On-policy Prediction with Approximation
Graded Quiz • 30 min

Due Nov 1, 11:59 PM MST

On-policy Prediction with Approximation

TOTAL POINTS 12

1. **Which of the following statements is true about function approximation in reinforcement learning? (Select all that apply)** 1 point

   □ **We only use function approximation because we have to for large or continuous state spaces. We would use tabular methods if we could, and learn an independent value per state.**

   □ **It allows faster training by generalizing between states.**

   □ **It can be more memory efficient.**

   □ **It can help the agent achieve good generalization with good discrimination, so that it learns faster and represent the values quite accurately.**

2. **We learned how value function estimation can be framed as supervised learning. But not all supervised learning methods are suitable. What are some key differences in reinforcement learning that can make it hard to apply standard supervised learning methods?** 1 point

   □ **When using bootstrapping methods like TD, the target labels change.**

   □ **Data is available as a fixed batch.**

   □ **Data is temporally correlated in reinforcement learning.**

3. **Value Prediction (or Policy Evaluation) with Function Approximation can be viewed as supervised learning mainly because ________. [choose the most appropriate completion of the proceeding statement]** 1 point

   ○ **We can learn the value function by training with batches of data obtained from the agent's interaction with the world.**

   ○ **Each state and its target estimate (used in the Monte Carlo update, TD(0) update, and DP update) can be seen as input-output training examples to estimate a continuous function.**

   ○ **We use stochastic gradient descent to learn the value function.**

4. **Which of the following is true about using Mean Squared Value Error ($\bar{V}E = \sum \mu(s)[v_\pi(s) - \hat{v}(s,w)]^2$) as the prediction objective?** 1 point

   $\mu(s)$ **represents the weighted distribution of visited states**

   (Select all that apply)

   □ **Gradient Monte Carlo with linear function approximation converges to the global optimum of this objective, if the step size is reduced over time.**

□ The agent can get zero MSVE when using a linear representation that cannot represent the true values of states visited ($\mu(s) \neq 0$)

□ The agent can get zero MSVE when using a tabular representation that can represent the true values.

□ This objective makes it explicit how we should trade-off accuracy of the value estimates across states, using the weighting μ.

5. Which of the following is true about $\mu(S)$ in Mean Squared Value Error? (Select all that apply)

1 point

□ If the policy is uniformly random, $\mu(S)$ would have the same value for all states.

□ It is a probability distribution.

□ It serves as a weighting to minimize the error more in states that we care about.

□ It has higher values for states that are visited more often.

6. The stochastic gradient descent update for the MSVE would be as follows.

1 point

Fill in the blanks (A), (B), (C) and (D) with correct terms. (Select all correct answers)

$$\mathbf{w_{t+1}} \doteq \mathbf{w_t}\ (A)\ \tfrac{1}{2}\alpha\nabla[\ (C)\ -\ (D)\ ]^2$$

$$= \mathbf{w_t}\ (B)\ \alpha[\ (C)\ -\ (D)\ ]\nabla\hat{v}(S_t, \mathbf{w_t})$$

$(\alpha > 0)$

□ $-, -, \hat{v}(S_t, \mathbf{w_t}), v_\pi(S_t)$

□ $+, -, v_\pi(S_t), \hat{v}(S_t, \mathbf{w_t})$

□ $+, +, \hat{v}(S_t, \mathbf{w_t}), v_\pi(S_t)$

□ $-, +, v_\pi(S_t), \hat{v}(S_t, \mathbf{w_t})$

7. In a Monte Carlo Update with function approximation, we do stochastic gradient descent using the following gradient:

1 point

$$\nabla[G_t - \hat{v}(s, \mathbf{w})]^2 = 2[G_t - \hat{v}(s, \mathbf{w})]\nabla(-\hat{v}(S_t, \mathbf{w}_t))$$

$$= (-1) * 2[G_t - \hat{v}(s, \mathbf{w})]\nabla\hat{v}(S_t, \mathbf{w}_t)$$

But the actual Monte Carlo Update rule is the following:

$$\mathbf{w_{t+1}} = \mathbf{w_t} +\ \alpha[G_t - \hat{v}(S_t, \mathbf{w}_t)]\nabla\hat{v}(S_t, \mathbf{w}_t), \qquad (\alpha > 0)$$

Where did the constant -1 and 2 go when α is positive? (Choose all that apply)

□ We are performing gradient ascent, so we subtract the gradient from the weights, negating -1.

