CS7643: Deep Learning
Fall 2019
HW4 Solutions

James Hahn

November 12, 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:

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

   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. First, let's observe how the value of γ changes the policy values. If we say, $\gamma = 0$, then this doesn't give any incentive to future rewards. If $\gamma = 1$, then this heavily incentivizes future rewards (i.e. reaching the goal state). The optimal policy depends on the value of γ. As such, the optimal policy when $\gamma \geq 0.232$, we want to choose the policy $(a_1, a_2) = (\text{“go”}, \text{“go”})$. Otherwise, when $\gamma < 0.232$, we want the policy $(a_1, a_2) = (\text{“stay”}, \text{“stay”})$. The sum of discounted rewards can be found below for both options:

   Assume we start at S_1 with policy $(a_1, a_2) = (\text{“stay”}, \text{“stay”})$ when $\gamma < 0.232$. This results in a sum of discounted rewards of $\sum_{t=0}^{\infty}(-1)\gamma^t = -\frac{1}{1-\gamma}$. So, when $\gamma = 0.232$, we get the final reward as -1.303. Now, when $\gamma = 0$, we get the final reward as -1. This means this case's sum of discounted rewards is bounded between -1 and -1.303.

   Assume we start at S_1 and use the policy $(a_1, a_2) = (\text{“go”}, \text{“go”})$ when $\gamma \geq 0.232$. This is optimal because a high γ value incentivizes future rewards 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 $\sum_{t=0}^{1}\gamma^t r_t(s_t, a_t) = \gamma^0 r_0(S_1, \text{“go”}) + \gamma^1 r_1(S_2, \text{“go”}) = -2 + \gamma(3) = 3\gamma - 2$. So, when $\gamma = 0.232$, we get the final reward as -1.303. Now, when $\gamma = 1$, we get the final reward as $+1$. This means this case's sum of discounted rewards is bounded between -1.303 and $+1$.

   As shown above, the cutoff point between the two action policies is at $\gamma = 0.232$, and the optimal policies are provided for the over and under cases.

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. $||T(V) - T(V')||_\infty$
$= ||max_a \sum_{s'} p(s'|s,a)[r(s,a) + \gamma V(s')] - max_a \sum_{s'} p(s'|s,a)[r(s,a) + \gamma V'(s')]||_\infty$
$\leq max_a || \sum_{s'} p(s'|s,a)[r(s,a) + \gamma V(s')] - \sum_{s'} p(s'|s,a)[r(s,a) + \gamma V'(s')]||_\infty$
$= max_a \gamma || \sum_{s'} p(s'|s,a)[r(s,a) + V(s')] - \sum_{s'} p(s'|s,a)[r(s,a) + V'(s')]||_\infty$
$= max_a \gamma \sum_{s'} p(s'|s,a)||[r(s,a) + V(s')] - [r(s,a) + V'(s')]||_\infty$
$= max_a \gamma \sum_{s'} p(s'|s,a)||V(s') - V'(s')||_\infty$
$\leq max_a \gamma \sum_{s'} p(s'|s,a)||V - V'||_\infty$
$= \gamma ||V - V'||_\infty max_a \sum_{s'} p(s'|s,a)$
$= \gamma ||V - V'||_\infty$ (We know $\sum_{s'} p(s'|s,a) = 1$)

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

3. We want to show $\forall \epsilon > 0, \exists N > 0 s.t. \forall n > N \ ||V^{n+1} - V^*||_\infty \leq \frac{\gamma}{1-\gamma}\epsilon$. This is shown below:

   $||V^{n+1} - V^*||_\infty$
   $= ||T(V^n) - T(V^*)||_\infty$
   $\leq \gamma ||V^n - V^*||_\infty$ (from the proof in question 2.2)
   $\leq \frac{\gamma}{1-\gamma} ||V^n - V^*||_\infty$ (we assume γ is between 0 and 1)

   From the above, we've shown $||V^{n+1} - V^*||_\infty \leq \frac{\gamma}{1-\gamma}||V^n - V^*||_\infty$. This indicates the distance between V^{n+1} and V^* shrinks over time (i.e. converges). As such, $\exists n$ s.t. $||V^n - V^*||_\infty \leq \epsilon$ (this n, in practice, is usually found when your program stops, indicating $||V^n - V^{n-1}||_\infty \leq \epsilon$). With this property, we see the following:

   $||V^{n+1} - V^*||_\infty \leq \frac{\gamma}{1-\gamma}||V^n - V^*||_\infty$
   $\implies ||V^{n+1} - V^*||_\infty \leq \frac{\gamma}{1-\gamma}\epsilon$

   $\therefore$ We have shown $||V^{n+1} - V^*||_\infty \leq \frac{\gamma}{1-\gamma}\epsilon \quad \square$

