Exercise Set 4 - Reinforcement Learning

Control with approximation and policy gradients

Giulio Starace - 13010840

October 4, 2022

Homework: Geometry of linear value-function approximation (Application)

1. To compute the Bellman error vector after initialization, we compute the Bellman error for each state. We first recall the definition of the Bellman operator B^π:

$$(B^\pi v)(s) \doteq \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma v(s')\right]. \tag{1}$$

We can plug this into the definition of the Bellman error:

$$\begin{aligned}\overline{\delta}_w(s) &= B^\pi v_w - v_w \\ &= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma v_w(s')\right] - v_w(s).\end{aligned}$$

For s_0, we have a single action available that is always taken, so our $\sum_a \pi(a|s)$ term disappears and we are left with:

$$\overline{\delta}_w(s) = \sum_{s',r} p(s',r|s)\left[r + \gamma v_w(s')\right] - v_w(s) \tag{2}$$

Our action always leads to the same state, with the same reward (of 0), so we can simplify further and write

$$\overline{\delta}_w(s) = \gamma v_w(s') - v_w(s). \tag{3}$$

Finally, we have that $v_{w,s} = w \cdot \phi_s$, so that we can write

$$\overline{\delta}_w(s) = \gamma w \cdot \phi_{s'} - w \cdot \phi_s. \tag{4}$$

The same arguments can be applied to s_1, so we can write the Bellman error *vector* as

$$\mathrm{BE}(w) = \left(\overline{\delta}_w(s_0),\ \overline{\delta}_w(s_1)\right)^T = \left(\gamma w \cdot \phi_{s_1} - w \cdot \phi_{s_0},\ \gamma w \cdot \phi_{s_0} - w \cdot \phi_{s_1}\right)^T. \tag{5}$$

We can plug in our values $w = 1$, $\phi_{s_0} = 2$, $\phi_{s_1} = 1$, and $\gamma = 1$ and obtain

$$\mathrm{BE}(w) = \left(1 \cdot 1 - 1 \cdot 2,\ 1 \cdot 2 - 1 \cdot 1\right)^T = \left(-1,\ 1\right)^T. \tag{6}$$

2. The Mean Squared Bellman Error $\overline{\text{BE}}(w)$ is the measure of the overall error in the value function, computed by taking the μ weighted norm of the Bellman error vector. Here, μ is a distribution $\mu : \mathcal{S} \to [0,1]$ specifying the extent to which each state is considered in the computation. Mathematically:

$$\overline{\text{BE}}(w) = \|\text{BE}(w)\|_\mu^2 = \sum_s \mu(s)\bar{\delta}_w(s)^2. \tag{7}$$

In our case, this would be expressed as

$$\overline{\text{BE}}(w) = \mu(s_0) \cdot (-1)^2 + \mu(s_1) \cdot 1^2 = \mu(s_0) + \mu(s_1) = \sum_s \mu(s). \tag{8}$$

3. The target values $B^\pi v_w$ we found in question 1. are 1 for s_0 and 2 for s_1. The w that results in the value function that is closest can be applied using a least-squares regression:

$$\begin{aligned} \beta &= \arg\min_\beta \|\mathbf{Y} - w\mathbf{X}\|^2 \\ w &= \arg\min_w \|(1,\ 2)^T - w(\phi_{s_0},\ \phi_{s_1})^T\|^2 \\ &= \arg\min_w \|(1,\ 2)^T - w(2,1)^T\|^2. \end{aligned} \tag{9}$$

This has a closed-form solution:

$$\beta = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}, \tag{10}$$

which for our numbers gives $w = 4/5$.

4. The plot of v_w, $B^\pi v_w$ and $\Pi B^\pi v_w$ is shown in Figure 1. TODO

Homework: Coding Assignment - Deep Q Networks

1. Coding answers have been submitted on codegra under the group “stalwart cocky sawly”.

2. hello world

Homework: REINFORCE

1. The update to the policy parameter θ under classical REINFORCE is given by

$$\theta_{t+1} \leftarrow \theta_t + \alpha \nabla J \tag{11}$$

2. hello world

3. hello world

4. hello world

5. hello world

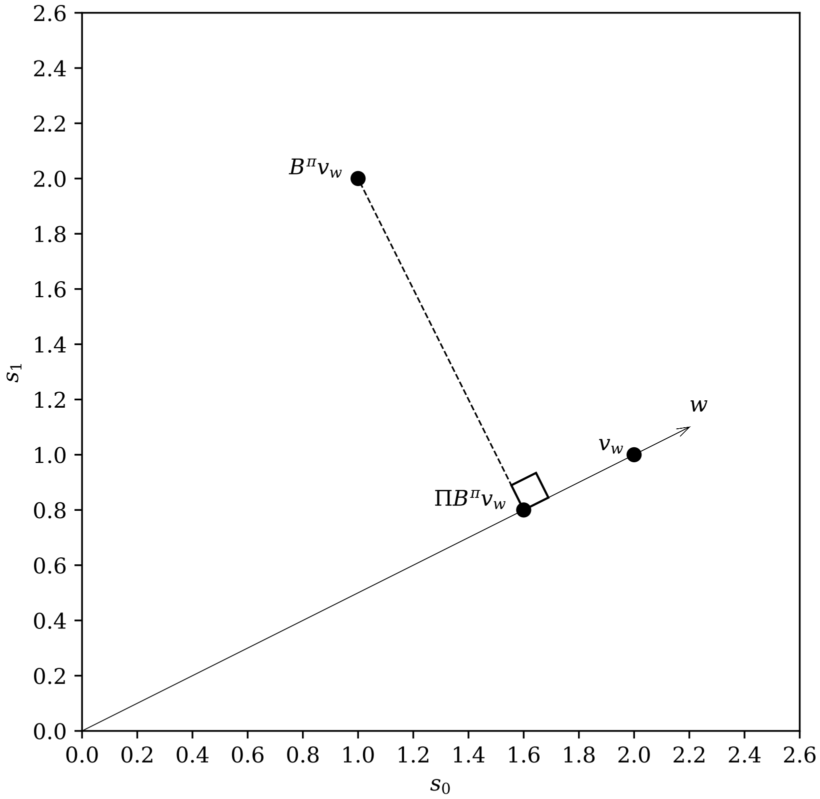

Figure 1: TODO