RL: HW 2

Raymond Koopmanschap (11925582), Benjamin Kolb(12416789)

November 7, 2019

5.4 Gradient Descent Methods

1. The value of a state S_t given policy π, $v_\pi(S_t)$ is the expected return starting from that state. Since G_t is exactly a sample return for this specific state, and by definition $\mathbb{E}[G_t|S_t] = v_\pi(S_t)$ it is an unbiased estimate. It does not rely on bootstrapping like TD.

2. We are not using the true value function to update, but some approximation to it. In particular a bootstrap target $r + \gamma\hat{v}(s', \mathbf{w}_t)$. Since this approximation depends on the current value of the weight vector $\mathbf{w}_t$, the results will be biased and not results in a true gradient-descent method. Because they don't take into account the effect on the target, therefore they only calculate a part of the gradient. Hence a semi-gradient.

3. In the mountain car problem the policy that has to be learned, needs to swing from one direction to the other to build momentum. So in order to hit the goal and end the episode, you already need a good policy, otherwise it will never end. Therefore MC methods who continue until the end of the episode can't be used and bootstrapping is necessary because you have to learn from experience without ending the episode.

6.1 Geometry of linear value-function approximation

1. First we can calculate v_w

$$\begin{bmatrix} v_w(s_0) \\ v_w(s_1) \end{bmatrix} = \begin{bmatrix} w\phi_{s_0} \\ w\phi_{s_1} \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Then we use the Bellman operator

$$B_\pi v(s) \doteq \sum_a \pi(a|s) \sum_{s',r} p\left(s', r|s, a\right)\left[r + \gamma v\left(s'\right)\right] \text{ (Sutton and Barto 11.21)}$$

to calculate the Bellman error.

$$\bar{\delta}_w = \begin{bmatrix} r + \gamma v(s_1) \\ r + \gamma v(s_2 \end{bmatrix} - \begin{bmatrix} v_w(s_0) \\ v_w(s_1) \end{bmatrix} = \begin{bmatrix} 2 - 1 \\ 1 - 2 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

2. The Mean Squared Bellman Error is gives by

$$\overline{\text{BE}}(\mathbf{w}) = \left\|\bar{\delta}_{\mathbf{w}}\right\|_{\mu}^{2} = \mu(s_0) * 1^2 + \mu(s_1) * (-1)^2 = \frac{1}{2} * 1 + \frac{1}{2} * 1 = 1$$

$\mu(s_0) = \mu(s_1) = \frac{1}{2}$ since both states are visited the same amount of times and therefore they are equally important.

3. We have to find the w that minimizes $\|B^{\pi} v_w - v_w\|_{\mu}^2$

$$\begin{aligned} \frac{d}{dw}(2-w)^2 + (1-2w)^2 &= 0 \\ \frac{d}{dw} 5w^2 - 8w + 5 &= 0 \\ 10w + 8 &= 0 \\ w &= 0.8 \end{aligned}$$

Thus the projected Bellman operator is

$$\Pi B^{\pi} v_w = \begin{bmatrix} 0.8 \\ 1.6 \end{bmatrix}$$

4. First the optimal Bellman point B_w^{π} is found from v_w. The distance between those two is the Bellman error. However this point doesn't lie in the v_w plane so can't be represented with our function approximation. Therefore it is projected back which produced ΠB_w^{π}, which our function approximation can now represent. Note that the yellow and blue line make a 90 degree angle with each other, since it is a projection

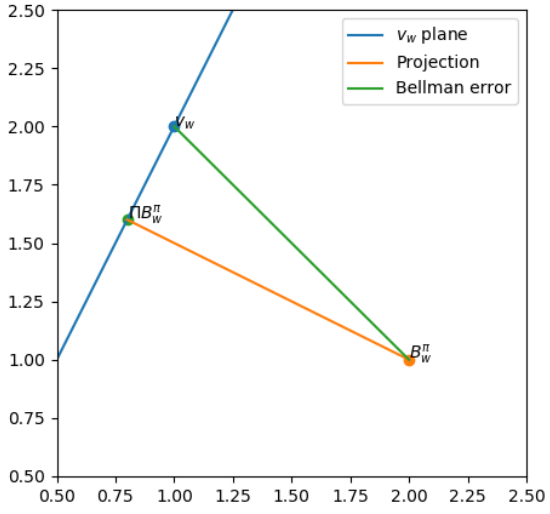

Figure 1: Visualization projected Bellman

8.2 Homework: Compatible Function Approximation Theorem

Part 1

In fact equality (14) is needed to prove this exercise. This is because equality (13) can equally be satisfied for example by $\hat{q}_w = c + w^T \nabla_\theta \log \pi_\theta(s,a)$, with $c \neq 0$ which would result in a result of c in our below computation.

