,$$

an average over states of per-state contributions, minus a term that does not depend on b (why not? recall which part of $\mathrm{Var} = \mathbb{E}[\widehat{g}^2] - (\mathbb{E}[\widehat{g}])^2$ the unbiasedness proof pins in place).

Because $b(s)$ can be chosen separately for each state, you may minimize each state's contribution on its own, which is the same parabola-in-b minimization as in class, now performed inside the conditional expectation. Derive

$$b^\star(s) = \frac{\mathbb{E}\left[G\,u^2 \mid s\right]}{\mathbb{E}\left[u^2 \mid s\right]} = \frac{\sum_a \pi_\theta(a|s)\,u(s,a)^2\,\mathbb{E}[G \mid s,a]}{\sum_a \pi_\theta(a|s)\,u(s,a)^2},$$

a *score-weighted average of the returns* that follow each action: action a's average return, weighted by $\pi_\theta(a|s)\,u(s,a)^2$.

Sanity-check the formula on the bandit at *general* θ (its one state, so drop s): show the weights normalize to $(1-p,\ p)$, flipped, so that each arm's reward is weighted by the *other* arm's probability, giving

$$b^\star = (1-p)\,r_1 + p\,r_2,$$

and then verify by enumeration that both possible values of $\widehat{g}_{b^\star}$ equal the exact gradient $p(1-p)(r_1 - r_2)$. Zero variance at every θ, not just at $\theta = 0$: for this bandit, one pull of one arm now computes the exact gradient.

Part C: The Value Function Is (Almost) the Perfect Control Variate

The formula $b^\star(s)$ is optimal but awkward: the weights $\pi_\theta(a|s)\,u(s,a)^2$ change with every coordinate of $\boldsymbol{\theta}$ (our CartPole network would need 226 different baselines per state), and computing them requires the very expectations we are sampling to estimate. This part identifies the approximation everyone actually uses.

First show that whenever the weights do not depend on a (as happened at $\theta = 0$ in Part A, where $u^2 = \frac{1}{4}$ for both arms), the weighted average collapses to the plain conditional average

$$b^\star(s) = \mathbb{E}[G \mid s] = V^\pi(s),$$

the value function you built the Bellman equation for on Problem 23. Explain in a sentence or two why $V^\pi(s)$ is the practical choice even when the weights do vary: one function serves every coordinate of $\boldsymbol{\theta}$, and it is a quantity an agent can estimate by averaging its own returns from s.

Finally, look at what multiplies the score once this baseline is subtracted. With $Q^\pi(s,a) = \mathbb{E}[G \mid s,a]$ from lecture, show

$$\mathbb{E}\left[G - V^\pi(s) \mid s,a\right] = Q^\pi(s,a) - V^\pi(s) = A^\pi(s,a),$$

the **advantage**, and show it recenters exactly: $\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[A^\pi(s,a)\right] = 0$ at every state s (use $V^\pi(s) = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}[Q^\pi(s,a)]$). Conclude in one sentence: the value function is (nearly) the variance-optimal control variate, state by state, which is Problem 2's coefficient upgraded from a constant to a function, thirteen weeks later.