Robot Learning	**Assignment 3**
CS 4756 Spring 2025	**Due 11:59pm Friday, March 21**

Bounding Error in Approximate Policy Iteration (10 points)

In this problem, we look at how errors in approximating the value function affect the performance of a policy during policy iteration. We aim to understand how the error ϵ in the value function propagates and impacts the overall performance of the policy.

Let's assume we are in an infinite horizon MDP with discount factor γ. We have a reference policy π whose true value function is $V^\pi(s)$. We collect rollouts with π, and fit a neural network to approximate this value function, where $\hat{V}(s) \approx V^\pi(s)$, given that the value function approximation error is bounded by ϵ.

Let's assume we did a really good job and can guarantee that the error from the fit is at most ϵ. More formally, let $||V^\pi - \hat{V}||_\infty \leq \epsilon$.[1]

We now choose a greedy policy to improve upon policy $\hat{\pi}$:

$$\hat{\pi}(s) = \underset{a}{\text{argmax}} \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a)\hat{V}(s') \right]$$

Note that this is exactly the policy improvement step, except the value function is substituted with our approximate value function. We want to know how the greedy policy $\hat{\pi}(s)$ performs with respect to $\pi(s)$.

In other words, $\hat{\pi}$ can end up doing much worse than π. This shows that even though the error in approximating the value function is at most ϵ, the performance error scales up by a factor of $\frac{1}{1-\gamma}$.

(a) Let $V^{\hat{\pi}}(s)$ be the value of the greedy policy $\hat{\pi}(s)$. Prove the following:

$$V^\pi(s) - V^{\hat{\pi}}(s) \leq \frac{2\gamma\epsilon}{1-\gamma} \text{ , for all } s$$

In other words, $\hat{\pi}$ can end up doing much worse than π. Additionally, even though the error from fitting the value was ϵ, the performance error scales up by a factor of $\frac{1}{1-\gamma}$.
Hint: One way to approach the question would be:

1. Start by using the Bellman equation for any policy:

   $$V^\pi(s) = R(s,\pi(s)) + \gamma \sum_{s'} P(s'|s,\pi(s))V^\pi(s')$$

   Use this substitution to expand $V^\pi(s) - V^{\hat{\pi}}(s)$.

2. Next, note that you need to establish a relationship between $\pi(s)$ and $\hat{\pi}(s)$. Exploit the following observation: $\hat{\pi}(s) = \text{argmax}_a f(s,a)$ must imply $f(s,\hat{\pi}(s)) \geq f(s,\pi(s))$ for any policy π.

3. Use these facts to obtain the following intermediate result. For any s,

   $$V^\pi(s) - V^{\hat{\pi}}(s) \leq 2\gamma\epsilon + \gamma \sum_{s' \in \mathcal{S}} \Pr[s'|s,\hat{\pi}(s)](V^\pi(s') - V^{\hat{\pi}}(s'))$$

   From the above, you should be able to prove the final result for all s.

[1] Note: $||x||_\infty$ is the L-infinity norm

(b) Explain why the performance error bound between $\pi(s)$ and $\hat{\pi}(s)$ is not simply ϵ. Does the error remain constant, scale with time, or compound over time? Why is the scaling factor $\frac{1}{1-\gamma}$ significant in this bound?

Hint: Think about how the value function accumulates rewards over an infinite horizon and how errors in each step affect the future value estimates.

(c) What would happen to the performance bound between if the time horizon T were finite instead of infinite? How would the bound change, and what implications does this have for practical applications of approximate policy iteration?