We have for all $s \in S$:

$$\begin{aligned}
\mathbb{E}_a\left[\hat{q}_w(s,a)\right] &= \mathbb{E}_a\left[w^T \nabla_\theta \log \pi_\theta(s,a)\right] \\
&= w^T \mathbb{E}_a\left[\nabla_\theta \log \pi_\theta(s,a)\right] \\
&= w^T \sum_a \pi_\theta(s,a) \nabla_\theta \log \pi_\theta(s,a) \\
&= w^T \sum_a \pi_\theta(s,a) \frac{\nabla_\theta \pi_\theta(s,a)}{\pi_\theta(s,a)} \\
&= w^T \sum_a \nabla_\theta \pi_\theta(s,a) \\
&= w^T \nabla_\theta \sum_a \pi_\theta(s,a) \\
&= w^T \nabla_\theta 1 \\
&= w^T \vec{0} \\
&= 0
\end{aligned}$$

This shows that in the case of the critic in question, the estimated q-values are equal to zero for the average action that is sampled in any state. This is desirable since it reduces the variance of the parameter updates and hence makes learning less slow.

Part 2

We have for all $s \in S$:

$$\begin{aligned}
\mathbb{E}_a\left[A(s,a)\right] &= \mathbb{E}_a\left[q_\pi(s,a) - v_\pi(s)\right] \\
&= \mathbb{E}_a\left[q_\pi(s,a) - \mathbb{E}_a\left[q_\pi(s,a)\right]\right] \\
&= \mathbb{E}_a\left[q_\pi(s,a)\right] - \mathbb{E}_a\left[\mathbb{E}_a\left[q_\pi(s,a)\right]\right] \\
&= \mathbb{E}_a\left[q_\pi(s,a)\right] - \mathbb{E}_a\left[q_\pi(s,a)\right] \\
&= 0
\end{aligned}$$

Part 3

In Part 2 we have seen that the desirable property from Part 1 holds for any policy if we would use $q_\pi(s,a) - v_\pi(s)$ instead of $\hat{q}_w(s,a)$ in our updates. In practice we would need to use $\hat{q}_w(s,a) - \hat{v}_u(s)$ with unbiased $\hat{q}_w$ and $\hat{v}_u$. It is intuitive to use $\hat{v}_u = \hat{v}_{w,\theta} = \mathbb{E}_a\left[\hat{q}_w(s,a)\right]$.

Part 4

We set:

$$\begin{aligned}
\hat{q}_w(s,a) &= w^T \nabla_\theta \log \pi_\theta(s,a) \\
&= w^T \nabla_\theta \log \frac{e^{\theta^T \phi_{sa}}}{\sum_b e^{\theta^T \phi_{sb}}} \\
&= w^T \nabla_\theta (\theta^T \phi_{sa} - \log(\sum_b e^{\theta^T \phi_{sb}})) \\
&= w^T \nabla_\theta \theta^T \phi_{sa} - w^T \nabla_\theta \log(\sum_b e^{\theta^T \phi_{sb}}) \\
&= w^T (\phi_{sa} - \frac{\nabla_\theta \sum_b e^{\theta^T \phi_{sb}}}{\sum_b e^{\theta^T \phi_{sb}}}) \\
&= w^T (\phi_{sa} - \frac{\sum_b \nabla_\theta e^{\theta^T \phi_{sb}}}{\sum_b e^{\theta^T \phi_{sb}}}) \\
&= w^T (\phi_{sa} - \frac{\sum_b \nabla_\theta \theta^T \phi_{sb} e^{\theta^T \phi_{sb}}}{\sum_b e^{\theta^T \phi_{sb}}}) \\
&= w^T (\phi_{sa} - \frac{\sum_b \phi_{sb} e^{\theta^T \phi_{sb}}}{\sum_b e^{\theta^T \phi_{sb}}}) \\
&= w^T (\phi_{sa} - \sum_b \phi_{sb} \frac{e^{\theta^T \phi_{sb}}}{\sum_b e^{\theta^T \phi_{sb}}}) \\
&= w^T (\phi_{sa} - \sum_b \phi_{sb} \pi_\theta(s,a))
\end{aligned}$$

