Homework 4

Elias Dubbeldam, Ankur Satya

October 10, 2022

7.4 Geometry of linear value-function approximation

1. The definition of the Bellman operator reads

$$B_\pi v_w(s) = \sum_a \pi(a|s) \cdot \sum_{s',r} p(s', r|s, a)[r + \gamma \cdot v(s')].$$

Let us define the left state as s_0 and the right state as s_1 such that

$$v_w(s_0) = 1 \cdot 2 = 2; \quad v_w(s_1) = 1 \cdot 1 = 1,$$

where we have used the approximation $v_w(s) = w \cdot \phi$. Using the above definitions we can calculate the values of the Bellman operator on the states:

$$B_\pi v_w(s_0) = 1 \cdot 1[0 + 1 \cdot 1 \cdot 1] = 1; \quad B_\pi v_w(s_1) = 1 \cdot 1[0 + 1 \cdot 1 \cdot 2] = 2.$$

Using those we can calculate the Bellman error vector:

$$\begin{aligned}\vec{\delta}_w &= B_\pi \overrightarrow{V_w} - \overrightarrow{V_w} \\ &= [B_\pi v_w(s_0) - v_w(s_0), B_\pi v_w(s_1) - v_w(s_1)]^T \\ &= [1 - 2, 2 - 1]^T \\ &= [-1, 1]^T\end{aligned}$$

2. Mean squared Bellman error is defined such that

$$\begin{aligned}\overrightarrow{VE} &= \sum_{s\epsilon} \mu(s)[B_\pi v_w - v_w(s)] \\ &= \frac{1}{2}\|(-1)^2 + 1^2\| = 1\end{aligned}$$

3. We find the w resulting in the value functions that is the closest to the target values by writing down a least-square expression

$$\begin{gathered}\frac{1}{2} \cdot (w \cdot \phi_0 - B_\pi v_w(s_0))^2 + \frac{1}{2} \cdot (w \cdot \phi_1 - B_\pi v_w(s_1))^2 \\ \frac{1}{2} \cdot (w \cdot 2 - 1)^2 + \frac{1}{2} \cdot (w \cdot 1 - 2)^2\end{gathered}$$

By differentiating the above expression w.r.t. w and setting it equal to zero, i.e. $\frac{\partial}{\partial w}[\ldots] = 0$, we find a value of $w = \frac{8}{10}$.

4. Figure 1 shows the required graph. The vectors v_w and $B_\pi v_w$ are created using the values obtained in part 1 of the question. The vector $\prod B_\pi v_w$ is the projection of v_w on the Bellman vector where projection is defined as the error vector. The distance between v_w and $B_\pi v_w$, i.e. $\prod B_\pi v_w$, is the smallest when $w = 8/10$.

