CMPUT 365: Introduction to Reinforcement Learning, Winter 2023
Worksheet #9: On-policy Prediction with Approximation

Manuscript version: *#38558e* - 2023-04-08 23:44:39-06:00

Question 1. Let $f(x, y) = (x + y)^2 + e^{xy}$. Recall that the gradient, assuming it exists, is composed of the partial derivatives for each variable,

$$\nabla f(x, y) = \begin{bmatrix} \frac{\partial f(x,y)}{\partial x} \\ \frac{\partial f(x,y)}{\partial y} \end{bmatrix}.$$

1. What is $\nabla f(x, y)$ for the f defined above?
2. What is $\nabla f(0, 1)$, the value that $\nabla f : (x, y) \mapsto \nabla f(x, y)$ takes at $(x, y) = (0, 1)$?

Question 2. Find the gradient of f for the following functions:

1. $f(x, y, z) = \frac{y^z}{x}$, where $x, y, z \in \mathbb{R}$, $x \neq 0$, $y > 0$;
2. $f(x) = e^{x^2+5}$, where $x \in \mathbb{R}$;
3. $f(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}$, where $\mathbf{x}, \boldsymbol{\beta} \in \mathbb{R}^N$;
4. $f(\mathbf{x}) = (\mathbf{x}^T\boldsymbol{\beta} - y)^2$, where $\mathbf{x}$, $\boldsymbol{\beta} \in \mathbb{R}^N$, $y \in \mathbb{R}$.

Question 3. (*Exercise 9.1 S&B*, slightly modified) Show that tabular methods (such as those presented in Course 2 of the MOOC and Part I of the RL book) are a special case of linear function approximation. In this case, what would the feature vectors be? Consider for example tabular value function prediction.

Question 4. Consider an episodic MDP with the state set $\mathcal{S} = \{1, 2, 3\}$ and a discount factor of $\gamma = 0.9$. Suppose $\hat{v}(s; w) = ws$, where $w \in \mathbb{R}$ and $s \in \mathcal{S}$. The agent observes the following episodic data before reaching the terminal state at time $t = 3$, while following some fixed policy:

$$S_0 = 1, R_1 = -7, S_1 = 3, R_2 = 5, S_2 = 1, R_3 = 10, S_3 = \texttt{<terminal>}.$$

Assume that w is initialized to 0 at the beginning of learning and the stepsize is fixed to $\alpha = 0.1$. Then

1. What is $\hat{v}(1; w)$ when w is updated using gradient Monte-Carlo on this data?
2. What is $\hat{v}(1; w)$ when w is updated using semi-gradient linear TD on this data?
3. Again compute the estimate $\hat{v}(1; w)$ for both gradient Monte-Carlo and semi-gradient Linear TD, when w is updated after seeing the same episode using the respective algorithms.

(Write down the numeric result to two decimal places.)

Question 5. Consider a discounted MDP $\mathcal{M}$ with state space $\mathcal{S}$. Let $\hat{v}(s;\mathbf{w}) = \mathbf{w}^\top \mathbf{x}(s)$, where $\mathbf{w} \in \mathbb{R}^d$ and $s \in \mathcal{S}$ and $\mathbf{x} : \mathcal{S} \to \mathbb{R}^d$ is some feature map. Fix a memoryless policy π of $\mathcal{M}$. For on-policy learning using TD while following π in $\mathcal{M}$, at the TD fixed point $\mathbf{w}_{\mathrm{TD}}$, as the book mentions (cf. Equation (9.4)), the mean-squared value error, $\overline{\mathrm{VE}}$, is within a constant factor of the lowest possible error:

$$\overline{\mathrm{VE}}(\mathbf{w}_{\mathrm{TD}}) \leq \frac{1}{1-\gamma} \min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w}) \,.$$

1. If $\gamma = 0.9$ and $\min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w}) = 1$, what is the minimum and maximum value of $\overline{\mathrm{VE}}(\mathbf{w}_{\mathrm{TD}})$? What is the minimum and maximum value of $\overline{\mathrm{VE}}(\mathbf{w}_{\mathrm{TD}})$ if $\min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w}) = 0$?

   (Hint: $\overline{\mathrm{VE}}$ is defined in Equation (9.1) in the book.)

2. We have seen that if $\hat{v}$ can perfectly represent the value function (in the sense that for some $\mathbf{w}_0 \in \mathbb{R}^d$, $\hat{v}(s;\mathbf{w}) = v_\pi(s)$ for all $s \in \mathcal{S}$) then $\min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w}) = 0$. How about the other direction: if $\min_{\mathbf{w}} \overline{\mathrm{VE}}(\mathbf{w}) = 0$, then does that mean $\hat{v}$ can perfectly represent v_π? Justify your answer.

Question 6. Assume that $\nabla f(\mathbf{x})$ exists. Then answer the following using true or false?

1. The gradient of a function is a function.
2. The gradient of a function at a point will point in the direction of steepest descent, i.e. following the direction of the gradient will locally decrease the function value.
3. For a function f, the gradient $\nabla f(\mathbf{x})$ at point $\mathbf{x}$ will be the zero vector if $\mathbf{x}$ globally *minimizes* f.
4. For a function f, the gradient $\nabla f(\mathbf{x})$ at point $\mathbf{x}$ will be the zero vector if $\mathbf{x}$ globally *maximizes* f.
5. For a function f, if the gradient $\nabla f(\mathbf{x}) = 0$ then $\mathbf{x}$ locally minimizes the function f.
