1. I read the indicated chapters from the book.

2. Part a) I chose the Adam optimizer as the optimization solver.
I defined the loss as follows:

$$loss = \left(V^*_{new}(H) - V^*_{old}(H)\right)^2 + \left(V^*_{new}(L) - V^*_{old}(L)\right)^2$$

In this way I found

$$V^*(H) = 15.748 \qquad V^*(L) = 14.17.32$$

For this

optimal Policy

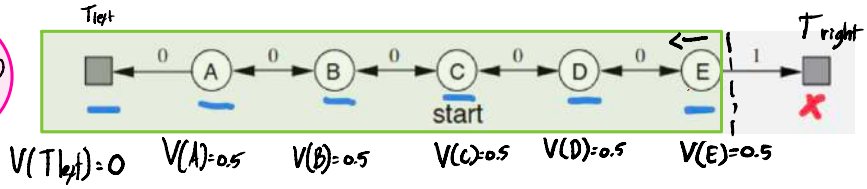
state: High => Action: Search

state: Low => Action: Recharge

Part b) The result for value iteration is the same.

More details are explained in the notebook file.

3. 6.3



Q1 Given that only the value of state A was changed we can infer that during the first episode, all states except the rigth terminal state could have been visited and it is trivial to see that the episode ended in the left terminal state. Thus, we can conclude only zeros rewards were received in the first episode.

Q2. A was the only state that had a change in value for the following two reasons

① The next states of A are T_{left} and B. There is equal probability to go either of them. But the key point is that if we end up a T_{left} we get

$$V(A) = V(A) + \alpha [r_{t+1} + \gamma V(T_{left}) - V(A)]$$

where $V(A) = 0.5$ $V(T_{left}) = 0$ $\alpha = 0.1$ $r_{t+1} = 0$ $\gamma = 1$

$$V(A) = 0.5 + 0.1 [0 + 0 - 0.5] = 0.45$$

② A is the only state of the possibly visited states that can go to T_{left}. The rest of the states have

next state which have a value of 0.5 and the reward would be zero. This combination results in $V(S_t) = V(S_t)$.

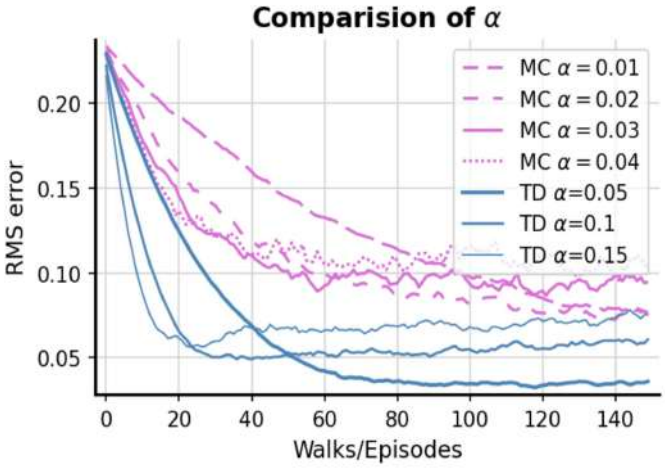
Regarding state E, it could have been visited in episode one, if and only if the next state for E was D.

Q_3. As we saw it changed by 0.05

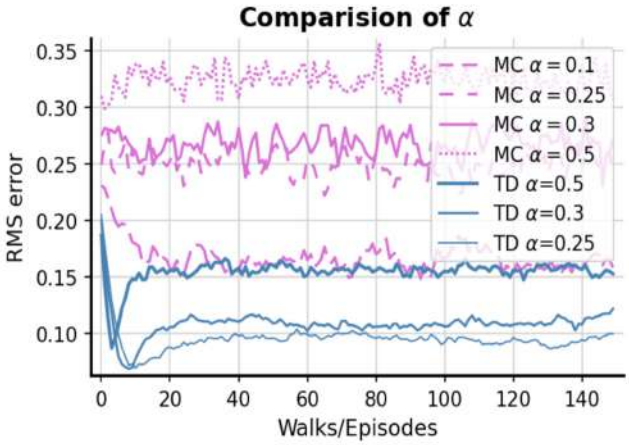
6.4 The conclusion does not change even if we use a wider range of values for α.

