Q1

A) we prove $U(V)$ is a contraction mapping under the real metric space with $\|.\|_\infty$ as the ditance metric:

let V_1 and V_2 be two value vetors, hence:

$$d(U(V_1), U(V_2)) = \|U(V_1) - U(V_2)\|_\infty$$

$$= \|R + \gamma P V_1 - R - \gamma P V_2\|_\infty$$

$$= \gamma \|P(V_1 - V_2)\|_\infty$$

$$\leq \gamma \|(V_1 - V_2)\|_\infty$$

$$= \gamma d(V_1, V_2)$$

since each entry of $P(V_1 - V_2)$ is a convex combination of $V_1 - V_2$, and hene no more than $\max(V_1 - V_2)$

assuming $\gamma < 1$, we are done.

B) By definition we have $U(V^\pi) = R + \gamma P V^\pi = V^\pi$. Thus we can say

$$\|U^n(V) - V^\pi\|_\infty = \|U^n(V) - U(V^\pi)\|_\infty$$

$$\leq \gamma \|U^{n-1}(V) - V^\pi\|_\infty$$

$$\vdots$$

$$\leq \gamma^n \|V - V^\pi\|_\infty$$

given an initial V, $\|V - V^\pi\|_\infty = C$

$$\Rightarrow \lim_{n\to\infty} \|U^n(V) - V^\pi\|_\infty \leq \lim_{n\to\infty} \gamma^n C = 0 \Longrightarrow \lim_{n\to\infty} U^n(V) = V^\pi$$

c)

$$\|v^{\pi} - U^{k}(v)\|_\infty = \|v^{\pi} - U^{k+1}(v) + U^{k+1}(v) - U^{k}(v)\|_\infty$$

$$\leq \|v^{\pi} - U^{k+1}(v)\|_\infty + \|U^{k+1}(v) - U^{k}(v)\|_\infty$$

$$\leqslant \gamma \|v^{\pi} - U^{k}(v)\|_\infty + \gamma \|U^{k}(v) - U^{k-1}(v)\|_\infty$$

$$\Longleftrightarrow (1-\gamma)\|v^{\pi} - U^{k}(v)\|_\infty \leqslant \varepsilon$$

$$\overset{\gamma \neq 1}{\Longleftrightarrow} \|v^{\pi} - U^{k}(v)\|_\infty \leqslant \frac{\varepsilon}{1-\gamma}$$

Q2

A) Starting from the last step of the episode, we assign the values. since 21 and 22 are the only states visited more than once. we average the score for them.

$V_{23} = 10$, $V_{18} = 10$, $V_{17} = 10$, $V_{22} = \frac{10+0}{2} = 5$, $V_{21} = \frac{0-10}{2} = -5$

$V_{20} = V_{16} = V_{12} = V_7 = V_8 = V_3 = V_2 = V_1 = -10$

B) since 21 and 22 are the only states visited more than once, these are the only states which chang in First-Visit-Monte-Carlo. We have

$V_{21} = 0$, $V_{22} = -10$

Q3

A) By definition, $V^{\pi}(s) = E^{\pi}[\sum_{t=0}^{\infty} \gamma^t r_t \mid S_0 = s]$. This expressin only depends on the probability distribution of the policy, and has nothing to do with the initial states distribution, as we condition over S_0. The V^{π} is the same for both M and M_0.

B) Correct. let b be a lower and upper bound for the absolute reward, i.e.

$$\forall\, s, a \,;\, |r(s,a)| \leq b$$

and $0<\alpha$ be a constant which all the the reward values are multiplied by. and $r'(s,a) = \alpha r(s,a)$. The new value function under the policy π becomes:

$$V'^{\pi}(s) = E^{\pi}[\sum_{t=0}^{\infty} \gamma^t r'_t \mid S_0 = s] = E^{\pi}[\sum_{t=0}^{\infty} \gamma^t \alpha r_t \mid S_0 = s] = \alpha V^{\pi}(s).$$

