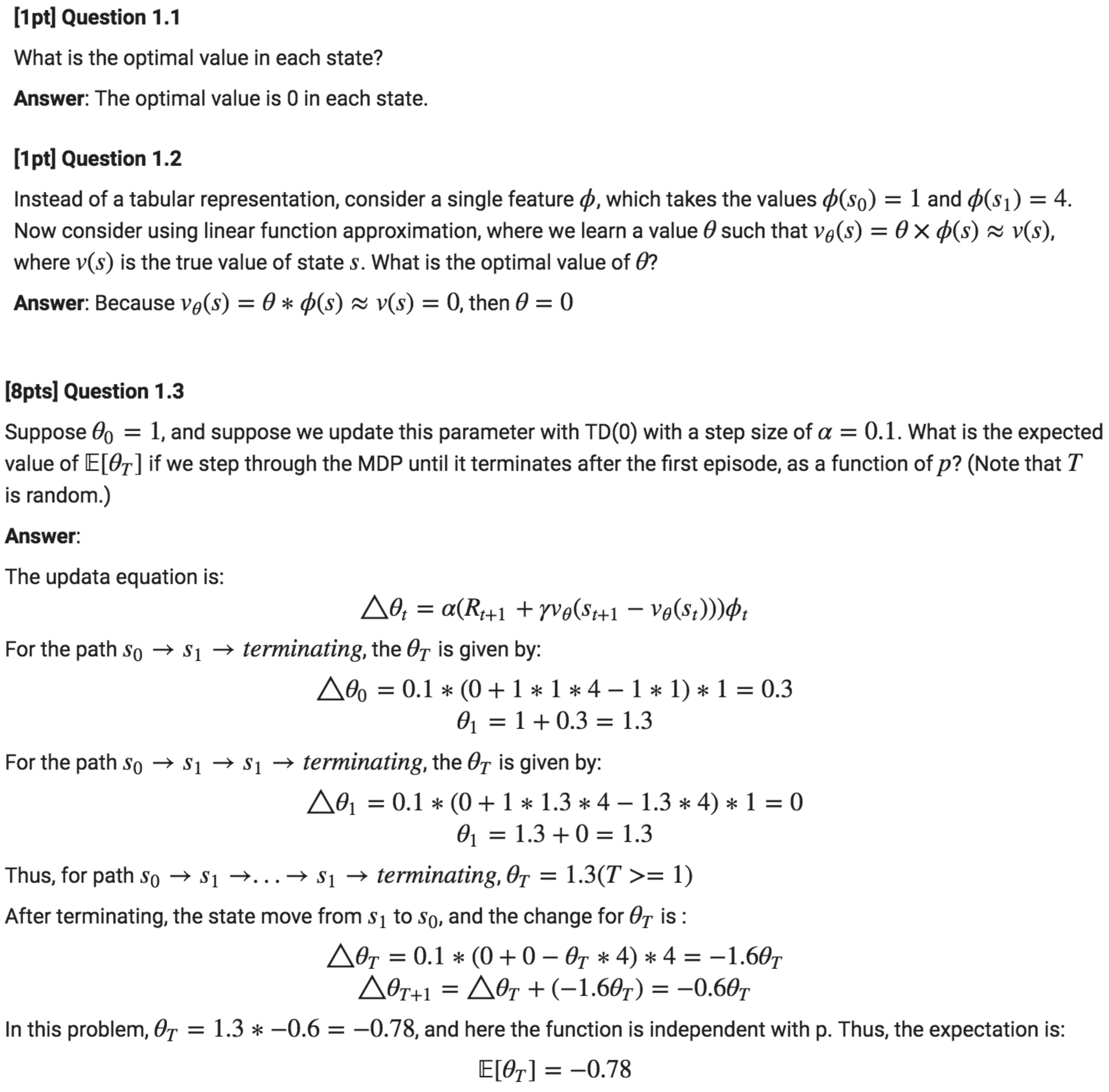
Reinforcement Learning Coursework 4

WU Zhi 17040772 UCABWUA@UCL.AC.UK

Department of Computer Science

University College London

Assignment 1



[1pt] Question 1.1

What is the optimal value in each state?

Answer: The optimal value is 0 in each state.

