Patrick Pegus
Mini Project 3
December 10, 2015
CMPSCI-689
Prof. Sridhar Mahadevan

1. (a)

$$\begin{aligned} w_{FP} &= argmin_w \left\| \Pi_\Phi T^\pi(\hat{V}) - \hat{V} \right\| \\ &= argmin_w \left\| \Pi_\Phi \left(R^\pi + \gamma P^\pi \Phi w \right) - \Phi w \right\| \end{aligned}$$

The norm is minimized if

$$\begin{aligned} \Phi w &= \Pi_\Phi \left(R^\pi + \gamma P^\pi \Phi w \right) \\ \Phi w &= \Phi \left(\Phi^T \Phi \right)^{-1} \Phi^T \left(R^\pi + \gamma P^\pi \Phi w \right) \\ w &= \left(\Phi^T \Phi \right)^{-1} \Phi^T \left(R^\pi + \gamma P^\pi \Phi w \right) \\ w &= \left(\Phi^T \Phi \right)^{-1} \Phi^T R^\pi + \gamma \left(\Phi^T \Phi \right)^{-1} \Phi^T P^\pi \Phi w \\ w - \gamma \left(\Phi^T \Phi \right)^{-1} \Phi^T P^\pi \Phi w &= \left(\Phi^T \Phi \right)^{-1} \Phi^T R^\pi \\ \left(I - \gamma \left(\Phi^T \Phi \right)^{-1} \Phi^T P^\pi \Phi \right) w &= \left(\Phi^T \Phi \right)^{-1} \Phi^T R^\pi \\ w &= \left(I - \gamma \left(\Phi^T \Phi \right)^{-1} \Phi^T P^\pi \Phi \right)^{-1} \left(\Phi^T \Phi \right)^{-1} \Phi^T R^\pi \\ w &= \left(\left(\Phi^T \Phi \right) \left(I - \gamma \left(\Phi^T \Phi \right)^{-1} \Phi^T P^\pi \Phi \right) \right)^{-1} \Phi^T R^\pi \\ w &= \left(\Phi^T \Phi - \gamma \Phi^T P^\pi \Phi \right)^{-1} \Phi^T R^\pi \\ w &= \left(\Phi^T \left(\Phi - \gamma P^\pi \Phi \right) \right)^{-1} \Phi^T R^\pi \\ w &= \left(\Phi^T \left(I - \gamma P^\pi \right) \Phi \right)^{-1} \Phi^T R^\pi \end{aligned}$$

(b)

$$\begin{aligned} w_{LS} &= argmin_w \left\| T^\pi(\hat{V}) - \hat{V} \right\| \\ &= argmin_w \left\| R^\pi + \gamma P^\pi \Phi w - \Phi w \right\| \end{aligned}$$

The norm is minimized if

$$\Phi w - \gamma P^\pi \Phi w = \left(I - \gamma P^\pi \right) \Phi w \approx R^\pi$$

This is equivalent to a weighted least squares problem $Ax \approx b$, whose solution is $x = \left(A^T C A \right)^{-1} A^T C b$, where C is a diagonal matrix with entries giving the non-uniform weights measuring the "length" in the space [1]. Let $A = \left(I - \gamma P^\pi \right) \Phi w$, $b = R^\pi$, and $C = I$, if the basis functions uniform importance. Therefore,

$$\begin{aligned} w &= \left(\left(\left(I - \gamma P^\pi \right) \Phi \right)^T I \left(\left(I - \gamma P^\pi \right) \Phi \right) \right)^{-1} \left(\left(I - \gamma P^\pi \right) \Phi \right)^T I R^\pi \\ &= \left(\Phi^T \left(I - \gamma P^\pi \right)^T \left(I - \gamma P^\pi \right) \Phi \right)^{-1} \Phi^T \left(I - \gamma P^\pi \right)^T R^\pi \end{aligned}$$

2. (a)

$$\begin{aligned}
\Delta_{\theta_i} L_i(\theta_i) &= \Delta_{\theta_i} E_{s,a\sim p(\cdot)}\left[(y_i - Q(s,a;\theta_i))^2\right] \\
&= -2E_{s,a\sim p(\cdot)}\left[(y_i - Q(s,a;\theta_i))\,\Delta_{\theta_i} Q(s,a;\theta_i)\right] \\
&= -2E_{s,a\sim p(\cdot)}\left[\left(E_{s'\sim\varepsilon}\left[r + \gamma \max_{a'} Q(s',a';\theta_{i-1}) \middle| s,a\right] - Q(s,a;\theta_i)\right)\Delta_{\theta_i} Q(s,a;\theta_i)\right] \\
&= -2E_{s,a\sim p(\cdot)}\left[\left(\sum_{s'}\left[\left(r + \gamma \max_{a'} Q(s',a';\theta_{i-1})\right)p(s')\right] - Q(s,a;\theta_i)\right)\Delta_{\theta_i} Q(s,a;\theta_i)\right] \\
&= -2E_{s,a\sim p(\cdot)}\sum_{s'}\left[\left(\left(r + \gamma \max_{a'} Q(s',a';\theta_{i-1})\right)p(s') - Q(s,a;\theta_i)\right)\Delta_{\theta_i} Q(s,a;\theta_i)\right] \\
&= -2E_{s,a\sim p(\cdot);s'\sim\varepsilon}\left[\left(r + \gamma \max_{a'} Q(s',a';\theta_{i-1}) - Q(s,a;\theta_i)\right)\Delta_{\theta_i} Q(s,a;\theta_i)\right]
\end{aligned}$$

We can disregard the -2, since a constant multiplier will not affect when the gradient is 0.

(b) TD-gammon approximated a value function for states instead of an action-value function approximation used in Deep Mind. Also, TD-gammon learnt this state value function on-policy, meaning that it follows an exploration policy. On the other hand, Deep Mind learnt about the greedy strategy of choosing the action with the greatest current value.

(c) Deep Mind uses experience replay to smooth the behavior distribution over many previously seen sequences. If experience replay was not used, then feedback loops could occur, meaning the behavior distribution could shift dramatically based on the most recent actions. These feedback loops may cause parameters choices that get stuck on poor local minima or diverge.

(d) The two left graphs show the average reward per episode vs. training epochs. For both Breakout and Seaquest, it is difficult to tell whether Deep Mind is creating more valuable policies with increased training time. In the two right graphs, they more directly measure this by showing the average action value vs. training epochs. Here we can see that the learning algorithm makes more consistent progress on Seaquest.

3. (a) Value did not increase much when learning rate reduced beyond 0.01, but the training time or time for the value to stablilize increased substantially. While values were low with higher training rates, there was no divergence.

(b)

References

[1] S. Mahadevan. *Learning Representation and Control in Markov Decision Processes.* Now Publishers Inc, 2009.