CS234 Problem Session Solutions

Week 2: Jan 20

1) [CA Session] To Converge Or Not To Converge

Recall the TD-learning update: $V(s_t) = V(s_t) + \alpha([r_t + \gamma \cdot V(s_{t+1})] - V(s_t))$ and the Incremental Monte-Carlo update: $V(s_t) = V(s_t) + \alpha(G_t - V(s_t))$.

The convergence of these value functions is highly dependent on the learning rate α. In fact, it may be convenient to vary the learning rate based on the timestep. In this problem, we will explore how varying the learning rate may affect convergence.

Let us define our timestep-varying learning rate as α_t. The above updates will converge with probability 1 if the following conditions are satisfied:

$$\sum_{i=1}^{\infty} \alpha_t = \infty \tag{1}$$

$$\sum_{i=1}^{\infty} \alpha_t^2 < \infty \tag{2}$$

Condition (1) is required to guarantee that the steps are large enough to eventually overcome any initial conditions or random fluctuations. Condition (2) guarantees that eventually the steps become small enough to assure convergence.

(a) Let $\alpha_t = \frac{1}{n}$. Does this α_t guarantee convergence?

Solution This is the Harmonic series. For condition (1), we see $\sum_{i=1}^{\infty} \frac{1}{n} = \infty$, as the Harmonic series diverges to infinity. For condition (2), we see $\sum_{i=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$, which is the solution to the Basel problem. Thus, both conditions are satisfied, so convergence is guaranteed.

(b) Let $\alpha_t = \frac{1}{2}$. Does this α_t guarantee convergence?

Solution For condition (1), we see $\sum_{i=1}^{\infty} \frac{1}{2} = \infty$. For condition (2), we see $\sum_{i=1}^{\infty} \frac{1}{4} = \infty$, so condition (2) is not satisfied. In general, any constant step size will not guarantee convergence. In nonstationary environments, it may be desirable to continue varying the value function in response to newly seen rewards.

(c) Let $\alpha_t = \frac{1}{n^3}$. Does this α_t guarantee convergence?

Solution For condition (1), we see $\sum_{i=1}^{\infty} \frac{1}{n^3} \approx 1.202$, known as Apery's constant. For condition (2), we know $\sum_{i=1}^{\infty} \frac{1}{n^6} < \infty$. Thus, convergence is not guaranteed as condition (1) is not satisfied, but condition (2) is satisfied.

2) [CA Session] Asynchronous Policy Evaluation

In class you have learned *synchronous* policy evaluation where the update operation (we omit the dependence on the action since the policy is fixed in policy evaluation) reads

$$\text{for } s = 1, 2, ..., S, \quad V^+(s) := r(s) + \gamma \sum_{s'} p(s' \mid s) V(s'). \tag{3}$$

where as usual V is the value function, r is the reward function and $0 < \gamma < 1$ the discount factor. We summarize the above operation using the Bellman operator B^π for policy π and write

$$V^+ = B^\pi V \tag{4}$$

This Bellman operator updates all states at once and writes the resulting value in V^+. This requires a temporary value vector V^+. After the update is complete in all states, V^+ is copied to V and the procedure is repeated until convergence.

Using the definition of B^π in equation 3 and 4, we have shown that the Bellman operator is a contraction in ∞-norm:

$$\|B^\pi V' - B^\pi V''\|_\infty \leq \gamma \|V' - V''\|_\infty. \tag{5}$$

Recall that the ∞-norm of a vector is the maximum component in absolute value, i.e., $\|V\|_\infty = \max_s |V(s)|$.

This is important, because it ensures the value function converges to the value function of policy π if the Bellman operator for policy π is repeatedly applied.

However, sometime it is preferable to avoid the construction of the temporary V^+ and directly overwrite the values in V; this procedure uses the *asynchronous* Bellman operator (we omit the policy dependence for brevity) B_s that updates the value function only at state s:

$$(B_s V)(i) = \begin{cases} r(s) + \gamma \sum_{s'} p(s' \mid s) V(s'), & \text{if } i = s \\ V(i), & \text{otherwise} \end{cases}$$

In other words, B_s only updates the state numbered s and leaves the remaining unaltered. This avoids creating the temporary V^+.

(a) B_s is no longer contractive

Show why equation 5 no longer holds in general (it should hold for any γ) when we use B_s (with a fixed s) instead of B^π.

Solution B_s only affects one state; so for example if $V' = [0, 0]$ and $V'' = [1, 1]$ then $\|V' - V''\|_\infty = 1$. However, since B_s only modifies one entry, we must have $\|B_s V' - B_s V''\|_\infty = 1$ due to the state B_s did not modify.