[1pt] Question 1.2

Instead of a tabular representation, consider a single feature ϕ, which takes the values $\phi(s_0) = 1$ and $\phi(s_1) = 4$. Now consider using linear function approximation, where we learn a value θ such that $v_\theta(s) = \theta \times \phi(s) \approx v(s)$, where $v(s)$ is the true value of state s. What is the optimal value of θ?

Answer: Because $v_\theta(s) = \theta * \phi(s) \approx v(s) = 0$, then $\theta = 0$

[8pts] Question 1.3

Suppose $\theta_0 = 1$, and suppose we update this parameter with TD(0) with a step size of $\alpha = 0.1$. What is the expected value of $\mathbb{E}[\theta_T]$ if we step through the MDP until it terminates after the first episode, as a function of p? (Note that T is random.)

Answer:

The updata equation is:

$$\triangle\theta_t = \alpha(R_{t+1} + \gamma v_\theta(s_{t+1} - v_\theta(s_t)))\phi_t$$

For the path $s_0 \rightarrow s_1 \rightarrow terminating$, the θ_T is given by:

$$\begin{aligned} \triangle\theta_0 &= 0.1 * (0 + 1 * 1 * 4 - 1 * 1) * 1 = 0.3 \\ \theta_1 &= 1 + 0.3 = 1.3 \end{aligned}$$

For the path $s_0 \rightarrow s_1 \rightarrow s_1 \rightarrow terminating$, the θ_T is given by:

$$\begin{aligned} \triangle\theta_1 &= 0.1 * (0 + 1 * 1.3 * 4 - 1.3 * 4) * 1 = 0 \\ \theta_1 &= 1.3 + 0 = 1.3 \end{aligned}$$

Thus, for path $s_0 \rightarrow s_1 \rightarrow \ldots \rightarrow s_1 \rightarrow terminating, \theta_T = 1.3(T >= 1)$

After terminating, the state move from s_1 to s_0, and the change for θ_T is :

$$\begin{aligned} \triangle\theta_T = 0.1 * (0 + 0 - \theta_T * 4) * 4 = -1.6\theta_T \\ \triangle\theta_{T+1} = \triangle\theta_T + (-1.6\theta_T) = -0.6\theta_T \end{aligned}$$

In this problem, $\theta_T = 1.3 * -0.6 = -0.78$, and here the function is independent with p. Thus, the expectation is:

$$\mathbb{E}[\theta_T] = -0.78$$

[5pts] Question 1.3

If $p = 0.1$, how many episodes does it take, starting from $\theta_0 = 1$, until $|v(s) - \mathbb{E}[v_\theta(s)]| < 0.5$ for all s, where the expectation is over the expected updates to θ?

Answer:

For each episode, the expectation of θ is -0.78: $\mathbb{E}[\theta_{T+1}] = -0.78 * \mathbb{E}[\theta_T]$

For the state s_1, we need make sure that:

$$|v(s_1) - \mathbb{E}[v_\theta(s_1)]| < 0.5$$
$$|0 - 4 * (-0.78)^n| < 0.5$$

Then for an integer n, we can get the answer: $n \geq 9$

For the state s_0, we need make sure that:

$$|v(s_0) - \mathbb{E}[v_\theta(s_0)]| < 0.5$$
$$|0 - 1 * (-0.78)^n| < 0.5$$

So, we can get the answer: $n \geq 3$

Overall, after at least **9** episodes $|v(s) - \mathbb{E}[v_\theta(s)]| < 0.5$.

[10pts] Question 1.4

What is the value of $\mathbb{E}[\theta_n]$, as a function of n and p?

