CS 747 (Autumn 2025) Week 10 Test (Batch 1)

5.35 p.m. – 6.00 p.m., October 21, 2025, LA 001

Name: _____ Roll number: _____

Note. There is one question in this test. You can use the space on both pages for your answer. Draw a line (either vertical or horizontal) and do all your rough work on one side of it.

Question 1. An MDP has 4 states, namely s_1, s_2, s_3, and s_4. The value function of a particular policy π is learned, using tile coding for function approximation. A *single* feature "ϕ" is employed. The feature value of each state s, as well as its value under π, is provided below.

s	$\phi(s)$	$V^\pi(s)$
s_1	0.8	5
s_2	1.1	3
s_3	2.1	2
s_4	2.6	7

The tile coding architecture uses two infinite tilings, both with a tile width of 1. The first tiling has tiles in the ranges $(1, 2]$, $(2, 3]$, $(3, 4]$, etc. The second tiling is offset by 0.5, and hence has tiles in the ranges $(0.5, 1.5]$, $(1.5, 2.5]$, $(2.5, 3.5]$, etc.

Suppose that the four states occur with equal probability in the long term: that is, the stationary probability of each state is $\frac{1}{4}$. What is the least possible value of mean-squared value error that this particular function approximation architecture can achieve? Provide a detailed justification, and suitably introduce/define any notation that you use. [3 marks]

Answer 1. Denote by w_α the weight corresponding to the tile that covers the interval $(\alpha, \alpha+1]$. The tile coding scheme implies the following approximation scheme.

$$\begin{aligned} V^\pi(s_1) &\approx w_0 + w_{0.5}, \\ V^\pi(s_2) &\approx w_1 + w_{0.5}, \\ V^\pi(s_3) &\approx w_2 + w_{1.5}, \\ V^\pi(s_4) &\approx w_2 + w_{2.5}. \end{aligned}$$

To what can we set these eight weights so that the apprixomation error is smallest? Upon visual inspection, it is apparent that we can achieve 0 error. Several possible configurations of the weight vector can do this, for example

$$\begin{aligned} w_0 &= V^\pi(s_1) = 5, \\ w_1 &= V^\pi(s_2) = 3, \\ w_2 &= V^\pi(s_3) = 2, \\ w_{0.5} &= 0, \\ w_{1.5} &= 0, \\ w_{2.5} &= V^\pi(s_4) - V^\pi(s_3) = 5. \end{aligned}$$

Consequently the mean-squared value error is

$$\frac{1}{4}\left((5 - w_0 - w_{0.5})^2 + (3 - w_1 - w_{0.5})^2 + (2 - w_2 - w_{1.5})^2 + (7 - w_2 - w_{2.5})^2\right) = 0.$$

CS 747 (Autumn 2025) Week 10 Test (Batch 2)

6.15 p.m. – 6.40 p.m., October 21, 2025, LA 001

Name: _____ Roll number: _____

Note. There is one question in this test. You can use the space on both pages for your answer. Draw a line (either vertical or horizontal) and do all your rough work on one side of it.

Question 1. An agent interacts with an MDP M in which every reward is 0. The agent follows a fixed policy π, whose value function it estimates using linear function approximation. In particular, the agent uses Linear TD(0) (which is the version with full bootstrapping) to make its learning updates.

For each state s, the value estimate is $\mathbf{w} \cdot \phi(s)$, where $\mathbf{w}$ is the weight vector and $\phi(s)$ is the feature vector, both of the same dimension. The dimension is smaller than the number of states, implying that there is generalisation across states. Linear TD(0) is initialised with a *non-zero* weight vector, and is run with an appropriate schedule for the learning rate.

Can the agent be guaranteed to converge to a weight vector that minimises the mean-squared value error? Answer yes or no, and provide sufficient justification for your answer. While working out your answer, consider the similarities and differences between this setup and that in Tsitsiklis and Van Roy's counterexample. [3 marks]

Answer 1. Yes.

In common with Tsitsiklis and Van Roy's counterexample, this setup involves two of the three elements of the deadly triad, namely bootstrapping and generalisation. However, by making updates along the trajectory that is encountered while following π, the updates are on-policy (not off-policy).

On the other hand, we notice that the value function of π is 0, and hence the weight vector $\mathbf{w}^\star = 0$ will achive 0 mean-squared value error. We know that that Linear TD(1) will therefore achieve 0 MSVE. In turn, this means that TD(0) will also achieve 0 MSVE, since the error of the latter is at most $\frac{1}{1-\gamma}$ of the former.

In summary, Linear TD(0) does achieve 0 mean-squared value error, which is clearly the least possible.