(b) Improvement in state s

Not all is lost, however. Show that B_s makes progress in state s, meaning that

$$|(B_s V')(s) - (B_s V'')(s)| \leq \gamma \|V' - V''\|_\infty.$$

Solution For a fixed state s:

$$|(B_s V')(s) - (B_s V'')(s)| = |r(s) + \gamma \sum_{s'} p(s' \mid s) V'(s') - r(s) - \gamma \sum_{s'} p(s' \mid s) V''(s')| \tag{6}$$

$$= |\gamma \sum_{s'} p(s' \mid s)[V'(s') - V''(s')]| \tag{7}$$

$$\leq \gamma \sum_{s'} p(s' \mid s) \max_{s''} |V'(s'') - V''(s'')| \tag{8}$$

$$= \gamma \|V' - V''\|_\infty. \tag{9}$$

(c) Order of Updates Consider applying B_1 first, then B_2, ... until B_S, i.e., consider doing one update to all states in sequence. One can show that (you don't need to show this):

$$\|B_S B_{S-1} \cdots B_2 B_1 V' - B_S B_{S-1} \cdots B_2 B_1 V''\|_\infty \leq \gamma \|V' - V''\|_\infty \tag{10}$$

Consider changing the order of the updates, i.e. you apply B_S first, and then B_{S-1}... until B_2, B_1. Can we still make the same claim, i.e., does

$$\|B_1 B_2 \cdots B_{S-1} B_S V' - B_1 B_2 \cdots B_{S-1} B_S V''\|_\infty \leq \gamma \|V' - V''\|_\infty \tag{11}$$

still hold? Justify your answer in 1-2 sentences (no need for a proof or counterexample).

Solution We can definitely make the same claim (although in general performance of the operators in different order gives different numerical results). In particular consider doing a permutation of the value function.

3) [Breakout Rooms] Certainty Equivalence Estimate

Consider the discounted ($\gamma < 1$), infinite-horizon MDP $M = (S, A, P, R, \gamma)$, where we do not know the true reward function $R(s, a) \in \mathbb{R}$ and state transition probabilities $P(s'|s, a)$. With a slight abuse of notation, we will also write $P(s, a) \in \mathbb{R}^{|S|}$ to denote the vector of transition probabilities of size $|S|$, whose values sum up to 1. For a given policy π, V_M^π denotes the value function of π in M.

Fortunately, we are given estimates of R and P, namely, $\widehat{R}$ and $\widehat{P}$, respectively, with the following properties:

$$\max_{s,a} |\widehat{R}(s, a) - R(s, a)| < \epsilon_R$$

$$\max_{s,a} \|\widehat{P}(s, a) - P(s, a)\|_1 < \epsilon_P$$

$$\text{where } \|\cdot\|_1 \text{ is the } L_1 \text{ norm for } \mathbf{x} \in \mathbb{R}^n\text{: } \|\mathbf{x}\| = \sum_{i=1}^{n} |x_i|$$

We can then define the *approximate MDP* $\widehat{M} = (S, A, \widehat{P}, \widehat{R}, \gamma)$ and let $V_{\widehat{M}}^\pi$ be the value function of policy π in $\widehat{M}$. For simplicity, assume that both reward functions are bounded within $[0, R_{\max}]$.

(a) Show that for all policies π and states s, the value function is bounded as follows:

$$0 \leq V^\pi(s) \leq \frac{R_{\max}}{1 - \gamma} \tag{12}$$

Solution All state values are at least 0 because reward is nonnegative. Also reward is at most $R_{\max}$, so state values are bounded by the one that has maximal ($R_{\max}$) reward at every time step. Thus for any policy π and state s, we have

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t \right] \leq \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R_{\max} \right] = R_{\max} \sum_{t=0}^{\infty} \gamma^t = \frac{R_{\max}}{1 - \gamma}$$

You are given that for any deterministic policy π, the error in state value is bounded as follows:

$$\|V_{\widehat{M}}^\pi - V_M^\pi\|_\infty \leq \frac{\epsilon_R}{1 - \gamma} + \gamma \epsilon_P \frac{R_{\max}}{2(1 - \gamma)^2} \tag{13}$$

where $\|\cdot\|_\infty$ is the L_∞ norm: $\|V_{\widehat{M}}^\pi - V_M^\pi\|_\infty = \max_s |V_{\widehat{M}}^\pi(s) - V_M^\pi(s)|$.

(b) Let π^* and $\widehat{\pi}^*$ be (deterministic) optimal policies in M and $\widehat{M}$, respectively. Using eq. (13), show that the following bound holds for all $s \in S$:

$$V_M^{\pi^*} - V_M^{\widehat{\pi}^*} \leq 2 \left(\frac{\epsilon_R}{1 - \gamma} + \gamma \epsilon_P \frac{R_{\max}}{2(1 - \gamma)^2} \right) \tag{14}$$

Note: What this means is that if we use a policy that is optimal in $\widehat{M}$, the amount we lose in value compared to the true optimal policy is bounded by the error in our approximation. In other words, the better the approximation, the closer we are to the true optimal policy!