7.5 Coding Assignment - Deep Q Networks

1. See the notebook

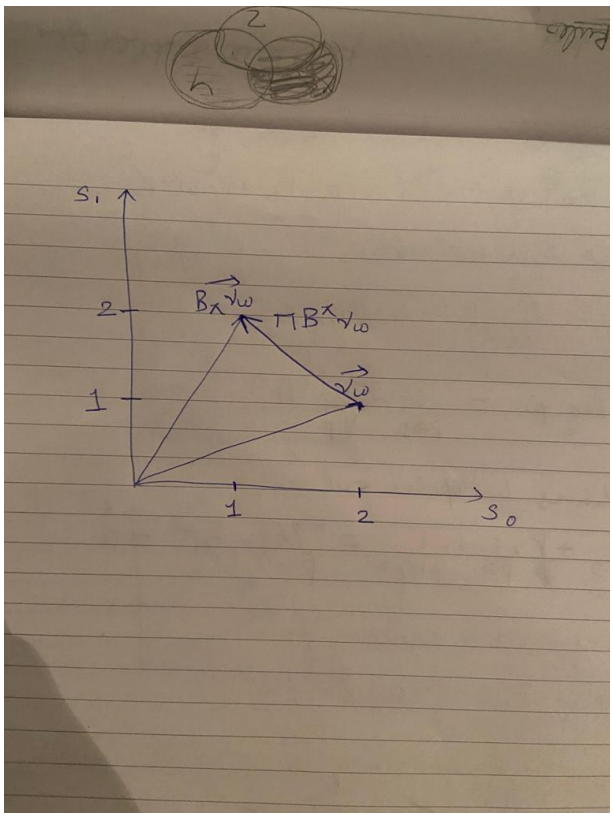

Figure 1: $v_w, B_\pi v_w, \prod B_\pi v_w$

2. A way to use the tabular approach is to divide the state space into bins. This would limit the state space to a finite value instead of the infinite state space in the case of continuous state space. Since our action space is already discretized, we don't need to worry about it. Discretizing the state space in the cart-pole problem would work because we have a hard limit on the state space features like the angle of the pole and the movement of the cart from the center position so using the bin method to discretize them would not create too many states and hence not create a memory storage problem while using the tabular approach.

   Tabular approach is not fit for the problems that would create memory issues while dealing with the Q-table. Also, number of states should be small enough so that they can be sampled multiple times. An example where this would not work is an agent trying to learn drive autonomously because of the following two reasons:

   - The action space would not be as simple as cart pole's problem i.e[+1, -1]. It would consist of a high number of actions.
   - The state space would be high as well which would make the Q-table big enough to create memory problems.

8.4 REINFORCE

1. a) Classical REINFORCE

$$\begin{aligned}\theta_{t+1} &= \theta_t + \alpha \nabla \widehat{J(\theta_t)} \\ &= \theta_t + \alpha \frac{1}{N} \sum_{i=1}^{N} [G(\tau_i) \cdot \sum_{t=0}^{T} \nabla_\theta log(p_\theta(a_t^i | s_t^i))]\end{aligned}$$

   We do it for each episode separately, so $N = 1$.

For Episode 1:

$$\begin{aligned}\theta^a_{t+1} = \theta^a_t + \alpha\frac{1}{1}\cdot(200-20-2-3)\cdot[&\nabla_{\theta_a} log(p_{\theta_a}(a_1=1|s=A))\\ &+\nabla_{\theta_a} log(p_{\theta_a}(a_2=2|s=A))\\ &+\nabla_{\theta_a} log(p_{\theta_a}(a_3=2|s=B))\\ &+\nabla_{\theta_a} log(p_{\theta_a}(a_4=2|s=B))]\\ \\ &= \theta^a_t + 175\alpha[\nabla_{\theta_a} log(p_{\theta_a}(a_1=1|s=A))\\ &+\nabla_{\theta_a} log(p_{\theta_a}(a_2=2|s=A))\\ &+0\\ &+0]\end{aligned}$$

$$\begin{aligned}\theta^b_{t+1} = \theta^b_t + \alpha\frac{1}{1}\cdot(200-20-2-3)\cdot[&\nabla_{\theta_b} log(p_{\theta_b}(a_1=1|s=A))\\ &+\nabla_{\theta_b} log(p_{\theta_b}(a_2=2|s=A))\\ &+\nabla_{\theta_b} log(p_{\theta_b}(a_3=2|s=B))\\ &+\nabla_{\theta_b} log(p_{\theta_b}(a_4=2|s=B))]\\ \\ &= \theta^b_t + 175\alpha[0\\ &+0\\ &+\nabla_{\theta_b} log(p_{\theta_b}(a_3=2|s=B))\\ &+\nabla_{\theta_b} log(p_{\theta_b}(a_4=2|s=B))]\end{aligned}$$

For Episode 2:

$$\begin{aligned}\theta^a_{t+1} = \theta^a_t + \alpha\frac{1}{1}\cdot(-100+20+10+10)\cdot[&\nabla_{\theta_a} log(p_{\theta_a}(a_1=1|s=A))\\ &+\nabla_{\theta_a} log(p_{\theta_a}(a_2=2|s=A))\\ &+\nabla_{\theta_a} log(p_{\theta_a}(a_3=1|s=B))\\ &+\nabla_{\theta_a} log(p_{\theta_a}(a_4=1|s=B))]\\ \\ &= \theta^a_t - 60\alpha[\nabla_{\theta_a} log(p_{\theta_a}(a_1=1|s=A))\\ &+\nabla_{\theta_a} log(p_{\theta_a}(a_2=2|s=A))\\ &+0\\ &+0]\end{aligned}$$

