Dry4 - 046203

Zohar Rimon

17.06.2020

1 Projected Bellman Operator

1.1

Let the length of the features vector be equal to one (k=1), $\phi_1(s) = 1$ and the MDP be an arbiraty two states MDP with a reward of 1 for the first state and 2 for second one (independent on the action and the next state). For this MDP and feature selection, we can show the wanted property. Indeed, for $v = [0, 0] \in S$ we get: $T^\pi(v) = r = [1, 2] \notin S$ (This is because $S = span[1, 1]$)

1.2

We can easily see that the features are:

$$\phi_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad \phi_2 = \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}$$

And thus:

$$\tilde{V} = \phi_1 w_1 + \phi_2 w_2 = (w_1, w_2, 2w_2)^\top$$

We can also write the feature vector for each state:

$$\phi(1) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \phi(2) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \phi(3) = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$$

1.3

Recall that $T^\pi \doteq r + \gamma P^\pi$. In our case:

$$P^\pi = \begin{pmatrix} 0 & p & 1-p \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad r = 0$$

So we get:

$$T^\pi = \gamma \begin{pmatrix} 0 & p & 1-p \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

Plugging in the general form of $\tilde{V}$:

$$T^\pi \tilde{V} = \gamma \begin{pmatrix} 0 & p & 1-p \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \cdot (w_1, w_2, 2w_2)^\top = \gamma \cdot \begin{pmatrix} 2w_2 - w_2 p \\ w_1 \\ w_1 \end{pmatrix}$$

1.4

The stationary distribution will hold $vP^{\pi} = v$. Plugging in the P in our case:

$$(v_1, v_2, v_3) \cdot \begin{pmatrix} 0 & p & 1-p \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} = (v_2 + v_3, p \cdot v_1, v_1 \cdot (1-p))$$

Thus:

$$v_2 = p \cdot v_1, \quad v_3 = v_1 \cdot (1-p)$$

We also want the vector to be a distribution so:

$$v_1 + v_2 + v_3 = v_1 + p \cdot v_1 + v_1 \cdot (1-p) = 1 \Longrightarrow v_1 = \frac{1}{2}$$

So we get:

$$v = (\frac{1}{2}, \frac{p}{2}, \frac{1-p}{2})$$

1.5

Recall the projection operator, as seen in class:

$$\Pi = \Phi \left(\Phi^T \Sigma \Phi\right)^{-1} \Phi^T \Sigma$$

In our case:

$$\Sigma = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{pmatrix}, \quad \Phi = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 2 \end{pmatrix}$$

Thus:

$$\Pi = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 2 \end{pmatrix} \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 2 \end{pmatrix}^T \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 2 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 2 \end{pmatrix}^T \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{4} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{5} & \frac{2}{5} \\ 0 & \frac{2}{5} & \frac{4}{5} \end{pmatrix}$$

1.6

Using the results from previous sections:

$$\tilde{V}' \doteq \Pi T^{\pi} \tilde{V} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{5} & \frac{2}{5} \\ 0 & \frac{2}{5} & \frac{4}{5} \end{pmatrix} \cdot \gamma \cdot \begin{pmatrix} 2w_2 - w_2 p \\ w_1 \\ w_1 \end{pmatrix} = \gamma \cdot \begin{pmatrix} 2w_2 - w_2 p \\ \frac{3}{5} \cdot w_1 \\ \frac{6}{5} \cdot w_1 \end{pmatrix}$$

We know that $\tilde{V}'$ is in the spanned subspace (due to the projection operator, thus we can write its weights:

$$w' = (w_1', w_2') = \gamma(2w_2 - w_2 p, \frac{3}{5} \cdot w_1)$$

1.7

Lets apply ΠT^{π} one more time, using the wight update formula that we got in the previous subsection:

$$w'' = (w_1'', w_2'') = \gamma(2w_2' - w_2' p, \frac{3}{5} \cdot w_1') = \gamma^2((2-p)\frac{3}{5}w_1, (2-p)\frac{3}{5}w_2)$$

Lets assume that $w_1 \neq 0$, and that we are on an even iteration, when continuing to update will be :

$$w_1^n = \gamma^n (2-p)^{n-1} w_1$$

If we will take $p < 2 - \frac{1}{\gamma}$ we will get divergence.