Solution From the definition of value function, we get:

$$
\begin{aligned}
& V_M^{\pi^*}(s) - V_M^{\widehat{\pi}^*}(s) \\
=& V_M^{\pi^*}(s) - V_{\widehat{M}}^{\pi^*}(s) + V_{\widehat{M}}^{\pi^*}(s) - V_M^{\widehat{\pi}^*}(s) \\
\leq& \left\| V_M^{\pi^*} - V_{\widehat{M}}^{\pi^*} \right\|_\infty + V_{\widehat{M}}^{\pi^*}(s) - V_M^{\widehat{\pi}^*}(s) \\
\leq& \left\| V_M^{\pi^*} - V_{\widehat{M}}^{\pi^*} \right\|_\infty + V_{\widehat{M}}^{\widehat{\pi}^*}(s) - V_M^{\widehat{\pi}^*}(s) && \widehat{\pi}^* \text{ is optimal in } \widehat{M} \\
\leq& \left\| V_M^{\pi^*} - V_{\widehat{M}}^{\pi^*} \right\|_\infty + \left\| V_{\widehat{M}}^{\widehat{\pi}^*} - V_M^{\widehat{\pi}^*} \right\|_\infty \\
\leq& 2\left(\frac{\epsilon_R}{1-\gamma} + \gamma \epsilon_P \frac{R_{\max}}{2(1-\gamma)^2} \right) && \text{from eq. (13)}
\end{aligned}
$$

4) **[Breakout Rooms] Into the Unknown**

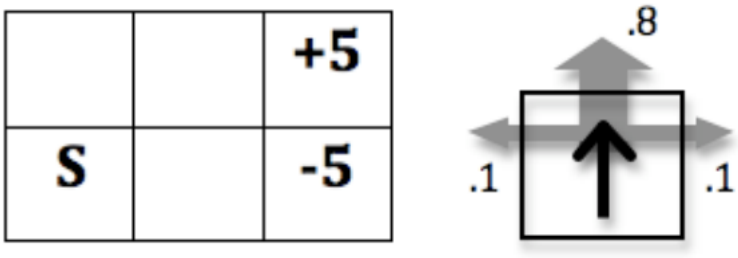

Figure 1: Left: Gridworld MDP, Right: Transition function

Let us define a gridworld MDP, depicted in Figure 1. The states are grid squares, identified by their row and column number (row first). The agent always starts in state (1,1), marked with the letter S. There are two terminal goal states, (2,3) with reward +5 and (1,3) with reward -5. Rewards are 0 in non-terminal states. (The reward for a state is received as the agent moves into the state.) The transition function is such that the intended agent movement (Up, Down, Left, or Right) happens with probability .8. With probability .1 each, the agent ends up in one of the states perpendicular to the intended direction. If a collision with a wall happens, the agent stays in the same state.

(a) Define the optimal policy for this gridworld MDP.

Solution

S=	(1, 1)	(1, 2)	(1, 3)	(2, 1)	(2, 2)	(2, 3)
$\pi^*(S)$	Up	Left	NA	Right	Right	NA

Since we know the transition function and the reward function, we can directly compute the optimal value function with value iteration. But, what if we **don't** know the transition and reward function?

(b) Suppose the agent does not know the transition probabilities. What does the agent need to be able do (or have available) in order to learn the optimal policy?

Solution The agent must be able to explore the world by taking actions and observing the effects.

(c) The agent starts with the policy that always chooses to go right, and executes the following three trajectories: **1)** (1,1)–(1,2)–(1,3), **2)** (1,1)–(1,2)–(2,2)–(2,3), and **3)** (1,1)–(2,1)–(2,2)–(2,3). What are the First-Visit Monte Carlo estimates for states (1,1) and (2,2), given these trajectories? Suppose $\gamma = 1$.

Solution To compute the estimates, average the rewards received in the trajectories that went through the indicates states.

$V((1,1)) = (-5+5+5)/3 = 5/3 = 1.666$

$V((2,2)) = (5+5)/2 = 5$

(d) Using a learning rate of $\alpha = 0.1$ and assuming initial values of 0, what updates does the TD-learning agent make after trials 1 and 2, above? For this part, suppose $\gamma = 0.9$.

Solution [1]

The general TD-learning update is: $V(s) = V(s) + \alpha(r + \gamma \cdot V(s') - V(s))$.

After trial 1, all of the updates will be zero, except for:

$V((1,2)) = 0 + .1(-5 + 0.9 \cdot 0 - 0) = -0.5$

After trial 2, the updates will be:

$V((1,1)) = 0 + .1(0 + 0.9 \cdot -0.5 - 0) = -0.045$

$V((1,2)) = -0.5 + .1(0 + 0.9 \cdot 0 + 0.5) = -0.45$

$V((2,2)) = 0 + .1(5 + 0.9 \cdot 0 - 0) = 0.5$

[1] This problem is borrowed from https://courses.cs.washington.edu/courses/cse573/10au/midterm1-solutions.pdf