$$\begin{aligned}\theta^b_{t+1} = \theta^b_t + \alpha\frac{1}{1}\cdot(-100+20+10+10)\cdot[&\nabla_{\theta_b} log(p_{\theta_b}(a_1=1|s=A))\\ &+\nabla_{\theta_b} log(p_{\theta_b}(a_2=2|s=A))\\ &+\nabla_{\theta_b} log(p_{\theta_b}(a_3=1|s=B))\\ &+\nabla_{\theta_b} log(p_{\theta_b}(a_4=1|s=B))]\\ \\ &= \theta^b_t - 60\alpha[0\\ &+0\\ &+\nabla_{\theta_b} log(p_{\theta_b}(a_3=1|s=B))\\ &+\nabla_{\theta_b} log(p_{\theta_b}(a_4=1|s=B))]\end{aligned}$$

b) REINFORCE/G(PO)MDP

$$\begin{aligned}\theta_{t+1} &= \theta_t + \alpha\nabla\widehat{J(\theta_t)}\\ &= \theta_t + \alpha\frac{1}{N}\sum_{i=1}^{N}[\sum_{t=1}^{T} r_t\cdot\sum_{t'=1}^{t}\nabla_\theta log(p_\theta(a^i_{t'}|s^i_{t'}))]\end{aligned}$$

We do it for each episode separately, so $N = 1$.

For Episode 1:

$$\begin{aligned}
\theta^a_{t+1} &= \theta^a_t + \alpha\frac{1}{1}[r_1[\nabla_a log(p(a_1|s_1))] \\
&\quad + r_2[\nabla_a log(p(a_1|s_1)) + \nabla_a log(p(a_2|s_2))] \\
&\quad + r_3[\nabla_a log(p(a_1|s_1)) + \nabla_a log(p(a_2|s_2)) + \nabla_a log(p(a_3|s_3))] \\
&\quad + r_4[\nabla_a log(p(a_1|s_1)) + \nabla_a log(p(a_2|s_2)) + \nabla_a log(p(a_3|s_3)) + \nabla_a log(p(a_4|s_4))]] \\
&= \theta^a_t + \alpha[200\nabla_a log(p(a_1|s_1)) \\
&\quad - 20[\nabla_a log(p(a_1|s_1)) + \nabla_a log(p(a_2|s_2))] \\
&\quad - 2[\nabla_a log(p(a_1|s_1)) + \nabla_a log(p(a_2|s_2)) + 0] \\
&\quad - 3[\nabla_a log(p(a_1|s_1)) + \nabla_a log(p(a_2|s_2)) + 0 + 0]] \\
&= \theta^a_t + \alpha[175 \cdot \nabla_a log(p(a_1 = 1|s_1 = A)) - 25 \cdot \nabla_a log(p(a_2 = 2|s_2 = A))]
\end{aligned}$$

$$\begin{aligned}
\theta^b_{t+1} &= \theta^b_t + \alpha\frac{1}{1}[r_1[\nabla_b log(p(a_1|s_1))] \\
&\quad + r_2[\nabla_b log(p(a_1|s_1)) + \nabla_b log(p(a_2|s_2))] \\
&\quad + r_3[\nabla_b log(p(a_1|s_1)) + \nabla_b log(p(a_2|s_2)) + \nabla_b log(p(a_3|s_3))] \\
&\quad + r_4[\nabla_b log(p(a_1|s_1)) + \nabla_b log(p(a_2|s_2)) + \nabla_b log(p(a_3|s_3)) + \nabla_a log(p(a_4|s_4))]] \\
&= \theta^b_t + \alpha[200 \cdot 0 \\
&\quad - 20[0 + 0] \\
&\quad - 2[0 + 0 + \nabla_b log(p(a_3|s_3))] \\
&\quad - 3[0 + 0 + \nabla_b log(p(a_3|s_3)) + \nabla_b log(p(a_4|s_4))]] \\
&= \theta^b_t + \alpha[-5 \cdot \nabla_b log(p(a_3 = 2|s_3 = B)) - 3 \cdot \nabla_b log(p(a_4 = 2|s_4 = B))]
\end{aligned}$$

