CS7643: Deep Learning
Fall 2019
HW4 Solutions

James Hahn

November 7, 2019

1 Optimal Policy and Value Function

1. First, let's take the sum of discounted rewards as $\sum_{t=0}^{\infty} \gamma^t r_t(s_t, a_t)$

   We know we start at S_1, so $s_0 = S_1$, and the we always choose "stay", so $a_0 = a_i =$"stay". Since we always "stay" at the same state, s_i will always be S_1. As such, this summation becomes:

   $\sum_{t=0}^{\infty} \gamma^t r_t(s_t, a_t)$
   $= \sum_{t=0}^{\infty} \gamma^t r_t(S_0, \text{"stay"})$
   $= \sum_{t=0}^{\infty} \gamma^t (-1)$
   $= -\sum_{t=0}^{\infty} \gamma^t$

   So, if we assume this simulation runs for an infinite number of steps, then the sum of discounted rewards is the value $-\infty \cdot \gamma$, which becomes $-\infty$ since $\gamma > 0$.

2. The optimal policy, assuming we start at S_1 is $(a_1, a_2) = (\text{"go"}, \text{"go"})$. This is because all other rewards in the MDP are negative, except for the termination reward, which is a reward of $+3$. Assuming we're forced to take an action at each iteration of the simulation, the only way to achieve that positive reward is to first traverse to S_2 and then terminate the program by choosing the "go" action. This results in a sum of discounted rewards of $\sum_{t=0}^{1} \gamma^t r_t(s_t, a_t) = \gamma^0 r_0(S_1, \text{"go"}) + \gamma^1 r_1(S_2, \text{"go"}) = -2 + \gamma(3) = 3\gamma - 2$.

3. $V_0 = [0, 0]$

   $V_1 = [max(r(s_1, \text{"stay"}) + \gamma V_0(s_1), r(s_1, \text{"go"}) + \gamma V_0(s_2)), max(r(s_2, \text{"stay"}) + \gamma V_0(s_2), r(s_2, \text{"go"}))] = [max(-1, -2), max(-1, 3)] = [-1, 3]$

   $V_2 = [max(r(s_1, \text{"stay"}) + \gamma V_1(s_1), r(s_1, \text{"go"}) + \gamma V_1(s_2)), max(r(s_2, \text{"stay"}) + \gamma V_1(s_2), r(s_2, \text{"go"}))] = [max(-1 - \gamma, -2 + 3\gamma), max(-1 + 3\gamma, 3)] = [max(-2, 1), max(2, 3)] = [1, 3]$

   $V_3 = [max(r(s_1, \text{"stay"}) + \gamma V_2(s_1), r(s_1, \text{"go"}) + \gamma V_2(s_2)), max(r(s_2, \text{"stay"}) + \gamma V_2(s_2), r(s_2, \text{"go"}))] = [max(-1 + 1\gamma, -2 + 3\gamma), max(-1 + 3\gamma, 3)] = [max(0, 1), max(2, 3)] = [1, 3]$

   The optimal V is V_2 or V_3 because they both provide the highest value returns for each state across all iterations of V. With that being said, V_3 can generally be seen as better, since we show that the values have converged, whereas if we stopped at V_2, we don't have any idea if the values were already their optimal values or not.

2 Value Iteration Convergence

1. $||V^0 - V^*||_\infty = ||[-1,-3]||_\infty = max(|-1|,|-3|) = max(1,3) = 3$

   $||V^1 - V^*||_\infty = ||[-2,0]||_\infty = max(|-2|,|0|) = max(2,0) = 2$

   $||V^2 - V^*||_\infty = ||[0,0]||_\infty = max(|0|,|0|) = max(0,0) = 0$

   $||V^3 - V^*||_\infty = ||[0,0]||_\infty = max(|0|,|0|) = max(0,0) = 0$

   Clearly, the error decreases monotonically.

2. $||T(V) - T(V')||_\infty$
   $= ||max_a \sum_{s'} p(s'|s,a)[r(s,a) + \gamma V_i(s')] - max_a \sum_{s'} p(s'|s,a)[r(s,a) + \gamma V_i'(s')]||_\infty$
   $= ||p(s^*|s,a^*)[r(s,a^*) + \gamma V(s^*)] - p(s^*|s,a^*)[r(s,a^*) + \gamma V'(s*)]||_\infty$ (Let a^* and s^* represent the optimal action and state respectively)
   $= ||\gamma \cdot p(s^*|s,a^*)V(s^*) - \gamma \cdot p(s^*|s,a^*)V'(s*)||_\infty$
   $= ||\gamma \cdot p(s^*|s,a^*)(V - V')||_\infty$
   $= \gamma||p(s^*|s,a^*)(V - V')||_\infty$
   $\leq \gamma||p(s^*|s,a^*)||_\infty||V - V'||_\infty$ (by Cauchy-Schwarz Inequality)
   $= \gamma||V - V'||_\infty$ (we know $max_s \sum_{s'} p(s'|s,a) = 1$)

   $\therefore$ We have shown $||T(V) - T(V')||_\infty \leq \gamma||V - V'||_\infty$ $\square$

3.