□ We are performing gradient descent, so we subtract the gradient from the weights, negating -1.

□ We assume that the 2 is included in the step-size.

□ We assume that the 2 is included in $\nabla\hat{v}(S_t, \mathbf{w}_t)$.

8. When using stochastic gradient descent for learning the value function, why do we only make a small update towards minimizing the error instead of fully minimizing the error at each encountered state? 1 point

○ Because we want to minimize approximation error for all states, proportionally to μ.

○ Because the target value may not be accurate initially for both TD(0) and Monte Carlo method.

○ Because small updates guarantee we can slowly reduce approximation error to zero for all states.

9. The general stochastic gradient descent update rule for state-value prediction is as follows: 1 point

$$\mathbf{w_{t+1}} \doteq \mathbf{w_t} + \alpha[U_t - \hat{v}(S_t, \mathbf{w_t})]\nabla\hat{v}(S_t, \mathbf{w_t})$$

For what values of U_t would this be a semi-gradient method?

○ $R_{t+1} + \hat{v}(S_{t+1}, w_t)$

○ G_t

○ $R_{t+1} + R_{t+2} + \ldots + R_T$

○ $v_\pi(S_t)$

10. Which of the following statements is true about state-value prediction using stochastic gradient descent? 1 point

$$\mathbf{w_{t+1}} \doteq \mathbf{w_t} + \alpha[U_t - \hat{v}(S_t, \mathbf{w_t})]\nabla\hat{v}(S_t, \mathbf{w_t})$$

(Select all that apply)

□ Using the Monte Carlo return or true value function as target results in an unbiased update.

□ Stochastic gradient descent updates with Monte Carlo targets always reduce the Mean Squared Value Error at each step.

□ Semi-gradient TD(0) methods typically learn faster than gradient Monte Carlo methods.

□ When using $U_t = R_{t+1} + \hat{v}(S_{t+1}, \mathbf{w_t})$, the weight update is not using the true gradient of the TD error.

□ Using the Monte Carlo return as target, and under appropriate stochastic approximation conditions, the value function will converge to a local optimum of the Mean Squared Value Error.

11. Which of the following is true about the TD fixed point? 1 point

(Select all correct answers)

□ At the TD fixed point, the mean squared value error is not larger than $\frac{1}{1-\gamma}$ times the minimal mean squared value error, assuming the same linear function approximation.

□ **Semi-gradient TD(0) with linear function approximation converges to the TD fixed point.**

□ **The weight vector corresponding to the TD fixed point is a local minimum of the Mean Squared Value Error.**

□ **The weight vector corresponding to the TD fixed point is the global minimum of the Mean Squared Value Error.**

12. **Which of the following is true about Linear Function Approximation, for estimating state-values? (Select all that apply)** 1 point

□ **The size of the feature vector is not necessarily equal to the size of the weight vector.**

□ **The gradient of the approximate value function $\hat{v}(s, \mathbf{w})$ with respect to $\mathbf{w}$ is just the feature vector.**

□ **Features are often called basis functions because every approximate value function we consider can be written as a linear combination of these features.**

□ **State aggregation is one way to generate features for linear function approximation.**

□ I, **Dhawal Gupta**, understand that submitting work that isn't my own may result in permanent failure of this course or deactivation of my Coursera account.

Learn more about Coursera's Honor Code

Save Submit