For this, we can plot the results of TD(0) and constant-α Mc as follows:

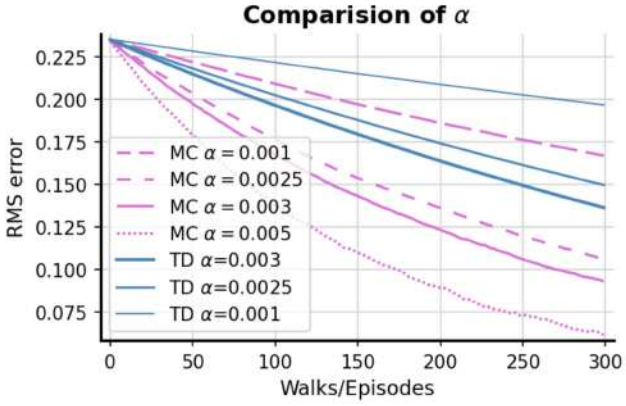
a) Same plot as the book



b) bigger α values



c) very small α values



We can see that the conclussion in the question does not change.

In plot a and b we can see that TD(0) reaches a lower RMS error than constant-α Mc, particularly lower as the value of α becomes smaller. On the other hand when the value of α is very small (plot c) the Constant-α Mc method performs better than TD(0) for same α values, however, it takes 250 episodes to reach the 0.075 error while TD(0) can reach 0.05 error in only approximately 50 episodes when $\alpha = 0.05$. Thus we can still conclude TD(0) is a better choice in this case.

One last important thing we need to notice is that when α is bigger the introduced noise causes both TD(0) and constant-α Mc to perform worse, particularly the values stuck (plot b) in a "wave" which is worse for Constant-α Mc because it forgets old episodes and the bigger α makes it change the value of visited states too much.

TD(0) is better in this case because, as we can conclude, it is not as affected by the noise as constant-α Mc which allows it to converge faster.

6.6 I think we can use either Policy iteration or Value iteration.

We can use them because we know $\pi(a|s) = 0.5$ for all actions and states and we can get the transition probabilities easily from the given MDP. This means that we can use Dynamic Programming.

I think the one used by the authors of the book was policy iteration. This is because the state space is pretty small, so the computational cost drawback of policy iteration that makes people use Value iteration is not something to worry.

4. 10.6

Q1. Ergodicity is sufficient but not necessary to guarantee the existence of the limit in (10.6). Thus the average reward can be defined but the differential return (10.9) cannot be defined for A and B

Q2. We can alternatively define the differential value as

$$v_\pi(s) \doteq \lim_{\gamma \to 1} \lim_{h \to \infty} \sum_{t=0}^{h} \gamma^t \left(E_\pi [R_{t+1} | S_0 = s] - r(\pi) \right)$$

To find the differential value for A and B we first employ (10.6) to find $r(\pi)$

$$r(\pi) \doteq \lim_{h \to \infty} \frac{1}{h} \sum_{t=1}^{h} E[R_t | S_0, A_{0:t-1}^{\pi}] \qquad (10.6)$$

here $E[R_t | S_0 A_{0:t-1}^{\pi}] = 0.5$ because it alternates between 1 and 0 infinitely

$$r(\pi) \doteq \lim_{h \to \infty} \frac{1}{h} \sum_{t=1}^{h} 0.5 = 0.5$$

because for all t 0.5 is our Constant expected reward

Now that we have the average reward we can obtain $V_\pi(A)$ and $V_\pi(B)$ $\quad A \to +1, 0, +1 \ldots$

$$V_\pi(A) = \lim_{\gamma \to 1} \lim_{h \to \infty} (1 - 0.5) + \gamma(0 - 0.5) + \gamma^2(1 - 0.5) + \ldots$$