Answer: The expectation of θ_{n+1} is:

$$\mathbb{E}[\theta_{n+1}] = \mathbb{E}[\theta_n] + \alpha\phi(s_0)\mathbb{E}[\delta_0] + \alpha\phi(s_1)\mathbb{E}[\delta_1]$$
$$where\delta = r + \gamma v_\theta(s_{t+1}) - v_\theta(s_t)$$

Then we can calaulate expectation of δ_0 and δ_1:

$$\mathbb{E}[\delta_0] = (\phi(s_1) - \phi(s_0))\mathbb{E}[\theta_n] = 3\mathbb{E}[\theta_n]$$
$$\mathbb{E}[\delta_1] = p * (0 - \phi(s_1)\mathbb{E}[\theta_n]) + (1-p) * 0 = -4p\mathbb{E}[\theta_n]$$

Thus the update equation is:

$$\mathbb{E}[\theta_{n+1}] = \mathbb{E}[\theta_n] + 3\mathbb{E}[\theta_n] + 4 * (-4p\mathbb{E}[\theta_n]) = (1 + 3\alpha - 16p\alpha)\mathbb{E}[\theta_n]$$

Assuming start from θ_0, the value of $\mathbb{E}[\theta_n]$ can be written as:

$$\mathbb{E}[\theta_n] = (1 + 3\alpha - 16\alpha p)^n\theta_0$$

[5pts] Question 1.5

For which values of p does not θ converge to the optimal solution?

Answer: The value of $\mathbb{E}[\theta_n]$ can be written as:

$$\mathbb{E}[\theta_n] = (1 + 3\alpha - 16\alpha p)^n\theta_0$$

If θ converge to the optimal solution which is 0, the requirement is:

$$(1 + 3\alpha - 16\alpha p)^n \simeq 0$$

Thus, for θ does not converge to 0, the requirement is

$$|1 + 3\alpha - 16\alpha p| \geq 1$$

The range of value of p is:

$$\frac{2+3\alpha}{16\alpha} \leq p \leq \frac{3}{16}$$

Hence, $\alpha = 0.1$ here, so $\frac{2+3\alpha}{16\alpha} = \frac{23}{16}$, so the final answer is

$$p \leq \frac{3}{16}$$

[10pts] Question 1.5

Why doesn't it? TD is known to converge, with linear function approximation, under certain assumptions. Explain for this concrete case why the algorithm does not converge, and explain which general assumption is violated that would otherwise ensure convergence of linear TD (in at most 200 words).

Answer:

$$\theta_{n+1} = \begin{cases} \theta_n + \frac{1}{10}(3\theta_n - 16\theta_n) = \theta_n - 1.3\theta_n & \text{with } p \\ \theta_n + \frac{1}{10}(3\theta_n) = \theta_n + 0.3\theta_n & \text{with } 1-p \end{cases}$$

So

$$\mathbb{E}[\theta_{n+1} \mid \theta_n] = \theta_n - A\theta_n = (1-A)\theta_n$$
$$where A = 1.3p - 0.3(1-p) = -0.3 + 1.6p$$

In general linear TD(0) algorithm, the convergence is guaranteed if $\mathbf{A}$ (with shape n by n, where n is the dimension of the feature space, in our case $n = 1$ and we have merge the step size α into A above) is positive definite so that in $(\mathbf{I} - \alpha\mathbf{A})\theta_n$, θ_n will be reduced toward zero and stability is assured.

In the above case, the convergence is guaranteed if $A > 0$ (actually we would also require $A < 2$ but since $p \le 1$ we can ignore this), that is $p > \frac{3}{16}$. So when $p \le \frac{3}{16}$, the assumption of the positive definiteness of A will be violated, resulting in divergence.

Answer:

Convergence can be obtained by setting different step sizes the two samples in each step, now

$$\theta_{n+1} = \theta_n + \alpha_0\delta_0\phi(s_0) + \alpha_1\delta_1\phi(s_1)$$

As before, $A = 16\alpha_1 p - 3\alpha_0$. Solve $0 < A < 2$ we have

$$\frac{3\alpha_0}{16\alpha_1} < p < \frac{2 + 3\alpha_0}{16\alpha_1}$$

In order to have this inequality holds for all $p \in (0, 1]$, we can set α_0 and α_1 as finctions of n such that