The optimal policy maximizes the value function for all states:

$$\pi^* = \arg\max_{\pi} V^{\pi}(s)$$

since rewards, and hence the value function are bounded (since $\gamma<1$), therefore

$$\pi'^* = \arg\max_{\pi} V'^{\pi}(s) = \arg\max_{\pi} \alpha V^{\pi}(s) = \arg\max_{\pi} V^{\pi}(s) = \pi^*.$$

C) Let us assume that we have terminating states and

$$V^{\pi}(s_1) = E^{\pi}[\sum_{t=0}^{L_1} \gamma^t r_t \mid S_0 = s_1]$$

(The problem is episodic)

$$V^{\pi}(s_2) = E^{\pi}[\sum_{t=0}^{L_2} \gamma^t r_t \mid S_0 = s_2]$$

and $V^{\pi}(s_1) = V^{\pi}(s_2)$ but $L_1 \neq L_2$. therefore the effect of adding a constant c varies based on the episode lengths. i.e.

$$V_c^{\pi}(s_1) = E^{\pi}\Big[\sum_{t=0}^{L_1}(\gamma^t r_t + \gamma^t c) \Big| S_0 = s_1\Big] = V^{\pi}(s_1) + c\sum_{t=1}^{L_1}\gamma^t$$

$$V_c^{\pi}(s_2) = E^{\pi}\Big[\sum_{t=0}^{L_2}(\gamma^t r_t + \gamma^t c) \Big| S_0 = s_2\Big] = V^{\pi}(s_2) + c\sum_{t=1}^{L_2}\gamma^t$$

So we can have $V_c^{\pi}(s_1) \neq V_c^{\pi}(s_2)$. Therfore the optimal policy can change.

D) If there are no terminating states, unlike part C, we cannot have L_1 and L_2 with different lengths, since for all states:

$$V^{\pi}(s) = E^{\pi}\Big[\sum_{t=0}^{\infty}\gamma^t r_t \Big| S_0 = s\Big]$$

In other words, the value functions for all states is add by a constant $c\sum_{t=0}^{\infty}\gamma^t$, thus the optimal policy doesn't change.

E) let π_1^* be the optimal policy for the MDP. By the convergence of policy iteration, this policy is the solution of bellman equation, hence

$$V_1^{\pi_1^*}(s) = \sum_{s'} P(s'|\pi_1^*(s), s)\left[r_1(s, \pi_1^*(s)) + \gamma V_1^{\pi_1^*}(s)\right]$$

No consider the new MDP. Since π_1^* is optimal, we have

$$r_2(s, \pi_1^*(s)) = r_1(s, \pi_1^*(s)) \Longrightarrow$$

$$V_2^{\pi_1^*}(s) = \sum_{s'} P(s'|\pi_1^*(s), s)\left[r_2(s, \pi_1^*(s)) + \gamma V_2^{\pi_1^*}(s)\right] = V_1^{\pi_1^*}(s)$$

So π_1^* also is correct for the new MDP, since $V_2^{\pi_1^*}(s) = V_1^{\pi_1^*}(s)$.

Now let π_2^* be the optimal policy for the new MDP and $\pi_1^* \neq \pi_2^*$.

Therefore $r_2(s, \pi_2^*(s)) = r_2(s, \pi_1^*(s)) - c$ and

$$V_2^{\pi_2^*}(s) = E^{\pi_2^*}\left[\sum_{t=0}^{\infty} \gamma^t (r_t' - c) \mid s_0 = s\right] = V_1^{\pi_2^*}(s) - c\sum_{t=0}^{\infty} \gamma^t < V_1^{\pi_2^*}(s) \leq V_1^{\pi_1^*}(s)$$

$$= V_2^{\pi_1^*}(s)$$

which is a contradiction. so $\pi_1^* = \pi_2^*$.