$$V_\pi(A) = \lim_{\gamma \to 1} \lim_{h \to \infty} 0.5 + \gamma(-0.5) + \gamma^2(0.5) + \ldots$$

$$V_\pi(A) = \lim_{\gamma \to 1} \lim_{h \to \infty} \sum_{t=0}^{h} \gamma^t \frac{(-1)^t}{2}$$

$$= \frac{1}{2} \lim_{\gamma \to 1} \lim_{h \to \infty} \sum_{t=0}^{h} (-\gamma)^t$$

here

$$\lim_{h \to \infty} \sum_{t=0}^{h} (-\gamma)^t$$

is a geometric series of the following form

$$\sum_{t=0}^{h} ar^t = a \frac{1 - r^{h+1}}{1 - r}$$

Then

$$\sum_{t=0}^{h} (-\gamma)^t = \frac{1 - (-\gamma)^{h+1}}{1 - (-\gamma)} = \frac{1 - (-\gamma)^{h+1}}{1 + \gamma}$$

$$\lim_{h\to\infty} \sum_{t=0}^{h} (-\gamma)^t = \lim_{h\to\infty} \left(\frac{1-(-\gamma)^{h+1}}{1+\gamma} \right)$$

Then if we take the limit, as h approaches to infinity the $|\gamma|<1$ for the series to converge because

$$\lim_{h\to\infty} (-\gamma)^{h+1} = 0 \quad \text{for } |\gamma|<1$$

then

$$\lim_{h\to\infty} \sum_{t=0}^{h} (-\gamma)^t = \frac{1}{1+\gamma}$$

Now if we get back to $V_*(A)$

$$V_{\pi}(A) = \frac{1}{2} \lim_{\gamma\to 1} \lim_{h\to\infty} \sum_{t=0}^{h} (-\gamma)^t$$

$$= \frac{1}{2} \lim_{\gamma\to 1} \frac{1}{1+\gamma}$$

$$= \frac{1}{4}$$

Now for $V_\pi(B)$ $\quad B = 0, +1, 0 \ldots$

$$V_\pi(B) = -\gamma \lim_{\gamma \to 1} \lim_{h \to \infty} \sum_{t=0}^{h} \gamma^t \frac{(-1)^t}{2}$$

$$V_\pi(B) = \lim_{\gamma \to 1} \lim_{h \to \infty} (0 - 0.5) + \gamma(1 - 0.5) + \gamma^2(0 - 0.5) + \ldots$$

$$V_\pi(B) = \lim_{\gamma \to 1} \lim_{h \to \infty} (-0.5) + \gamma(0.5) + \gamma^2(-0.5) + \ldots$$

$$V_\pi(B) = \lim_{\gamma \to 1} \lim_{h \to \infty} (-1)[(0.5) + \gamma(-0.5) + \gamma^2(0.5) + \ldots$$

$$V_\pi(B) = -\lim_{\gamma \to 1} \lim_{h \to \infty} \sum_{t=0}^{h} \gamma^t \frac{(-1)^t}{2}$$

$$V_\pi(B) = -\frac{1}{2} \lim_{\gamma \to 1} \lim_{h \to \infty} \sum_{t=0}^{h} (-\gamma)^t$$

$$= -V_\pi(A)$$

$$= -\frac{1}{4}$$

10.7 The average reward for this case would be $\frac{1}{3}$ because we only get reward upon arrival in A but we have three states

$$V_\pi(s) = \lim_{\gamma \to 1} \lim_{h \to \infty} \sum_{t=0}^{h=\infty} \gamma^t \left(E_\pi[R_{t+1} | S_0 = s] - r(\pi) \right)$$

$$= \lim_{\gamma \to 1} \gamma^0 (R - r(\pi) + \gamma (R_2 - r(\pi)) + \ldots + \gamma^\infty (R_{\infty+1} - r(\pi))$$

$$= \gamma^0 (R - r(\pi) + \gamma (R_2 - r(\pi)) + \ldots + \gamma^\infty (R_{\infty+1} - r(\pi))$$

