

CSCI 145 Problem 24: Same Objective, Different Noise

August 27, 2026

Submission Instructions

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

- Turn in a single PDF of *handwritten* solutions. Typed (including L^AT_EX) 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; bring your work and be ready to explain any of them.

Problem 24: Same Objective, Different Noise

2 points

A two-armed bandit chooses arm 1 with probability $p = \sigma(\theta)$ and arm 0 otherwise. The arms initially pay deterministic rewards r_1 and r_0 . A teammate says that adding the same constant c to both rewards is irrelevant because it does not change which policy is optimal.

Decide whether “irrelevant” is correct for one-sample REINFORCE training. Derive the exact objective gradient and the complete distribution of the one-sample estimator after the reward shift. Find the value of c that minimizes its variance for a general $p \in (0, 1)$, and determine whether zero variance is possible without making the true gradient zero.

Then replace each arm’s deterministic reward by an independent draw with the same mean as before and conditional variance s_a^2 on arm a . Determine the best constant shift again and identify the part of the estimator variance that no constant can remove. Use the result to state precisely which aspects of the learning problem are unchanged by a reward shift and which are not.

Results you may use. For this Bernoulli policy, $\partial_\theta \log \pi_\theta(a) = a - p$. You may use the law of total variance.