2 Approximate Value Iteration

2.1

Using the fact that V^* is a fixed point of the optimal bellman operator T.

$$\left|v^k - v^*\right|_\infty = \left|v^k - Tv^*\right|_\infty = \left|v^k - Tv^{k-1} + Tv^{k-1} - Tv^*\right|_\infty$$

Using the triangle inequality:

$$\left|v^k - Tv^{k-1} + Tv^{k-1} - Tv^*\right|_\infty \leq \left|v^k - Tv^{k-1}\right|_\infty + \left|Tv^{k-1} - Tv^*\right|_\infty$$

Using the approximate Value Iteration scheme:

$$\left|v^k - Tv^{k-1}\right|_\infty + \left|Tv^{k-1} - Tv^*\right|_\infty \leq \epsilon + \left|Tv^{k-1} - Tv^*\right|_\infty$$

And Using the contraction property of the bellman operator:

$$\epsilon + \left|Tv^{k-1} - Tv^*\right|_\infty \leq \epsilon + \gamma \cdot \left|v^{k-1} - v^*\right|_\infty$$

So we get:

$$\left|v^k - v^*\right|_\infty \leq \epsilon + \gamma \cdot \left|v^{k-1} - v^*\right|_\infty$$

We can iteratively use this connection in order to bound $\left|v^k - v^*\right|_\infty$ using $\left|v^0 - v^*\right|_\infty$:

$$\begin{aligned}\left|v^k - v^*\right|_\infty &\leq \epsilon + \gamma \cdot \left|v^{k-1} - v^*\right|_\infty \leq \epsilon + \gamma\epsilon + \gamma^2\epsilon + \cdots + \gamma^{k-1}\epsilon + \gamma^k \cdot \left|v^0 - v^*\right|_\infty \\ &= \epsilon \cdot \sum_{i=0}^{k-1} \gamma^i + \gamma^k \cdot \left|v^0 - v^*\right|_\infty = \epsilon \cdot \frac{1-\gamma^k}{1-\gamma} + \gamma^k \cdot \left|v^0 - v^*\right|_\infty \leq \epsilon \cdot \frac{1}{1-\gamma} + \gamma^k \cdot \left|v^0 - v^*\right|_\infty\end{aligned}$$

Now we can consider v^k as an approximation to the optimal value function, and by using the proposition we saw in class, we get:

$$\left|v^{\pi_k} - v^*\right|_\infty \leq \left[\epsilon \cdot \frac{1}{1-\gamma} + \gamma^k \cdot \left|v^0 - v^*\right|_\infty\right] \cdot \frac{2 \cdot \gamma}{1-\gamma} = \frac{2\gamma^{k+1}}{1-\gamma}\left|v_0 - v^*\right|_\infty + \frac{2\gamma\epsilon}{(1-\gamma)^2}$$

2.2

We can easily see that we get a constant upper bound for $k \longrightarrow \infty$:

$$\frac{2\gamma^{k+1}}{1-\gamma}\left|v_0 - v^*\right|_\infty + \frac{2\gamma\epsilon}{(1-\gamma)^2} \longrightarrow \frac{2\gamma\epsilon}{(1-\gamma)^2}$$

Thus the guaranty for large ks:

$$\left|v^{\pi_k} - v^*\right|_\infty \leq \frac{2\gamma\epsilon}{(1-\gamma)^2}$$

3 Multiple Step Return and Algorithms

3.1

In class we saw the TD(0) algorithm:

$$\hat{V}\left(s_n\right) := \hat{V}\left(s_n\right) + \alpha_n\left[r_n + \gamma\hat{V}\left(s_{n+1}\right) - \hat{V}\left(s_n\right)\right]$$

