CMPUT 609 Assignment 3

Shaurya Seth

February 5, 2023

Question 1

It is better to minimize the projected Bellman error because the Bellman error is not learnable. The projected Bellman error can be learned directly from the data distribution and will have a unique minimizer.

Question 2

False. Baird's counterexample shows a case where TD(0) diverges.

Question 3

$$\begin{aligned}
\overline{\text{RE}}(\mathbf{w}) &= \mathbb{E}[(G_t - \hat{v}(S_t, \mathbf{w}))^2] \\
&= \sum_S \mu(s)(G_t - \hat{v}(S_t, \mathbf{w}))^2 \\
&= \sum_S \mu(s)([G_t - v^*(s)] + [v^*(s) - \hat{v}(S_t, \mathbf{w})])^2 \\
&= \sum_S \mu(s)([G_t - v^*(s)]^2 + [v^*(s) - \hat{v}(S_t, \mathbf{w})]^2) \\
&= \overline{\text{VE}}(\mathbf{w}) + \mathbb{E}[(G_t - v_\pi(S_t))^2]
\end{aligned}$$

Question 4

a) In the case of genuine approximation, $\overline{\text{VE}}$ is not expected to have a zero.

b) The $\overline{\text{VE}}$ is zero given by $\overline{\text{VE}}(\mathbf{w}) = ||v_{\mathbf{w}} - v_\pi||_\mu^2$.

c) No, the $\overline{\text{VE}}$ is not learnable. It is not a unique function of the data distribution.

d) Yes, the $\overline{\text{VE}}$ is optimizable.

e) The minimum of $\overline{\text{RE}}$ is greater than the minimum of $\overline{\text{VE}}$ because of the variance term. This can be seen from the equation $\overline{\text{RE}}(\mathbf{w}) = \overline{\text{VE}}(\mathbf{w}) + \mathbb{E}[(G_t - v_\pi(S_t))^2]$.

f) Yes, $\overline{\text{RE}}$ and $\overline{\text{VE}}$ have the same minimizer.

g) Yes, the $\overline{\text{RE}}$ is learnable.

h) If $\mathbf{w}$ is a zero of the $\overline{\text{VE}}$ then $v_{\mathbf{w}}(s) = v_\pi(s)$. Since v_π solves the Bellman equation exactly the $\overline{\text{BE}}(\mathbf{w})$ is zero.

i) If $\mathbf{w}$ is a zero of $\overline{\text{BE}}$ then it solves the Bellman equation exactly and $v_{\mathbf{w}}(s) = v_\pi(s)$. So it is also a zero of the $\overline{\text{VE}}$.

j) No, the $\overline{\text{BE}}$ is not learnable.

k) Yes, the $\overline{\text{PBE}}$ is learnable. It can be directly determined from the data.

l) There will always be a zero of $\overline{\text{PBE}}$ which is $\mathbf{w}_{\text{TD}}$.

m) The $\overline{\text{BE}}$ will generally not be zero at the $\mathbf{w}_{\text{TD}}$.

n) There will always be a zero of $\overline{\text{PBE}}$ which is $\mathbf{w}_{\text{TD}}$.

o) Yes, the $\overline{\text{TDE}}$ is learnable. It can be directly determined from the data.

p) In general, the minimums of the other four measures are non-zero.

q) The minimum of $\overline{\text{TDE}}$ is unlikely to be zero. Minimizing the $\overline{\text{TDE}}$ is naive.

r) No, in general the minimizer of $\overline{\text{TDE}}$ is different from the other measures.

s) Linear semi-gradient TD(0) converges to the TD fixed point. It minimizes the $\overline{\text{PBE}}$ to zero.

Question 5

For each state, the expected number of times it is visited is:

$$\mu(a) = 2$$
$$\mu(b) = 1$$
$$\mu(c) = 2$$

For each state, the Bellman error is:

$$\bar{\delta}_{\mathbf{w}}(a) = (0.5[2+4] + 0.5[1+4]) - 4 = 1.5$$
$$\bar{\delta}_{\mathbf{w}}(b) = (1[1+4]) - 4 = 1$$
$$\bar{\delta}_{\mathbf{w}}(c) = (0.5[2+4] + 0.5[2+4]) - 4 = 2$$

The overall $\overline{\text{BE}}$ for the first MRP is 4.5 while for the second MRP is 9. So these identical appearing MRPs have a different $\overline{\text{BE}}$. Since it is not unique for the data distribution, it is not learnable.

For $\gamma = 1$ the expression for the $\overline{\text{BE}}$ for both MRPs does not depend on w. So the value for w that minimizes the $\overline{\text{BE}}$ cannot be found. The $\overline{\text{BE}}$ is not optimizable.