CSCI 145 Problem Set 10

November 6, 2025

Submission Instructions

Please upload *your* work by **11:59pm Monday November 10, 2025.**

- You are encouraged to discuss ideas and work with your classmates. However, you **must acknowledge** your collaborators at the top of each solution on which you collaborated with others and you **must write** your solutions independently.
- Your solutions to theory questions must be written legibly, or typeset in LaTeX or markdown. If you would like to use LaTeX, you can import the source of this document (available from the course webpage) to Overleaf.
- I recommend that you write your solutions to coding questions in a Jupyter notebook using Google Colab.
- You should submit your solutions as a **single PDF** via the assignment on Gradescope.

Grading: The point of the problem set is for *you* to learn. To this end, I hope to disincentivize the use of LLMs by **not** grading your work for correctness. Instead, you will grade your own work by comparing it to my solutions. This self-grade is due the Friday *after* the problem set is due, also on Gradescope.

Problem 1: Policy Gradient Reward Shift

Using the policy gradient strategy we described in class, our strategy is to run gradient descent using the gradient:

$$\mathbb{E}_{\tau\sim\pi}[R(\tau)\nabla_\theta \log \pi(\tau)]. \tag{1}$$

Computing the expectation exactly is infeasible because we can't (in general) enumerate every possible trajectory, so we *estimate* the expectation by sampling trajectories τ from the policy π. We'd like our estimator to be right in expectation, and to have low variance. In order to reduce the variance, we often subtract a constant baseline b from the reward $R(\tau)$. In this problem, we will investigate how to choose b.

Part A: Correct in Expectation

Consider a *discrete* distribution, i.e., τ is one of a finite set of values. With this assumption, we can write the expectation as a summation e.g.,

$$\mathbb{E}[R(\tau)] = \sum_\tau \pi(\tau)R(\tau) \tag{2}$$

where $\pi(\tau)$ is the probability that the policy produces the trajectory τ. Show that

$$\mathbb{E}[(R(\tau) - b)\nabla_\theta \log \pi(\tau)] = \mathbb{E}[R(\tau)\nabla_\theta \log \pi(\tau)]. \tag{3}$$

Hint: Use the log-derivative trick from class in reverse.

Part B: Variance Reduction

For mathematical simplicity, define $\mathbf{g} = \nabla_\theta \log \pi(\tau)$, and $R = R(\tau)$. Consider the random variable

$$\mathbf{X} = (R - b)\mathbf{g} \tag{4}$$

Derive the optimal b^* to minimize the variance of X.

Hint: Recall that we can write the variance of a random variable X as $\mathrm{Var}(\mathbf{X}) = \mathbb{E}[\|\mathbf{X}\|_2^2] - \|\mathbb{E}[\mathbf{X}]\|^2$.

Suppose that $\mathbf{g}$ is independent of R, what is the optimal b^*?

Problem 2: Reinforcement Learning in Action

Part A: A New Environment

Adapt the code from class to a different Gymnasium environment.

Part B: Baseline Shift Strategies

Define the per-step *cost-to-go*

$$R_\ell(\tau) \;=\; \sum_{t=\ell}^{L-1} (-r_t)\,\gamma^{\,t-\ell},$$

and let

$$g_\ell \;=\; \nabla_\Theta \log \pi_\Theta(a_\ell).$$

Using a constant (per-episode) baseline b, the REINFORCE gradient is

$$\hat{\mathbf{g}} \;=\; \sum_{\ell=0}^{L-1} \bigl(R_\ell - b\bigr)\, g_\ell.$$

Compare the following choices of b:

(a) No baseline

$$b \;=\; 0, \qquad \hat{\mathbf{g}} \;=\; \sum_{\ell=0}^{L-1} R_\ell\, g_\ell.$$

(b) Mean (episode-average) baseline

$$\hat{b} \;=\; \frac{1}{L}\sum_{\ell=0}^{L-1} R_\ell, \qquad \hat{\mathbf{g}} \;=\; \sum_{\ell=0}^{L-1} \bigl(R_\ell - \hat{b}\bigr)\, g_\ell.$$

(c) Grad-norm (variance-minimizing constant) baseline

$$\hat{b} \;=\; \frac{\sum_{\ell=0}^{L-1} R_\ell \, \|g_\ell\|_2^2}{\sum_{\ell=0}^{L-1} \|g_\ell\|_2^2}, \qquad \hat{\mathbf{g}} \;=\; \sum_{\ell=0}^{L-1} \bigl(R_\ell - \hat{b}\bigr)\, g_\ell.$$

Plot the running average (window size 10) of *episode reward.*