3 Learning the Model

1.

4 Policy Gradients Variance Reduction

1. Let the approximation of the policy gradient be $\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^N R(\tau_i)\nabla_\theta log\, \pi_\theta(\tau_i)$.

   Now, let's show when $R(\tau) := R(\tau) - b$ does not change this estimate:

   $\frac{1}{N}\sum_{i=1}^N R(\tau_i)\nabla_\theta log\, \pi_\theta(\tau_i)$
   $\implies \frac{1}{N}\sum_{i=1}^N (R(\tau_i) - b)\nabla_\theta log\, \pi_\theta(\tau_i)$
   $= \frac{1}{N}\big[\sum_{i=1}^N \nabla_\theta log\, \pi_\theta(\tau_i)\big]\big[\sum_{i=1}^N R(\tau_i) - b\big]$
   $= \frac{1}{N}\big[\sum_{i=1}^N \nabla_\theta log\, \pi_\theta(\tau_i)R(\tau_i)\big] - \frac{1}{N}\big[\sum_{i=1}^N \nabla_\theta log\, \pi_\theta(\tau_i)b\big]$

   Now, in order to prove $\frac{1}{N}\sum_{i=1}^N R(\tau_i)\nabla_\theta log\, \pi_\theta(\tau_i) = \frac{1}{N}\big[\sum_{i=1}^N \nabla_\theta log\, \pi_\theta(\tau_i)R(\tau_i)\big] - \frac{1}{N}\big[\sum_{i=1}^N \nabla_\theta log\, \pi_\theta(\tau_i)b\big]$, we must show $\frac{1}{N}\sum_{i=1}^N \nabla_\theta log\, \pi_\theta(\tau_i)b = 0$. This can be seen below:

   $\frac{1}{N}\sum_{i=1}^N \nabla_\theta log\, \pi_\theta(\tau_i)b$
   $= \mathrm{E}_{\tau\sim\pi_\theta\theta}\big[\nabla_\theta log\, \pi_\theta(\tau_i)b\big]$
   $= \mathrm{E}_{s_{0:t},a_{0:(t-1)}}\big[\mathrm{E}_{s_{(t+1):T},a_{t:(T-1)}}[\nabla_\theta log\, \pi_\theta(\tau_i)b]\big]$

$= \mathrm{E}_{s_{0:t}, a_{0:(t-1)}} \big[b \cdot \mathrm{E}_{s_{(t+1):T}, a_{t:(T-1)}} [\nabla_\theta log\, \pi_\theta(\tau_i)] \big]$
$= \mathrm{E}_{s_{0:t}, a_{0:(t-1)}} \big[b \cdot \mathrm{E}_{a_t} [\nabla_\theta log\, \pi_\theta(\tau_i)] \big]$
$= \mathrm{E}_{s_{0:t}, a_{0:(t-1)}} \big[b \cdot \int \frac{\nabla_\theta \pi_\theta \theta(a_t|s_t)}{\pi_\theta \theta(a_t|s_t)} \pi_\theta(a_t|s_t) da_t \big]$
$= \mathrm{E}_{s_{0:t}, a_{0:(t-1)}} \big[b \cdot \nabla_\theta \int \pi_\theta(a_t|s_t) da_t \big]$
$= \mathrm{E}_{s_{0:t}, a_{0:(t-1)}} \big[b \cdot \nabla_\theta 1 \big]$
$= \mathrm{E}_{s_{0:t}, a_{0:(t-1)}} \big[b \cdot 0 \big]$
$= \mathrm{E}_{s_{0:t}, a_{0:(t-1)}} \big[0 \big]$
$= 0$

In the above mini-proof, we have shown for any t, the product of the gradient with b is 0.

As such, since the second term of $\frac{1}{N} \big[\sum_{i=1}^{N} \nabla_\theta log\, \pi_\theta(\tau_i) R(\tau_i) \big] - \frac{1}{N} \big[\sum_{i=1}^{N} \nabla_\theta log\, \pi_\theta(\tau_i) b \big]$ is 0, we can reduce it to $\frac{1}{N} \big[\sum_{i=1}^{N} \nabla_\theta log\, \pi_\theta(\tau_i) R(\tau_i) \big]$, and we observe that $\frac{1}{N} \sum_{i=1}^{N} (R(\tau_i) - b) \nabla_\theta log\, \pi_\theta(\tau_i) = \frac{1}{N} \sum_{i=1}^{N} R(\tau_i) \nabla_\theta log\, \pi_\theta(\tau_i)$. $\square$

2. $\mathrm{Var}(R(\tau_i) \nabla_\theta log\, \pi_\theta(\tau_i))$