CSCI 145 Problem 25: When Is the Value Baseline Variance-Optimal?

October 27, 2026

Submission Instructions

Please upload your work to Gradescope by **11:59pm on Wednesday, December 16**.

- Turn in a single PDF of *handwritten* solutions. Typed (including LaTeX) solutions will not be accepted.
- A complete, good-faith attempt earns submission credit. The grader may select any recent assigned problem for a brief conversation, so keep your work and be prepared to explain each solution and answer a related question. Scheduling details will be announced in class.

Problem 25: When Is the Value Baseline Variance-Optimal?

At one policy decision, let S be the observed state, G the subsequent return, and $\mathbf{u} = \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A \mid S)$ the policy score. Assume

$$\mathbb{E}[\mathbf{u} \mid S] = \mathbf{0}, \qquad 0 < \mathbb{E}[\|\mathbf{u}\|_2^2 \mid S] < \infty \quad \text{almost surely},$$

and assume $\mathbb{E}[G^2(1 + \|\mathbf{u}\|_2^2)] < \infty$. Every deterministic state-only baseline $b(S)$ preserves the mean update, but an actor–critic implementation typically uses the value baseline $V^\pi(S) = \mathbb{E}[G \mid S]$. A colleague claims that this conditional mean must also be the variance-minimizing state-only baseline.

For any measurable baseline $b(S)$ satisfying $\mathbb{E}[b(S)^2 \|\mathbf{u}\|_2^2] < \infty$, define the total coordinate variance

$$\mathcal{V}(b) = \sum_j \mathrm{Var}(((G - b(S))\mathbf{u})_j)\,.$$

1. Which state-only baseline minimizes $\mathcal{V}(b)$, and what exact variance penalty does any other baseline incur?
2. What if-and-only-if condition involving $\mathrm{Cov}(G, \|\mathbf{u}\|_2^2 \mid S)$ characterizes when $b^*(S) = V^\pi(S)$ almost surely?

Hint: You may use the tower property of conditional expectation and

$$\sum_j \mathrm{Var}(H_j) = \mathbb{E}[\|\mathbf{H}\|_2^2] - \|\mathbb{E}[\mathbf{H}]\|_2^2$$

for a random vector $\mathbf{H}$.

Takeaway: *A baseline can predict the average return correctly and still leave unnecessary noise in the gradient. The best baseline averages returns with weights given by the squared lengths of the policy scores. It accounts for how much each return can affect the update.*