where γ approches 1

$$= G_t$$

This is differential return

We also know that return can be expressed as

$$G_t = r_{t+1} + \gamma V(S_{t+1})$$

in this case $r_{t+1} = R_{t+1} - r(\pi)$

$$G_t = R_{t+1} - r(\pi) + \gamma V(S_{t+1})$$

Then we have the following

$$V(A) = R_{t+1} - r(\pi) + \gamma V(B)$$
$$V(B) = R_{t+1} - r(\pi) + \gamma V(C)$$
$$V(C) = R_{t+1} - r(\pi) + \gamma V(A)$$

Since we only get a reward when arriving to A, $r(\pi) = \frac{1}{3}$ and

$$V(A) = 0 - \frac{1}{3} + \gamma V(B)$$
$$V(B) = 0 - \frac{1}{3} + \gamma V(C)$$
$$V(C) = 1 - \frac{1}{3} + \gamma V(A)$$

Then we can obtain $V(A)$ as

$$V(A) = 0 - \frac{1}{3} + \gamma \left[0 - \frac{1}{3} + \gamma \left[1 - \frac{1}{3} + \gamma V(A) \right] \right]$$

$$V(A) = -\frac{1}{3} - \frac{1}{3}\gamma + \frac{2}{3}\gamma^2 + \gamma^3 V(A)$$

$$V(A) = \frac{1}{3}(-1 - \gamma + 2\gamma^2) + \gamma^3 V(A)$$

$$V(A) - \gamma^3 V(A) = \frac{1}{3}(-1 - \gamma + 2\gamma^2)$$

$$V(A) = \frac{1}{3} \frac{(-1 - \gamma + 2\gamma^2)}{1 - \gamma^3}$$

$$V(A) = \frac{1}{3}\frac{(-1-\gamma+2\gamma^2)}{1-\gamma^3}$$

$$V(A) = \frac{1}{3}\frac{(2\gamma+1)(\gamma-1)}{-(\gamma^2+\gamma+1)(\gamma-1)}$$

$$V(A) = -\frac{1}{3}\frac{2\gamma+1}{\gamma^2+\gamma+1}$$

Then as we know that γ approaches 1

$$V(A) = -\frac{1}{3}\left(\frac{3}{3}\right) = -\frac{1}{3}$$

We can then get $V(c)$

$$V(c) = 1 - \frac{1}{3} + \gamma V(A)$$

$$V(c) = \frac{2}{3} - \frac{1}{3} = \frac{1}{3}$$

and $V(B)$

$$V(B) = 0 - \frac{1}{3} + \gamma V(c)$$

$$V(B) = -\frac{1}{3} + \frac{1}{3} = 0$$

10.8 First if we use only $R_{t+1} - r(\pi)$ we get the following if we start from state A

$R_{A\to B} - r(\pi), R_{B\to C} - r(\pi), R_{C\to A} - r(\pi), R_{A\to B} - r(\pi), R_{B\to C} - r(\pi), \ldots$

$0 - \frac{1}{3}, 0 - \frac{1}{3}, 1 - \frac{1}{3}, 0 - \frac{1}{3}, 0 - \frac{1}{3}, \ldots$

$-\frac{1}{3}, -\frac{1}{3}, \frac{2}{3}, -\frac{1}{3}, -\frac{1}{3}, \frac{2}{3}, -\frac{1}{3}, -\frac{1}{3}, \frac{2}{3} \ldots$

But if we use the differential form of TD error

$$\delta_t = R_{t+1} - r(\pi) + V(S_{t+1}, W_t) - V(S_t, W_t)$$

recall $V(A) = -\frac{1}{3}$, $V(B) = 0$ and $V(C) = \frac{1}{3}$

then

$-\frac{1}{3} + V(B) - V(A), -\frac{1}{3} + V(C) - V(B), \frac{2}{3} + V(A) - V(C), \ldots$