4. Did not do this bonus question.

3 Learning the Model

1. Did not do this bonus question.

2. Did not do this bonus question.

3. Did not do this bonus question.

4. Did not do this bonus question.

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. We are first going to calculate the variance $\text{Var}(\frac{1}{N}\sum_{i=1}^{N}(R(\tau_i)-b)\nabla_\theta log\,\pi_\theta(\tau_i))$. We can use the rule that $Var(x) = \text{E}[x^2] - \text{E}[x]^2$ to solve this:

$$
\begin{aligned}
&\text{Var}(\tfrac{1}{N}\textstyle\sum_{i=1}^{N}(R(\tau_i)-b)\nabla_\theta log\,\pi_\theta(\tau_i)) \\
&= \text{Var}(\text{E}_{\tau\sim\pi_\theta\theta(\tau)}\big[(R(\tau)-b)\nabla_\theta log\,\pi_\theta(\tau)\big]) \\
&= \text{Var}(\text{E}_{\tau\sim\pi_\theta\theta(\tau)}\big[\nabla_\theta log\,\pi_\theta(\tau)(R(\tau)-b)\big]) \\
&= \text{E}_{\tau\sim\pi_\theta\theta(\tau)}\big[(\nabla_\theta log\,\pi_\theta(\tau)(R(\tau)-b))^2\big] - \text{E}_{\tau\sim\pi_\theta\theta(\tau)}\big[\nabla_\theta log\,\pi_\theta(\tau)(R(\tau)-b)\big]^2 \\
&= \text{E}_{\tau\sim\pi_\theta\theta(\tau)}\big[(\nabla_\theta log\,\pi_\theta(\tau)(R(\tau)-b))^2\big] - \text{E}_{\tau\sim\pi_\theta\theta(\tau)}\big[\nabla_\theta log\,\pi_\theta(\tau)R(\tau)\big]^2 \quad \text{(Baseline is unbiased, as we showed in the previous question)} \\
&= \text{E}_{\tau\sim\pi_\theta\theta(\tau)}\big[(\nabla_\theta log\,\pi_\theta(\tau)(R(\tau)-b))^2\big] - \text{E}_{\tau\sim\pi_\theta\theta(\tau)}\big[R(\tau)\nabla_\theta log\,\pi_\theta(\tau)\big]^2
\end{aligned}
$$

As such, the variance is $\text{E}_{\tau\sim\pi_\theta\theta(\tau)}\big[(\nabla_\theta log\,\pi_\theta(\tau)(R(\tau)-b))^2\big] - \text{E}_{\tau\sim\pi_\theta\theta(\tau)}\big[R(\tau)\nabla_\theta log\,\pi_\theta(\tau)\big]^2$. We can see the baseline will impact the first term's values, reducing its values and thus reducing the variance. As such, subtracting b helps reduce the variance of $\nabla_\theta J(\theta)$.

Now, we will calculate the baseline value leading to the least variance. To do this, we need to calculate the gradient of the variance with respect to the baseline b and set it to 0 and solve for b:

$$
\begin{aligned}
&\frac{\delta \text{Var}}{\delta b} \\
&= \frac{\delta}{\delta b}\text{E}_{\tau\sim\pi_\theta\theta(\tau)}\big[(\nabla_\theta log\,\pi_\theta(\tau)(R(\tau)-b))^2\big] - \text{E}_{\tau\sim\pi_\theta\theta(\tau)}\big[R(\tau)\nabla_\theta log\,\pi_\theta(\tau)\big]^2 \\
&= \frac{\delta}{\delta b}\text{E}_{\tau\sim\pi_\theta\theta(\tau)}\big[(\nabla_\theta log\,\pi_\theta(\tau)(R(\tau)-b))^2\big] \quad \text{(second term doesn't depend on } b\text{)} \\
&= \frac{\delta}{\delta b}\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau)(R(\tau)-b))^2\big] \\
&= \frac{\delta}{\delta b}\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2(R(\tau)-b)^2\big] \\
&= \frac{\delta}{\delta b}\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2(R(\tau)^2-2R(\tau)b+b^2)\big] \\
&= \frac{\delta}{\delta b}\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2R(\tau)^2\big] - 2\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2(R(\tau)b)\big] + \text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2(b^2)\big] \\
&= \frac{\delta}{\delta b}\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2R(\tau)^2\big] - 2b\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2R(\tau)\big] + b^2\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2\big] \\
&= \frac{\delta}{\delta b}(-2b)\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2R(\tau)\big] + b^2\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2\big] \quad \text{(first term doesn't depend on } b\text{)} \\
&= -2\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2R(\tau)\big] + 2b\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2\big] = 0 \\
&\implies -2\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2R(\tau)\big] + 2b\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2\big] = 0 \\
&\implies 2b\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2\big] = 2\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2R(\tau)\big] \\
&\implies b\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2\big] = \text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2R(\tau)\big] \\
&\implies b = \frac{\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2R(\tau)\big]}{\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2\big]}
\end{aligned}
$$

As such, the value of b that reduces the variance the most is $b = \frac{\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2R(\tau)\big]}{\text{E}\big[(\nabla_\theta log\,\pi_\theta(\tau))^2\big]}$.