CSCI 145 Problem 24: Same Objective, Different Noise

November 4, 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 24: Same Objective, Different Noise

A policy with parameters $\boldsymbol{\theta}$ generates a trajectory τ through an environment whose dynamics need not be differentiable. Write

$$\mathbf{u}(\tau) = \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tau), \qquad \mathbf{g} = \mathbb{E}[G\mathbf{u}]$$

for its trajectory score and exact policy gradient, where G is the scalar return. Two evaluation systems encode the same outcome with different reward units and origins, replacing G by

$$\widetilde{G} = aG + c$$

for constants a and c. A teammate claims that this recoding is irrelevant to policy-gradient training. Choosing a variance-reducing reference is reserved for the next problem, so c is not to be optimized here.

1. For each sign of a, how do the recoded objective and its exact gradient change the policy optimization problem?
2. Assuming $\mathbb{E}[\|\mathbf{u}\|_2^2] > 0$ and the needed second moments exist, how do the mean and total coordinate variance of the one-trajectory REINFORCE estimator depend on c, and can changing only the reward origin make it arbitrarily noisy for fixed $a \neq 0$?

Hint: You may use $\mathbb{E}[\mathbf{u}] = \mathbf{0}$ 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: *Adding a constant to every return leaves the exact policy gradient unchanged. It can still make a sampled gradient arbitrarily noisy. How we encode rewards matters when training uses sampled gradients.*