$-\frac{1}{3} + 0 - \left(-\frac{1}{3}\right), -\frac{1}{3} + \frac{1}{3} - 0, \frac{2}{3} + -\frac{1}{3} - \frac{1}{3}$

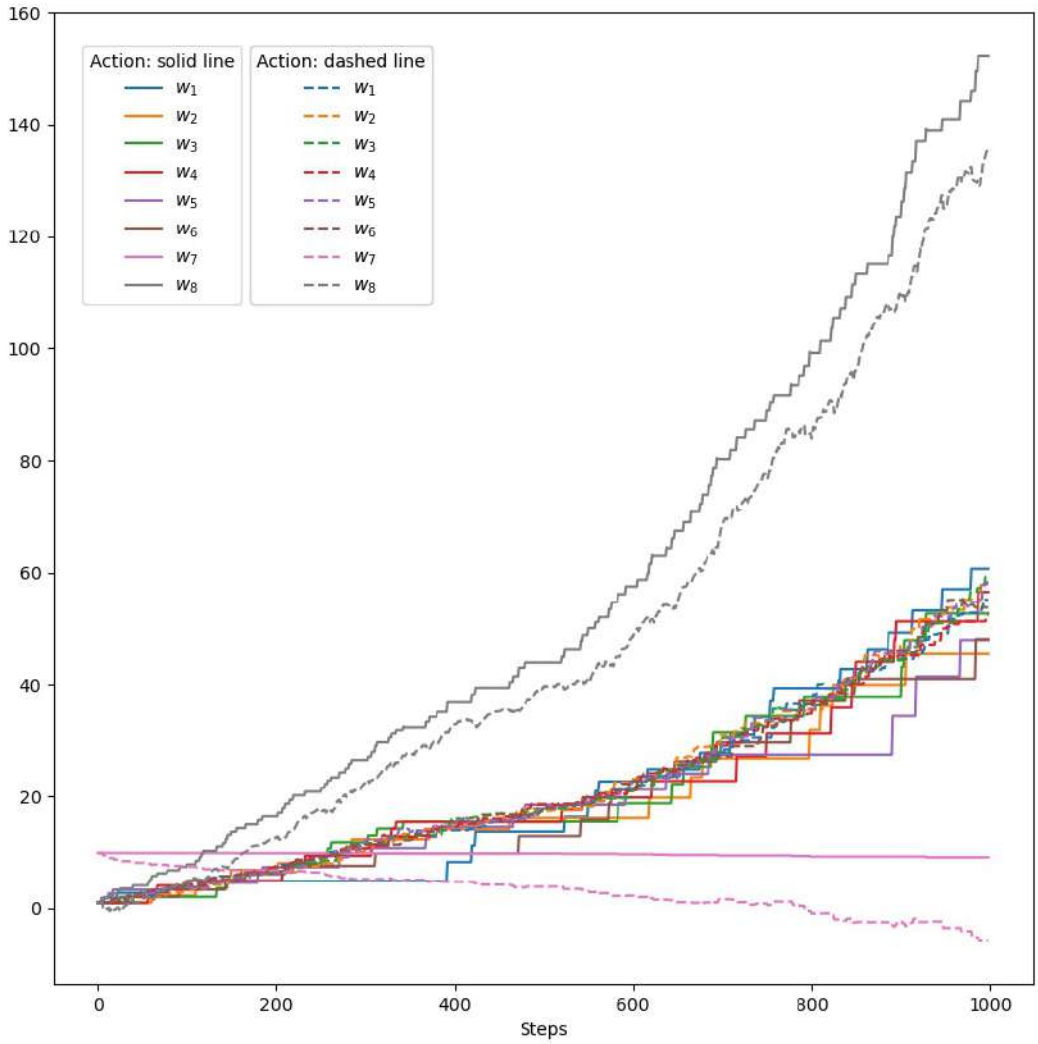
$0, 0, 0 \ldots = 0, 0, 0, 0, 0, 0, 0, 0, 0, \ldots$

Continued →

the differential form of TD error is the one that has more stable updates because it uses TD to reflect on the speed of change required. As we can see we obtained zeroes in the computation of errors because we already have the true value function. In other words the size of change decreases as we approach the true value function.

