NAME1: Nikita Tokovenko
STUDENTID1:12185892
MAIL1:
nikita.tokovenko@student.uva.nl

Homework Assignment 2
Reinforcement Learning, 19/20

2019-10-06

1 Gradient Descent Methods

1.1

Monte Carlo methods use sampled episodes to obtain real rewards by directly interacting with an environment. Thus, we can obtain real target.

$$\mathbb{E}_\pi[G_t - V_\pi(s)|s_t = s] = \mathbb{E}_\pi[G_t - \mathbb{E}_\pi[G_t|s_t = s]|s_t = s] = \mathbb{E}_\pi[G_t|s_t = s] - \mathbb{E}_\pi[G_t|s_t = s] = 0$$

1.2

DP target depends on the current value of the weight vector $\mathbf{w_t}$, that is used for approximating the estimated value that might not be the same as the true value. Thus, taking into account the effect of changing the weight vector on the estimate, but ignore its effect on target. This affects the form of an update rule for the weight vector including only the part of the gradient thus resulting in a semi-gradient method.

1.3

One of the main disadvantages of using MC approaches is that we have to wait until the end of every episode to learn from it. This can cause problems for the Mountain Car problem because reaching the top of the hill might take too long for initial bad policy. In turn, bootstrapping methods allow us to learn at each timestep by estimating future rewards. Although these methods are biased, using them leads our training model to converge to the optimal policy within reasonable time that is really advantageous w.r.t approaches that might not be able to learn anything for a long time because of ceasing to reach the terminal state thus accomplishing at least one episode.

2 Geometry of linear value-function approximation

2.1

$$\overline{\delta_w} = B^\pi v_w - v_w = R^\pi + \gamma P^\pi v_w - v_w = [v_w = w\phi] = \begin{bmatrix}0\\0\end{bmatrix} + 1 * \begin{vmatrix}0 & 1\\1 & 0\end{vmatrix}\begin{bmatrix}1\\2\end{bmatrix} - \begin{bmatrix}1\\2\end{bmatrix} = \begin{bmatrix}2\\1\end{bmatrix} - \begin{bmatrix}1\\2\end{bmatrix} = \begin{bmatrix}1\\-1\end{bmatrix}$$

2.2

$$MSBE = \frac{\|\overline{\delta_w}\|_{l_2}^2}{\#states} = \frac{1^2 + (-1)^2}{2} = 1$$

2.3

$$\begin{aligned}
w^* &= argmin_w(\|B^\pi v_w - v_w\|_{l_2}^2)\\
\|B^\pi v_w - v_w\|_{l_2}^2 &= \|\begin{bmatrix}2\\1\end{bmatrix} - \begin{bmatrix}w\\2w\end{bmatrix}\|_{l_2}^2 = (2-w)^2 + (1-2w)^2\\
\frac{\partial}{\partial w}(\|B^\pi v_w - v_w\|_{l_2}^2) &= -2(2-w) - 4(1-2w) = 10w - 8 = 0\\
w^* = 0.8 \Rightarrow v_{w^*} &= \Pi B^\pi v_w = \begin{bmatrix}0.8\\1.6\end{bmatrix}
\end{aligned}$$

2.4

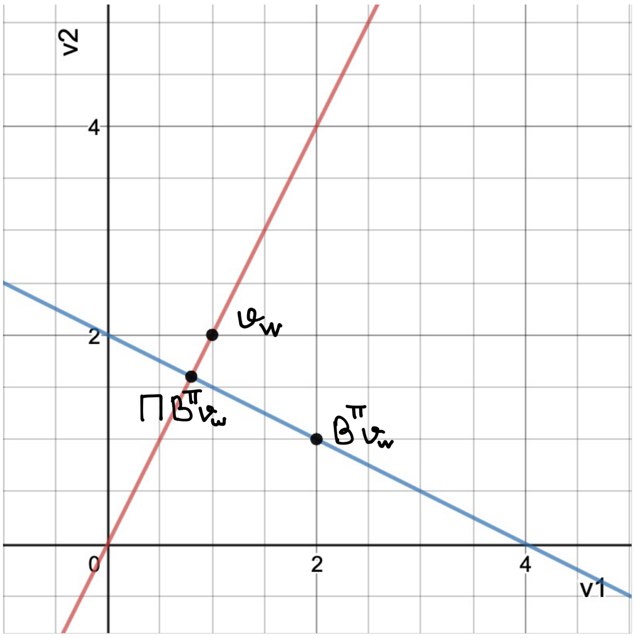

Figure 1: Plotted $v_w, B^\pi v_w$ and $\Pi B^\pi v_w$

In Figure 1 you can see value function, defined for $w = 1$. By projecting it onto the $B^\pi v_w$ we obtain the point that corresponds to the value of w, that is closest to $B^\pi v_w$ in least-square sense.

3 Compatible Function Approximation Theorem

3.1

$$\mathbb{E}_a[\hat{q}_w(s,a)] = \mathbb{E}_a[w^T \nabla_\theta \log \pi_\theta(s,a)] = w^T \sum_a \pi_\theta(s,a) \nabla_\theta \log \pi_\theta(s,a) =$$
$$w^T \sum_a \pi_\theta(s,a) \frac{\nabla_\theta \pi_\theta(s,a)}{\pi_\theta(s,a)} = w^T \sum_a \nabla_\theta \pi_\theta(s,a) = w^T \nabla_\theta \sum_a \pi_\theta(s,a) =$$
$$w^T \nabla_\theta 1 = w^T 0 = 0, \forall s \in S$$

This result shows that in such formulation approximator only gives biased estimates (unless true state-action value function is zero-function).

3.2

$$\mathbb{E}_a[q_\pi(s,a) - v_\pi(s)] = \mathbb{E}_a[q_\pi(s,a)] - \mathbb{E}_a[v_\pi(s)] = v_\pi(s) - v_\pi(s) = 0$$

3.3

In such formulation approximator $\hat{q}_w(s,a)$ can be thought of as an unbiased approximator of the Advantage function.

3.4

$$\pi_\theta(s,a) = \frac{e^{\theta^T \phi_{sa}}}{\sum_b e^{\theta^T \phi_{sb}}}$$

$$\nabla_\theta \log \pi_\theta(s,a) = \frac{\nabla_\theta \pi_\theta(s,a)}{\pi_\theta(s,a)} = \frac{\sum_b e^{\theta^T \phi_{sb}}}{e^{\theta^T \phi_{sa}}} \frac{e^{\theta^T \phi_{sa}} \phi_{sa} \sum_b e^{\theta^T \phi_{sb}} - e^{\theta^T \phi_{sa}} \sum_b e^{\theta^T \phi_{sb}} \phi_{sb}}{(\sum_b e^{\theta^T \phi_{sb}})^2} =$$

$$\frac{\phi_{sa} \sum_b e^{\theta^T \phi_{sb}} - \sum_b e^{\theta^T \phi_{sb}} \phi_{sb}}{\sum_b e^{\theta^T \phi_{sb}}} = \phi_{sa} - \sum_b \frac{e^{\theta^T \phi_{sb}} \phi_{sb}}{\sum_c e^{\theta^T \phi_{sc}}} =$$

$$\phi_{sa} - \sum_b \pi_\theta(s,b) \phi_{sb}$$

Thus,

$$\hat{q}_w(s,a) = w^T [\phi_{sa} - \sum_b \pi_\theta(s,b) \phi_{sb}]$$