1)

a) $$A = \begin{array}{cc} & \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \end{array} & \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

b) $$X = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -2 & 0.5 \\ 1 & 3 \\ 0 & -1 \end{bmatrix}$$

$$X' = \sigma(AX) \qquad \sigma(\;) = \text{Relu}$$

$$AX = \begin{bmatrix} -1 & -0.5 \\ -1 & 2.5 \\ -1 & 3.5 \\ 2 & 1 \end{bmatrix}$$

$$\therefore X' = \begin{bmatrix} 0 & 0 \\ 0 & 2.5 \\ 0 & 3.5 \\ 2 & 1 \end{bmatrix}$$

c) $D = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$

$\therefore D^{-1} = \begin{bmatrix} 0.5 & 0 & 0 & 0 \\ 0 & 0.33 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.33 \end{bmatrix}$

$\therefore D^{-1}AX = \begin{bmatrix} -0.5 & -0.25 \\ -0.33 & 0.83 \\ -0.5 & 1.75 \\ 0.67 & 0.33 \end{bmatrix}$

$\therefore X' = \begin{bmatrix} 0 & 0 \\ 0 & 0.83 \\ 0 & 1.75 \\ 0.67 & 0.33 \end{bmatrix}$

2a) Starting from state 3

$$R_3^A = 0.25 \times 6 + 0.75 \times 2$$
$$= \frac{3}{2} + \frac{3}{2}$$
$$= 3$$

$$R_3^B = 0.5 \times 3 + 0.5 \times 1$$
$$= 2$$

∴ From this we can see that the expected reward for the next turn is higher for action A than B

b) $\gamma = 0.8$

$$v(\text{state } 1) = 6 + \gamma \times 1 \times v(\text{state } 3)$$
$$= 6 + 0$$
$$= 6$$

$$v(\text{state } 2) = 2 + \gamma \times 1 \times v(\text{state } 3)$$
$$= 2$$

$$v(\text{state }4) = 3 + \gamma\left(0.5\, v(\text{state }4) + 0.5\, v(\text{state }5)\right)$$
$$= 3 + 0.4\, v(\text{state }4) + 0.4\, v(\text{state }5)$$
$$0.6\, v(\text{state }4) - 0.4\, v(\text{state }5) = 3 \quad - (1)$$

Similarly

$$v(\text{state }5) = 1 + \gamma\left(0.5\, v(\text{state }5) + 0.5\, v(\text{state }4)\right)$$

$$0.6\, v(\text{state }5) - 0.4\, v(\text{state }4) = 1 \quad - (2)$$

solving (1) & (2) we get
$v(\text{state }4) = 11$ &
$v(\text{state }5) = 9$

$$q(3,A) = \gamma\left(0.25 \times v(\text{state }1) + 0.75\, v(\text{state }2)\right)$$
$$= \gamma(0.25 \times 6 + 0.75 \times 2)$$
$$= 0.8 \times 3 = 2.4$$

$$q(3,B) = \gamma\left(0.5 \times v(\text{state }4) + 0.5\, v(\text{state }5)\right)$$
$$= \left(\frac{11}{2} + \frac{9}{2}\right) \times 0.8$$
$$= 8$$

Thus taking action B leads to a higher reward.

c) From above

$$v(\text{state }4) = 3 + \gamma\left(0.5v(\text{state }4) + 0.5v(\text{state }5)\right)$$
$$\& \ v(\text{state }5) = 1 + \gamma\left(0.5v(\text{state }5) + 0.5v(\text{state }4)\right)$$

$$\therefore \left(1-\frac{\gamma}{2}\right)v(\text{state }4) + \frac{\gamma}{2}v(\text{state }5) = 3$$
$$\& \left(1-\frac{\gamma}{2}\right)v(\text{state }5) + \frac{\gamma}{2}v(\text{state }4) = 1$$

Solving we get:

$$v(\text{state }4) = \frac{\gamma-3}{\gamma-1}$$
$$v(\text{state }5) = \frac{\gamma+1}{1-\gamma}$$

$$q(3,A) = 3\gamma \quad (\text{from part B})$$
$$q(3,B) = \gamma\left(\frac{1}{2}\left(\frac{\gamma-3}{\gamma-1}\right) + \frac{1}{2}\left(\frac{\gamma+1}{1-\gamma}\right)\right)$$

for equality

$$6 = \frac{\gamma-3}{\gamma-1} + \frac{\gamma+1}{1-\gamma} = \frac{\gamma-3-\gamma-1}{\gamma-1}$$

$$\Rightarrow \gamma = 1-\frac{4}{6} = \frac{1}{3}$$

$\therefore$ for $\gamma < 1/3$ expected reward from A is greater than B

3)

a) all Q are initialized to zero

$$\therefore Q(S_1, A) = Q(S_1, A) + \alpha(r(S_1) + \gamma Q(S_2, a) - Q(S_1, A))$$

$$= 0 + 1(1 + 0.5 \times 0 - 0)$$

$$= 1$$

as we break ties by choosing A $Q(S_1, B)$ is not updated

$\therefore Q(S_1, B) = 0$

b) After 5 steps we are at S_5 and so $Q(S_1, A)$ & $Q(S_1, B)$ will not change

$\therefore Q(S_1, A) = 1$ & $Q(S_1, B) = 0$

c) After N+1 steps we reach back at S_1 & since $Q(S_1, A) > Q(S_1, B)$ we choose action A.

$$\therefore Q(S_1, A) = 1 + 1(1 + 0.5(1) - 1)$$

$$= 1.5$$

$Q(S_1, B) = 0$ (did not change)

at N+5 steps this value will be the same as above as we reach S_4 at N+5

d) As $N \to \infty$

$Q(S_1, B)$ will never change as $Q(S_1, A)$ is always greater than $Q(S_1, B)$

For $Q(S_1, A)$ after $2N$ steps

$$Q(S_1, A) = 1 + \frac{1}{2} + \frac{1}{4}$$

$\therefore$ as $N \to \infty$ $Q(S_1, A) = \dfrac{1}{1 - \frac{1}{2}} = 2$

4)

a) In on-policy RL we update our Q-values based on the actions according to the current policy. The same policy which is used to select actions is also used to update the algorithms. On the other hand off policy algorithms evaluate a policy that may be different from the one being used to select the action.

Examples

On policy: Policy Iteration, value iteration

Off policy: Q learning, expected sarsa

b) If we use just the latest interaction data to train our model we risk getting swayed by correlation. To countract this experience replay provides a large pool of data to train from which helps break correlation & allows the same sample to be used multiple times thereby improving training. It also allows recalling rare occurrences & improves the usage of experience

c) Q learning is off policy while policy gradients are on policy. Q learning aims to learn a single deterministic action from a discrete set by finding the maxima. policy gradients on the other hand learn to map states to actions & can be stochastic thereby working in continuous spaces

d) In actor-critic method the actor is the policy & the critic is the estimated value function. The critic evaluates actions taken by the actor based on the given policy. By interacting with the environment the critic & actor both learn thereby improving them. The critic forces the actor to improve by criticizing the actor. This is akin to a GAN where the discriminator forces the generator to produce a better output