Can Shifting Returns Reduce Policy Gradient Noise?

Due: Wednesday, December 16 at 11:59pm

Instructions: Please submit your *handwritten* solutions as a single PDF on Gradescope.

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.

1. A policy's expected return after recoding is $\mathbb{E}[aG + c]$. How does the sign of a affect which policies maximize it?
2. With one sampled trajectory, the recoded gradient estimate is $(aG + c)\mathbf{u}$. What is the mean of this estimate?
3. Suppose $a \neq 0$ is fixed, $\mathbb{E}[\|\mathbf{u}\|_2^2] > 0$, and the needed second moments exist. Can changing only c make the estimate's total coordinate variance arbitrarily large?

Hint: You may use $\mathbb{E}[\mathbf{u}] = \mathbf{0}$ and

$$\sum_{j=1}^{d} \mathrm{Var}(H_j) = \mathbb{E}[\|\mathbf{H}\|_2^2] - \|\mathbb{E}[\mathbf{H}]\|_2^2$$

for a random vector $\mathbf{H} \in \mathbb{R}^d$.

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.*