This algorithm is based on the fixed policy bellman operator, and the fixed-policy value iteration algorithm:

$$V_{n+1}(s) = E^\pi\left(r(s,a) + \gamma V_n\left(s'\right)\right)$$

We can see that instead of taking the expectancy over s', in the TD(0) we take a one-sample target to target.
We can build a similar algorithm that is based on the h-step bellman operator, estimating the expectancy with:

$$\left(T^\pi\right)^h \hat{V}\left(s_n\right) \approx \sum_{m=0}^{h-1} \gamma^m r_{n+m} + \gamma^h \hat{V}\left(s_{n+h}\right)$$

This is an unbiased estimate to the real h-step bellman operator. Like we did in the TD(0) case, we do not want to take this estimate directly due to the noisiness of r and the transitions. We will use it to update the current value function slightly:
The TD(h) algorithm:

- Start with arbitrary $\hat{V}(s_n)$
- For every iteration n, simulate the environment and observe a trajectory of length h and save the rewards $[r_{i+h}]_{i=0}^{h-1}$ and the final trajectory state s_{n+h}
- Update the current value function: $\hat{V}(s_n) := \hat{V}(s_n) + \alpha_n \left(\sum_{m=0}^{h-1} \gamma^m r_{n+m} + \gamma^h \hat{V}(s_{n+h}) - \hat{V}(s_n) \right)$

Now we also want to add the linear function approximation of the value function:
The TD(h) algorithm (with function approximation):

- Start with arbitrary $\hat{V}(s_n)$
- For every iteration n, simulate the environment and observe a trajectory of length h and save the rewards $[r_{i+h}]_{i=0}^{h-1}$ and the final trajectory state s_{n+h}
- Update the current value function: $\theta_{n+1} = \theta_n + \alpha_n \left(\sum_{m=0}^{h-1} \gamma^m r_{n+m} + \gamma^h \phi(s_{n+h})^\top \theta_n - \phi(s_n)^\top \theta_n \right) \phi(s_n)$

3.2

When we wolved the PBE with the 1-step bellman operator we solved:

$$\theta^* = \underset{\theta \in \mathbb{R}^k}{\operatorname{argmin}} \left\| \Phi\theta - (R^\mu + \gamma P^\mu \Phi\theta^*) \right\|_\epsilon^2$$

We can generalize this to the h-step bellman operator:

$$\theta^* = \underset{\theta \in \mathbb{R}^k}{\operatorname{argmin}} \left\| \Phi\theta - \left(\sum_{i=0}^{h} \gamma^i r_i^\mu + \gamma^h \left(P^\mu\right)^h \Phi\theta^* \right) \right\|_\epsilon^2$$

And the solution is similar to the 1-step pbe:

$$\theta^* = C^{-1} d$$

Where:

$$C = \Phi^T \Sigma \left(I - \gamma^h \left(P^\mu\right)^h \right) \Phi, \quad d = \Phi^T \Sigma \sum_{i=0}^{h} \gamma^i r_i^\mu$$

Similar to the LSTD algorithm, we can approximate C and d with batch smaplint in this case as well:

$$\hat{d}_n = \frac{1}{n} \sum_{t=1}^{n} \phi(s_t) \sum_{i=0}^{h} \gamma^i r_i^\mu \left(s_{t+i}, \mu(s_{t+i}) \right)$$

$$\hat{C}_n = \frac{1}{n} \sum_{t=1}^{n} \phi(s_t) \left(I - \gamma^h \left(P^\mu\right)^h \right) \phi^T(s_t)$$

Where n are the number of samples in the batch, where each sample is a trajectory of length h. The batches have to be sampled iid in order to achieve an unbiased estimator.