Q4

A) The standard Bellman equation is:

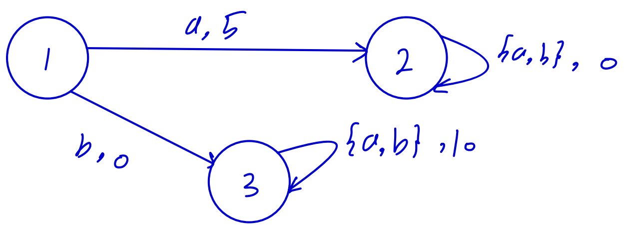
$$v^{\pi}(s) = \sum_a \pi(s,a) \sum_{s'} P(s',a,s)[R(s,a) + \gamma V^{\pi}(s')]$$

Now Let's consider the hypothetical "reversed" Bellman equation the attemts to determine the value of a state based on the future states:

$$v^{\pi}(s) = \sum_{s'} \sum_{a'} P(s',a,s) \left[\frac{v^{\pi}(s') - R(s,a)}{\gamma} \right]$$

The problem with the above method is that it is not consistent with the opimality principle, as the sequence derived from this operator is not nessesarily increasing.

Now let's see a counterexample. Consider this MDP:



$S = \{1, 2, 3\}$
$A = \{a, b\}$

Let $\gamma = 0.5$ and states 2 and 3 are terminal. First let's use the standard bellman equation. Since 2 and 3 are both termial states, $V^{\pi}(2) = 0$, $V^{\pi}(3) = 10$ regardles of the action

And $V^{\pi}(1) = P(2,a,1)[\underbrace{R(1,a) + \gamma V^{\pi}(2)}_{5 + 0.5\times 0 = 5}] + P(3,b,1)[\underbrace{R(1,b) + \gamma V^{\pi}(3)}_{0 + 0.5\times 10 = 5}] = 5$

But by the reversed Bellman equation we get:

$V^{\pi}(2) = \frac{V^{\pi}(2) - 0}{0.5} \Rightarrow V^{\pi}(2) = 0$, $V^{\pi}(3) = \frac{V^{\pi}(3) - 10}{0.5} \Rightarrow V^{\pi}(3) = -10$

$V^{\pi}(1) = P(2,a,1) \times \frac{V^{\pi}(2) - 5}{0.5} + P(3,a,1) \times \frac{V^{\pi}(3) - 0}{0.5} = -10$

But it is clearly incorrect. Since the state values of an MDP with all positive rewards cannot be negative.

B) Consider a Markov decision process. Since the value of each state contians the value of previous states, we have Markov property. Because given, $s_1, s_2, \dots, s_{t-1}, s_t, s_{t+1}$, s_t contains $s_1, \dots, s_{t-1}$ and s_{t+1} is base on $s_1, \dots, s_{t-1}, s_t$. so give s_t, s_{t+1} has all the information it needs. Thus if $s_t = s$, we can leave out the previous rewards and have: $v^{\pi}(s) = E[G_t \mid S_t = s, \pi] = E[G \mid S_0 = s, \pi]$.

Alternatively we can say

$$V^{\pi}(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R_t \mid S_0 = s\right]$$

on the other hand

$$E[G_t \mid S_t = s, \pi] = E\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \mid S_t = s, \pi\right]$$

$$= \sum_a \pi(s,a)\left[R(s,a) + E\left[\sum_{k=1}^{\infty} \gamma^k R_{t+k} \mid S_t = s, \pi\right]\right]$$

$$= \sum_a \pi(s,a) R(s,a) + \sum_a \pi(s,a) E\left[\sum_{k=1}^{\infty} \gamma^k R_{t+k} \mid S_t = s, \pi\right]$$

$$= \sum_a \pi(s,a) R(s,a) + \sum_a \pi(s,a) \sum_{s'} P(s',a,s) \sum_{a'} \pi(s',a')\left(\gamma R(s',a') + E\left[\sum_{k=2}^{\infty} \gamma^k \mid S_t = s, \pi\right]\right)$$

$$\vdots$$

$$= \sum_a \pi(s,a) R(s,a) + \gamma \sum_a \pi(s,a) \sum_{s'} P(s',a,s) \sum_{a'} \pi(s',a') R(s',a')$$