5. 11.3 The code is in the notebook file.

In the below figure from the code we can confirm that the weights diverge.



6. (13.4) First we obtain $\ln \pi(a|s,\theta)$

Part a) $$\pi(a|s,\theta) \doteq \frac{1}{\sigma(s,\theta)\sqrt{2\pi}} \exp\left(-\frac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2}\right)$$

$$\ln \pi(a|s,\theta) \doteq \ln\left(\frac{1}{\sigma(s,\theta)\sqrt{2\pi}}\right) + \ln \exp\left(-\frac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2}\right)$$

$$\ln \pi(a|s,\theta) \doteq -\frac{1}{2}\ln(2\pi\sigma(s,\theta)^2) - \left(\frac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2}\right)$$

Now we obtain the first part of the elegibility vector

$$\nabla \ln \pi(a|s,\theta_\mu) = \left(-\frac{1}{2}\ln(2\pi\sigma(s,\theta)^2) - \left(\frac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2}\right)\right)'$$

$$\nabla \ln \pi(a|s,\theta_\mu) = 0 + \frac{1}{2\sigma(s,\theta)^2} - \left((a-\mu(s,\theta))^2\right)'$$

$$= \frac{-2}{2\sigma(s,\theta)^2}(a-\mu(s,\theta))(-x_\mu(s))$$

$$= \frac{1}{\sigma(s,\theta)^2}(a-\mu(s,\theta))\, x_\mu(s)$$

Part b)

Then we obtain the second part of the elegibility vector

$$\nabla \ln \pi(a|s,\theta_\sigma) \doteq \left(-\frac{1}{2} \ln(2\pi\sigma(s,\theta)^2) - \left(\frac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2} \right) \right)'$$

first term of derivative

$$\left(-\frac{1}{2} \ln(2\pi\sigma(s,\theta)^2) \right)'$$

$$= \left(-\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln \sigma(s,\theta)^2 \right)'$$

$$= 0 - \frac{1}{2} \frac{1}{\sigma(s,\theta)^2} (\sigma(s,\theta)^2)'$$

$$= 0 - \frac{1}{2} \frac{1}{\sigma(s,\theta)^2} 2\sigma(s,\theta)(\sigma(s,\theta))'$$

$$= -\frac{1}{\sigma(s,\theta)} (\sigma(s,\theta))'$$

Continued →

Second term for derivative

$$\left(-\left(\frac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2}\right)\right)'$$

$$-\frac{(a-\mu(s,\theta))^2}{2}\left(\frac{1}{\sigma(s,\theta)^2}\right)'$$

$$-\frac{(a-\mu(s,\theta))^2}{2}\left(\sigma(s,\theta)^{-2}\right)'$$

$$-\frac{(a-\mu(s,\theta))^2}{2}\left(-2\,\sigma(s,\theta)^{-3}\right)\left(\sigma(s,\theta)\right)'$$

$$\frac{(a-\mu(s,\theta))^2}{\sigma(s,\theta)^3}\left(\sigma(s,\theta)\right)'$$

Then the derivate with both terms

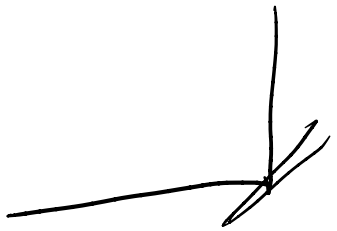
$$-\frac{1}{\sigma(s,\theta)}\left(\sigma(s,\theta)\right)' + \frac{(a-\mu(s,\theta))^2}{\sigma(s,\theta)^3}\left(\sigma(s,\theta)\right)'$$