$$\lim_{n\to\infty} \frac{3\alpha_0(n)}{16\alpha_1(n)} = 0$$

and

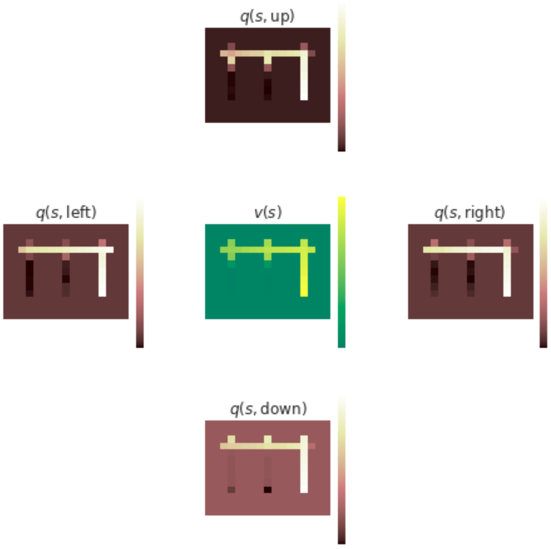
$$\lim_{n\to\infty} \frac{2 + 3\alpha_0(n)}{16\alpha_1(n)} \ge 1$$

We can set $\alpha_0(n)$ as a non-negative decreasing function of n and $\alpha_1(n)$ as an increasing function of n with supremum $\frac{2}{16}$, such that the two limits above will hold. Then the convergence is guaranteed whenever $p > 0$. Note that we cannot reach convergence if $p = 0$ as θ_n will go to ∞ as n goes to ∞.

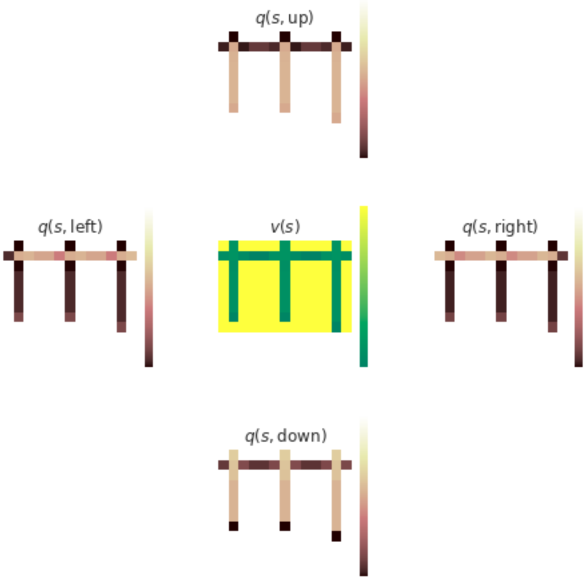
Assignment 2

Plots

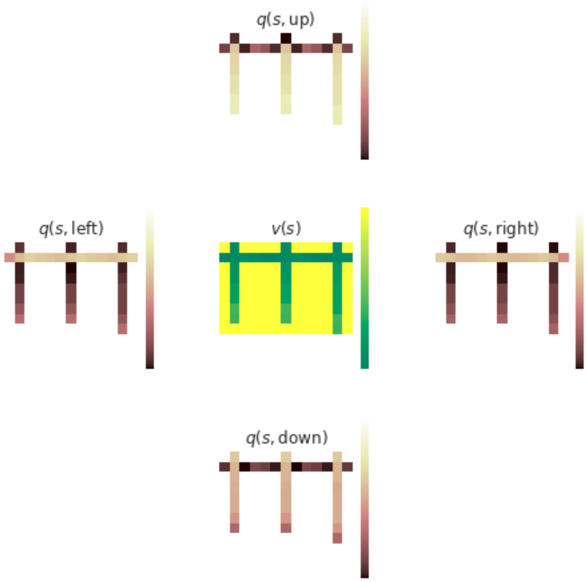
SARSA



Vision size of 1



Vision size of 2



Questions

Consider the greedy policy with respect to the estimated values

[5 pts] Which algorithm performed best? Why?

Answer:

Tabular Sarsa performed the best since the environment is deterministic and the tabular can update the q values with greater precision compared to that using of neural network no cunts leave here till we find out what cunts did it.

[5 pts] Is there a difference in the solution found by Neural Sarsa with a vision size of 1 (so 3x3 local observations), and a vision size of 2 (so 5x5 local observations)? Why?

Answer:

There is no obvious difference between the two plots above.

[10 pts] How could we improve the performance of the Neural Sarsa agent on this domain (for both vision sizes)? Identify the main issue, and propose a concrete solution (in max 200 words).

[10 BONUS pts] Implement your proposed improvement and show that it actually helps performance.