Course 3, Module 1
On-policy Prediction with Approximation

CMPUT 397
Fall 2019

Worksheet Question

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

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

where $\frac{\partial f(x,y)}{\partial x}$ is the derivative of $f(x, y)$ w.r.t. x assuming that y is fixed.

(a) What is $\nabla f(x, y)$ for the f defined above? Hint: Recall that the derivative of e^z is e^z.

(b) What is $\nabla f(0, 1)$?

Pop Quiz!

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} p(x)x$$

- Imagine you want to estimate the expected value E[X]
  - Example: X = height in cms for a person, E[X] = average height in population
- How would you estimate E[X], if you do not have p?
  - But, you can sample from p

Estimating the Gradient

- Whiteboard to discuss how we estimate the gradient

Mean Squared Value Error

$$\sum_{s} \mu(s)[v_\pi(s) - \hat{v}(s, \mathbf{w})]^2$$

The fraction of time we spend in S when following policy π

Gradient of Value Error, with Linear Fcn Approx

$$\nabla \overline{VE}(\mathbf{w}) = \nabla \sum_{s \in \mathcal{S}} \mu(s)[v_\pi(s) - \mathbf{w}^T \mathbf{x}(s)]^2$$

$$= \sum_{s \in \mathcal{S}} \mu(s) \nabla [v_\pi(s) - \mathbf{w}^T \mathbf{x}(s)]^2$$

$$= -\sum_{s \in \mathcal{S}} \mu(s) 2[v_\pi(s) - \mathbf{w}^T \mathbf{x}(s)] \nabla \mathbf{w}^T \mathbf{x}(s)$$

$$= -\sum_{s \in \mathcal{S}} \mu(s) 2[v_\pi(s) - \mathbf{w}^T \mathbf{x}(s)] \mathbf{x}(s)$$

Sample of Gradient

$$\nabla \overline{VE}(\mathbf{w}) = -\sum_{s \in \mathcal{S}} \mu(s) 2[v_\pi(s) - \mathbf{w}^T \mathbf{x}(s)]\mathbf{x}(s)$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \alpha 2[v_\pi(s) - \mathbf{w}^T \mathbf{x}(s)]\mathbf{x}(s)$$

Exercise Questions

$$\min_{\mathbf{w}\in\mathbb{R}^d} \sum_s \mu(s)[v_\pi(s) - \hat{v}(s, \mathbf{w})]^2$$

- Why can't we directly optimize the MSVE? We know the stochastic gradient descent update would be the following

$$\mathbf{w}_t + \alpha[v_\pi(S_t) - \hat{v}(S_t, \mathbf{w})]\,\nabla\,\hat{v}(S_t, \mathbf{w})$$

- Further, why doesn't the TD fixed point minimize the MSVE?
    - Let's do this on the whiteboard