3.3

First, we will show the contraction property of the projected h-step fixed policy bellman operator (which will be easy because we already know that the projected 1-step fixed-policy bellman operator is a γ contraction operator):

$$\left| \left(T^\pi\right)^h (V_1)(s) - \left(T^\pi\right)^h (V_2)(s) \right| = \left| R_h^\pi(s) + \gamma^h \left(P^\pi\right)^h (V_1)(s) - R_h^\pi(s) - \gamma^h \left(P^\pi\right)^h (V_2)(s) \right|$$

Where:

$$R_h^\pi(s_0) = E^\pi \left[\sum_{i=0}^{h-1} \gamma^i r(s_i, \pi(s_i) | s_0) \right]$$

Thus:

$$\left|\left(T^{\pi}\right)^{h}\left(V_1\right)(s)-\left(T^{\pi}\right)^{h}\left(V_2\right)(s)\right|=\gamma^{h}\left|\left(P^{\pi}\right)^{h}\left(V_1\right)(s)-\left(P^{\pi}\right)^{h}\left(V_2\right)(s)\right|$$

We can observe that $(P^{\pi})^{h}$ is a transition matrix and remember that we have shown in class that a transition matrix is a non-expansions operator. So:

$$\left|\left(T^{\pi}\right)^{h}\left(V_1\right)(s)-\left(T^{\pi}\right)^{h}\left(V_2\right)(s)\right|\leq\gamma^{h}\left|V_1(s)-V_2(s)\right|$$

So $(T^{\pi})^{h}$ is a γ^{h}-contraction operator.

Recall that we showed in class that $\|V^{\mu}-\Phi\theta^{*}\|_{\epsilon}^{2}\leq\frac{1}{1-\gamma^{2}}\|V^{\mu}-\Pi_{\epsilon}V^{\mu}\|_{\epsilon}^{2}$ such that $\Phi\theta^{*}$is the unique fixed point of $\Pi_{\epsilon}T^{\mu}$. In our case, the algorithm will result in the fixed point of $\Pi_{\epsilon}\left(T^{\mu}\right)^{h}$, so:

$$\|V^{\mu}-\Phi\theta^{*}\|_{\epsilon}^{2}\leq\frac{1}{1-\gamma^{2\cdot h}}\|V^{\mu}-\Pi_{\epsilon}V^{\mu}\|_{\epsilon}^{2}$$

Now we will compare the upper bound for different hs.

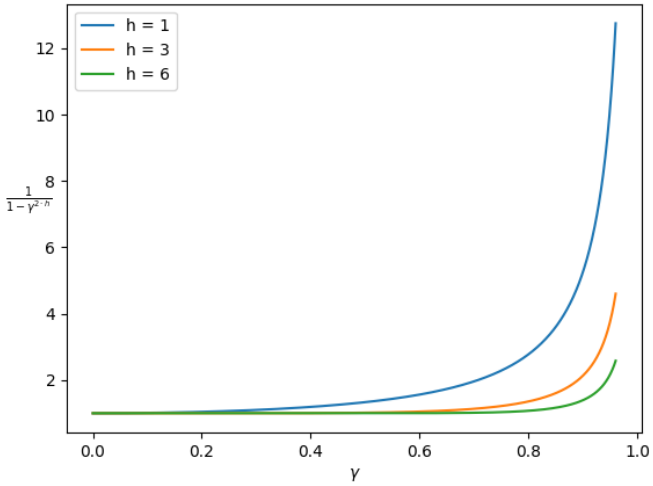

Figure 1: The upper bound for different look-ahead lengths

As you can see in Figure 1, as h increases we get a better upper bound (as it guarantees a lower value for each γ). Still, we have a trade-off due to the fact that as h increases the data collection becomes harder (the trajectories needed are longer).

We can see that as h goes to infinity:

$$\|V^{\mu}-\Phi\theta^{*}\|_{\epsilon}^{2}\leq\frac{1}{1-\gamma^{2\cdot h}}\|V^{\mu}-\Pi_{\epsilon}V^{\mu}\|_{\epsilon}^{2}\longrightarrow\|V^{\mu}-\Pi_{\epsilon}V^{\mu}\|_{\epsilon}^{2}$$

This means that the approximated value from our algorithm ($\Phi\theta^{*}$) is reaching the best approximation that can be achieved using the subspace defined by our features. While this is great, we actually wanted to develop this algorithm due to the fact that collecting such long trajectories are problematic and we wanted an online algorithm that doesn't need to wait until the end of the episode to update the value function.