Continued →

$$= (\delta(s,\theta))' \left(\frac{(a-\mu(s,\theta))^2}{\delta(s,\theta)^3} - \frac{1}{\delta(s,\theta)} \right)$$

$$= \delta(s,\theta) X_\delta(s) \left(\frac{(a-\mu(s,\theta))^2}{\delta(s,\theta)^2 \delta(s,\theta)} - \frac{1}{\delta(s,\theta)} \right)$$

$$= \left(\frac{(a-\mu(s,\theta))^2}{\delta(s,\theta)^2} - 1 \right) X_\delta(s)$$



13.5 Part a $\pi(a|s,\theta) \doteq \dfrac{e^{h(s,a,\theta)}}{\sum_b e^{(s,b,\theta)}}$

Then

$$P_t = \pi(1|s,\theta) = \frac{e^{h(s,1,\theta)}}{e^{h(s,1,\theta)} + e^{h(s,0,\theta)}}$$

$$P_t = \pi(1 \mid s, \theta) = \frac{e^{h(s,1,\theta)}}{e^{h(s,1,\theta)}\left(1 + \frac{e^{h(s,0,\theta)}}{e^{h(s,1,\theta}}\right)}$$

$$= \frac{1}{1 + \frac{e^{h(s,0,\theta)}}{e^{h(s,1,\theta}}} = \frac{1}{1 + e^{h(s,0,\theta) - h(s,1,\theta)}}$$

$$= \frac{1}{1 + e^{h(s,0,\theta) - h(s,1,\theta)}} = \frac{1}{1 + e^{-(h(s,1,\theta) - h(s,0,\theta))}}$$

We know

$$h(s,1,\theta) - h(s,0,\theta) = \theta^T X(s)$$

Then

$$\frac{1}{1 + e^{-\theta_t^T X(s)}}$$

Part b) From the algorithm in page 328 we know that

$$\theta_{t+1} = \theta_t + \alpha \gamma^t G \nabla \ln \pi(A_t | S_t, \theta)$$

Now for the Bernoulli logistic unit, we need to bring here the results we obtain later in part c, since we know from it that

$$\nabla \ln \pi(a|s,\theta) = (a - P) X(s)$$

Then

$$\theta_{t+1} = \theta_t + \alpha \gamma^t G_t (a - P) X(s)$$

$$\theta_{t+1} = \theta_t + \alpha \gamma^t G_t \left(a - \frac{1}{1 + e^{-\theta^T X(s)}} \right) X(s)$$

Part c) $p = \pi(1|s,\theta) = \frac{1}{1+e^{-\theta^T x(s}}$

and we know that $\pi(0|s,\theta) = 1 - p$

First, let's take the derivative $\ln \pi(1|s,\theta)$

$$\nabla \ln \pi(1|s,\theta) = (\ln p)' = \frac{1}{p} \nabla p$$

In part a we showed that

$$p = \sigma(\theta^T x(s))$$

And we know that

$$\nabla \sigma(x) = \left(\frac{1}{1+e^{-x}}\right)' = \sigma(x)(1-\sigma(x))(x)'$$

Then

$$\nabla p = p(1-p)(\theta^T x(s))'$$

$$\nabla p = p(1-p)\, x(s)$$

In this way

$$\nabla \ln \pi(1|s,\theta) = \frac{1}{p} \nabla p = \frac{1}{p} p(1-p) x(s)$$

$$\nabla \ln \pi(1|s,\theta) = (1-p) x(s)$$

We need to also find

$$\nabla \ln \pi(0|s,\theta) = \nabla \ln(1-p)$$

$$= \frac{1}{1-p}(1-p)' = \frac{-1}{1-p} \nabla p$$

$$= \frac{-1}{1-p} p(1-p) x(s) = -p x(s)$$

Then we need to combine them in an expression including a.

For this we can notice

$$\nabla \ln \pi(1|s,\theta) = (1-p) x(s) = (a-p) x(s)$$

continued →

If we use

$$\nabla \ln \pi(a|s,\theta) = (a-p)\, x(s)$$

In

$$\nabla \ln \pi(0|s,\theta) = (0-p)\, x(s)$$

$$= -p\, x(s)$$

Then we can conclude that

$$\nabla \ln \pi(a|s,\theta) = (a-p)\, x(s)$$