For Episode 2:

$$\begin{aligned}
\theta^a_{t+1} &= \theta^a_t + \alpha\frac{1}{1}[r_1[\nabla_a log(p(a_1|s_1))] \\
&\quad + r_2[\nabla_a log(p(a_1|s_1)) + \nabla_a log(p(a_2|s_2))] \\
&\quad + r_3[\nabla_a log(p(a_1|s_1)) + \nabla_a log(p(a_2|s_2)) + \nabla_a log(p(a_3|s_3))] \\
&\quad + r_4[\nabla_a log(p(a_1|s_1)) + \nabla_a log(p(a_2|s_2)) + \nabla_a log(p(a_3|s_3)) + \nabla_a log(p(a_4|s_4))]] \\
&= \theta^a_t + \alpha[-100\nabla_a log(p(a_1|s_1)) \\
&\quad + 20[\nabla_a log(p(a_1|s_1)) + \nabla_a log(p(a_2|s_2))] \\
&\quad + 10[\nabla_a log(p(a_1|s_1)) + \nabla_a log(p(a_2|s_2)) + 0] \\
&\quad + 10 = [\nabla_a log(p(a_1|s_1)) + \nabla_a log(p(a_2|s_2)) + 0 + 0]] \\
&= \theta^a_t + \alpha[-60 \cdot \nabla_a log(p(a_1 = 1|s_1 = A)) + 40 \cdot \nabla_a log(p(a_2 = 2|s_2 = A))]
\end{aligned}$$

$$
\begin{aligned}
\theta^b_{t+1} &= \theta^b_t + \alpha\frac{1}{1}[r_1[\nabla_b log(p(a_1|s_1))] \\
&\quad + r_2[\nabla_b log(p(a_1|s_1)) + \nabla_b log(p(a_2|s_2))] \\
&\quad + r_3[\nabla_b log(p(a_1|s_1)) + \nabla_b log(p(a_2|s_2)) + \nabla_b log(p(a_3|s_3))] \\
&\quad + r_4[\nabla_b log(p(a_1|s_1)) + \nabla_b log(p(a_2|s_2)) + \nabla_b log(p(a_3|s_3)) + \nabla_a log(p(a_4|s_4))]] \\
&= \theta^b_t + \alpha[-100 \cdot 0 \\
&\quad + 20[0 + 0] \\
&\quad + 10[0 + 0 + \nabla_b log(p(a_3|s_3))] \\
&\quad + 10[0 + 0 + \nabla_b log(p(a_3|s_3)) + \nabla_b log(p(a_4|s_4))]] \\
&= \theta^b_t + \alpha[20 \cdot \nabla_b log(p(a_3 = 1|s_3 = B)) + 10 \cdot \nabla_b log(p(a_4 = 1|s_4 = B))]
\end{aligned}
$$

2. For the classical REINFORCE updates, the probability of taking action 1 in state B will decrease. This is because when the action 1 is taken in state B, the weights are updated with a negative return of -60. In a similar fashion, we expect the probability to increase for action 2. Because when this action is taken, the weights are updated with a positive return (+175)

   In the REINFORCE/G(PO)MDP case it is the other way around: the probability of taking action 2 in state B will decrease (negative rewards of -5 and -3) and taking action 1 in state B will increase (positive rewards of +10 and +10).

3. REINFORCE/G(PO)MDP provides a better update since it only takes the future rewards into account. The classical REINFORCE method takes the rewards of all timesteps into account. In this example this leads to overestimated and underestimated values of respectively action 2 and 1 in state B. The rewards when taking action from state B are incorrect because the rewards of earlier states are taken into account. This is not the case for REINFORCE/G(PO)MDP, as we can also see in the update values above.

4. The variance would be less in the REINFORCE/G(PO)MDP method since the number of steps over which reward is summed up are less compared to the classical REINFORCE method. taking more timesteps into account results in more random variables which result in a higher variance. Therefore, the classical REINFORCE method has more variance.

5. We can include the value function along with policy update methods. The most straightforward way to do this would be to include a baseline. The baseline reduces the variance. Typically, an estimated value function is used for the baseline. Using the notation of Equation (27) from the problem sheet, the update rule would look like

$$
\nabla_\theta J(\theta) \approx \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta \left(A_t \mid S_t\right) \sum_{k=t}^{T} \left(R\left(S_k, A_k\right) - b(S_k)\right)
$$

   with $b(S_k)$ the baseline, which is the the value function in this case.