This of course implies the first condition:

$$\begin{aligned}
\nabla_w \hat{q}_w(s,a) &= \nabla_w w^T \nabla_\theta \log \pi_\theta(s,a) \\
&= \nabla_\theta \log \pi_\theta(s,a) \\
\frac{\partial \hat{q}_w(s,a)}{\partial w}
\end{aligned}$$

Now to fulfill the second condition, we need to have:

$$\begin{aligned}
0 &= \mathbb{E}_{s\sim\mu, a\sim\pi(\cdot,s)} \left((q_\pi(s,a) - \hat{q}_w(s,a)) \nabla_w \hat{q}_w(s,a) \right) \\
\mathbb{E}_{s\sim\mu, a\sim\pi(\cdot,s)} \left(\hat{q}_w(s,a) \nabla_w \hat{q}_w(s,a) \right) &= \mathbb{E}_{s\sim\mu, a\sim\pi(\cdot,s)} \left(q_\pi(s,a) \nabla_w \hat{q}_w(s,a) \right) \\
\mathbb{E}_{s\sim\mu, a\sim\pi(\cdot,s)} \left(\hat{q}_w(s,a) \nabla_\theta \log \pi_\theta(s,a) \right) &= \mathbb{E}_{s\sim\mu, a\sim\pi(\cdot,s)} \left(q_\pi(s,a) \nabla_\theta \log \pi_\theta(s,a) \right) \\
\mathbb{E}_{s\sim\mu, a\sim\pi(\cdot,s)} \left(w^T (\phi_{sa} - \sum_b \phi_{sb} \pi_\theta(s,a)) \nabla_\theta \log \pi_\theta(s,a) \right) &= \mathbb{E}_{s\sim\mu, a\sim\pi(\cdot,s)} \left(q_\pi(s,a) \nabla_\theta \log \pi_\theta(s,a) \right) \\
w^T \mathbb{E}_{s\sim\mu, a\sim\pi(\cdot,s)} \left((\phi_{sa} - \sum_b \phi_{sb} \pi_\theta(s,a)) \nabla_\theta \log \pi_\theta(s,a) \right) &= \mathbb{E}_{s\sim\mu, a\sim\pi(\cdot,s)} \left(q_\pi(s,a) \nabla_\theta \log \pi_\theta(s,a) \right) \\
w^T \mathbb{E}_{s\sim\mu, a\sim\pi(\cdot,s)} \left((\phi_{sa} - \sum_b \phi_{sb} \pi_\theta(s,a)) \nabla_\theta \log \pi_\theta(s,a) \right) &= \mathbb{E}_{s\sim\mu, a\sim\pi(\cdot,s)} \left(q_\pi(s,a) \nabla_\theta \log \pi_\theta(s,a) \right)
\end{aligned}$$

This is a linear equation to be solved for w. This is however only possible if the matrix $\mathbb{E}_{s\sim\mu,a\sim\pi(\cdot,s)}\left((\phi_{sa}-\sum_b \phi_{sb}\pi_\theta(s,a))\nabla_\theta \log \pi_\theta(s,a)\right)$ is invertible. In this case we get:

$$w = ((\mathbb{E}_{s\sim\mu,a\sim\pi(\cdot,s)}\left((\phi_{sa}-\sum_b \phi_{sb}\pi_\theta(s,a))\nabla_\theta \log \pi_\theta(s,a)\right))^{-1}\mathbb{E}_{s\sim\mu,a\sim\pi(\cdot,s)}\left(q_\pi(s,a)\nabla_\theta \log \pi_\theta(s,a)\right))^T$$

Note that the second condition is in particular fulfilled in the case where we have some w so that $q_\pi(s,a) = \hat{q}_w(s,a) = w^T(\phi_{sa}-\sum_b \phi_{sb}\pi_\theta(s,a))$ for all a and s.