CS 747 (Spring 2025) Week 9 Test (Batch 1)

5.35 p.m. – 6.00 p.m., March 20, 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. Suppose the TD(0) algorithm is being run with learning rate $\alpha_t \in [0, 1]$ for time steps $t = 0, 1, 2, \ldots$. The algorithm is run on a continuing MDP $M = (S, A, T, R, \gamma)$, with notations as usual, in which discount factor $\gamma \in (0, 1)$. A policy $\pi : S \to A$ is being evaluated. At time step t (that is, after t updates have been made), let $V^t(s)$ denote the value function estimate for state $s \in S$; thus the initial values are $V^0 : S \to \mathbb{R}$. Learning rate α_t is used in the update to get V^{t+1} from V^t.

Recall that one of the conditions required on the learning rate sequence $(\alpha_t)_{t=0}^{\infty}$ for V^t to converge to V^π as $t \to \infty$ is that $\sum_{t=0}^{\infty} \alpha_t = \infty$. Call this the unbounded-sum condition. On the other hand, suppose that the learning rate sequence we are using is such that its sum is upper-bounded by a constant. In other words, our sequence satisfies

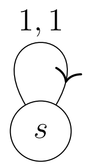
$$\sum_{t=0}^{\infty} \alpha_t < c$$

for some positive constant c.

Show that there exist M, π, and V^0 such that V^t *does not* converge to V^π as $t \to \infty$. Your demonstration will serve as a proof that the unbounded-sum condition is necessary in general for TD(0) to converge to the true value function. You are encouraged to think of a "simple" choice of MDP M for this proof; to focus on the relationship between V^{t+1} and V^t, which will depend on α_t; and to examine resulting constraints on the sequence of value estimates. [3 marks]

Answer 1.

Consider an MDP with a single state s, a single action a, and discount factor $\gamma \in (0,1)$ that we will specify later. The MDP is shown in the figure below—the sole transition has transition probability and reward both as 1.



Clearly the value of state s is $V(s) \stackrel{\text{def}}{=} \frac{1}{1-\gamma}$. Suppose our sequence of estimates of the value of s is $V^0, V^1, V^2, \ldots$. We set our initial estimate to be $V^0 = 0$. For $t = 0, 1, 2, \ldots$, the TD(0) update rule yields

$$V^{t+1} = V^t(1-\alpha_t) + \alpha_t(1 + \gamma V^t) = V^t + \alpha_t - V^t(1-\gamma)\alpha_t.$$

We shall prove by induction that for $t = 0, 1, 2, \ldots$, (1) $V^{t+1} \geq 0$ and (2) $V^{t+1} \leq V^t + \alpha_t$. The base case of $t = 0$ is easily verified, since $V^1 = \alpha_0$. If the hypothesis is true for t, it must follow for $t+1$ since (1) V^{t+1} is the sum of $V^t(1-(1-\gamma)\alpha_t)$ and α_t, which are both non-negative, and (2) the quantity $1-(1-\gamma)\alpha_t$ lies in $(0,1)$, and hence $V^{t+1} \leq V^t(1) + \alpha_t$.

It follows from the proof above that for $t = 0, 1, 2, \ldots$,

$$V^{t+1} \leq \sum_{i=0}^{t} \alpha_i \leq \sum_{i=0}^{\infty} \alpha_i < c.$$

On the other hand, if $\gamma > 1 - \frac{1}{c}$, then the true value

$$V(s) = \frac{1}{1-\gamma} > c.$$

We have shown that the sequence $V^0, V^1, V^2, \ldots$ cannot converge to $V(s)$.

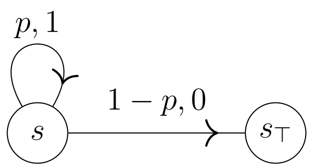
CS 747 (Spring 2025) Week 9 Test (Batch 2)

6.15 p.m. – 6.40 p.m., March 20, 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 a single non-terminal state s and a terminal state $s_\top$. Starting from s and following some fixed policy π, the probability of staying in s is p, while the probability of terminating is $1-p$, for some $p \in (0,1)$. The reward for transitioning from s to s is 1, while that for terminating is 0. Transitions in the MDP, under this fixed policy π, are shown below; arrows are annotated with "transition probability, reward". There is no discounting.



Suppose we use the TD(0) algorithm to estimate the value of s under π, starting with an initial estimate $V^0 = 0$. Also suppose that each learning update is performed with a constant learning rate $\alpha \in [0,1]$. A *single* episode is executed, starting from s and following π, until $s_\top$ is reached. A learning update using the TD(0) rule is performed after each transition. Let V denote the value estimate for s obtained at the end of the episode. Naturally V is a random variable, obtained after 1 or more time steps (the length of the episode). Calculate $\mathbb{E}[V]$—that is, the expectation of V. Comment on its dependence on α (if any). [3 marks]

Answer 1.

V is determined by the number of transitions in the episode, which is itself a random variable. Suppose an episode ends after m transitions from s to s, and then a final transition to $s_\top$, for some $m \geq 0$. The probability of such an episode is $p^m(1-p)$. On such an episode, we have

$$\begin{aligned}
V^0 &= 0; \\
V^1 &= V^0(1-\alpha) + \alpha(1 + V^0) = V^0 + \alpha = \alpha. \\
V^2 &= V^1(1-\alpha) + \alpha(1 + V^1) = V^1 + \alpha = 2\alpha. \\
&\vdots \\
V^m &= V^{m-1}(1-\alpha) + \alpha(1 + V^{m-1}) = V^{m-1} + \alpha = m\alpha. \\
V &= V^m(1-\alpha) + \alpha(0) = V^m(1-\alpha) = m\alpha(1-\alpha).
\end{aligned}$$

Having characterised the probability distribution over V, we obtain its expectation:

$$\begin{aligned}
\mathbb{E}[V] &= \sum_{m=0}^{\infty} \mathbb{P}\{\text{The episode has } m \text{ transitions from } s \text{ to } s\} \times m\alpha(1-\alpha) \\
&= \sum_{m=0}^{\infty} p^m(1-p)m\alpha(1-\alpha) \\
&= \frac{p\alpha(1-\alpha)}{1-p}.
\end{aligned}$$

We observe that $\mathbb{E}[V]$ does depend on α, and V is a *biased* estimator for all values of $\alpha \in [0, 1]$, since $V^\pi(s) = \frac{p}{1-p}$.