$$+ \gamma^2 \sum_a \pi(s,a) \sum_{s'} P(s',a,s) \sum_{a'} \pi(s',a') \sum_{s''} P(s'',a',s') \sum_{a''} \pi(s'',a'') R(s'',a'')$$

$$+ \dots$$

$$= E\left[\sum_{t=0}^{\infty} \gamma^t R_t \mid S_0 = s, \pi\right]$$

$$= E[G \mid S_0 = s, \pi]$$

$$= V^{\pi}(s)$$

C) For all $1 \leq i \leq L-2$ we have

$$V^{\pi}(s_i) = \sum_{s'} P(s', \pi(s_i), s_i)\left[E[R(\pi(s_i), s_i) + V^{\pi}(s')]\right]$$

$$= E[R(\pi(s_i), s_i)] + V^{\pi}(s_{i+1})$$

(since the transitions are definit, $P(s_{i+1}, \pi(s_i), s_i) = 1$)

$$< V^{\pi}(s_{i+1})$$

since $\forall i,\ R(\pi(s_i), s_i) < 0 \Rightarrow E[R(\pi(s_i), s_i)] < 0$

Q5

A) The relation fo TD(n) is :

$$G_t^{(n)}(s) = R_{t+1} + \gamma R_{t+2} + \dots + \gamma^{n-1} R_{t+n} + \gamma^n V(S_{t+n})$$

$$V(S_t) \leftarrow V(S_t) + \alpha(G_t^{(n)}(s) - V(S_t)]$$

Now let assum $\gamma = 1$. Then for $n = 1$, thus

$$G_t^{(1)}(C) = R_1 + \gamma V(D)$$

$$V(C) \leftarrow V(C) + \alpha(G_t^{(1)}(C) - \gamma V(C))$$

since $R_1 = 0$, $\gamma = 1$ and $V(C) = V(D) = 0.5$, the value of $V(C)$ doesn't update. similarly $V(D)$ doesn't change, and only $V(E)$ does. with the same argument we can see for $n=2$, only $V(E)$ and $V(D)$ are updated and for $n \geq 3$ the state value for all 3 states change.

B) The value of α, adjusts how much the observation and the previous estimation contribute to updating the next estimation. If α is very large, we give so much weight to the observations, neglecting the previous estimates and causing high variance. If α is very small, it's vice versa and it causes high bias. In both cases, the total error increases.

C) 1) By increasing the number of states, we add to the complexity and the need for more data and episodes. Thus by keeping the othe parameters, the variance and hence the error increases.

2) having more episodes, means having more data, which reduces the variance and so the error. By the "Law of Larg Numbers", the more episodes, e.g. samples we have, our estimation gets closer to the true value.

3) Increasing the number of repetions, eventhough doesn't have the effect of more episodes, as we do not continue the updating process and we reset the experiment each time, but still it reduces varinace and slighty redues the error.

D) The recursive relation for eligibity trace is

$$E_0(s) = 0 \quad , \; E_t(s) = \gamma\lambda E_{t-1}(s) + \mathbb{1}(S_t = s)$$

Now let's denote the so called state by s. The eligibility trace fo this state is maximum, as for all t we have $\mathbb{1}(S_t = s) = 1$. Thus

$$E_t(s) = \gamma\lambda E_{t-1}(s) + 1 = \gamma\lambda(\gamma\lambda E_{t-2}(s) + 1) + 1$$

$$= \dots = \gamma\lambda(\gamma\lambda(\dots(\gamma\lambda E_0(s) + 1) + 1) + 1) + 1$$

$$= \sum_{n=0}^{t} (\gamma\lambda)^n$$

as $t \to \infty$ we have

$$\underset{t\to\infty}{E_t(s)} = \sum_{n=0}^{\infty} (\gamma\lambda)^n = \frac{1}{1 - \gamma\lambda} = 1.25$$