Soit un MDP avec $S = \{s_0, s_1, s_2, s_3\}$ où s_2 est terminal, l'ensemble d'actions $\{a_1, a_2, a_3\}$ et le facteur d'escompte $\gamma = 0.5$. On suppose que toutes les actions sont possibles à partir de chaque état.

Soit une politique donnée π ayant généré les essais suivants :

$(s_0)_1 \rightarrow (s_0)_1 \rightarrow (s_1)_1 \rightarrow (s_1)_1 \rightarrow (s_2)_{10}$
$(s_0)_1 \rightarrow (s_0)_1 \rightarrow (s_3)_2 \rightarrow (s_1)_1 \rightarrow (s_2)_{10}$

1. Estimez les valeurs $V(s)$ de la politique π par estimation directe.

$$\begin{array}{lll}
V(s_0) & = \frac{1}{4}(& 1 + 0.5 \times 1 + 0.5^2 \times 1 + 0.5^3 \times 1 + 0.5^4 \times 10 + \\
& & 1 + 0.5 \times 1 + 0.5^2 \times 1 + 0.5^3 \times 10 + \\
& & 1 + 0.5 \times 1 + 0.5^2 \times 2 + 0.5^3 \times 1 + 0.5^4 \times 10 \\
& & 1 + 0.5 \times 2 + 0.5^2 \times 1 + 0.5^3 \times 10) \\
& = & 2.9375 \\
V(s_1) & = \frac{1}{3}(& 1 + 0.5 \times 1 + 0.5^2 \times 10 + \\
& & 1 + 0.5 \times 10 \\
& & 1 + 0.5 \times 10) \\
& = & 5.3333 \\
V(s_2) & = & 10 \\
V(s_3) & = & 2 + 0.5 \times 1 + 0.5^2 \times 10 \\
& = & 5
\end{array}$$

2. Donnez le système d'équations des valeurs $V(s)$ pour π tel qu'estimé par apprentissage par programmation dynamique adaptative.

$$\begin{array}{lll}
V(s_0) & = & 1 + 0.5 \times \left(\frac{2}{4} V(s_0) + \frac{1}{4} V(s_1) + \frac{1}{4} V(s_3)\right) \\
V(s_1) & = & 1 + 0.5 \times \left(\frac{1}{3} V(s_1) + \frac{2}{3} V(s_2)\right) \\
V(s_2) & = & 10 \\
V(s_3) & = & 2 + 0.5 \times V(s_1)
\end{array}$$

3. Estimez les valeurs $V(s)$ de la politique π par apprentissage par différence temporelle à l'aide d'un taux d'apprentissage $\alpha = 0.1$.

$$
\begin{aligned}
V(s_0) &\leftarrow 1 \text{ (initialisation)} \\
V(s_0) &\leftarrow 1 + 0.1 \times (1 + 0.5 \times 1 - 1) = 1.05 \\
V(s_1) &\leftarrow 1 \text{ (initialisation)} \\
V(s_0) &\leftarrow 1.05 + 0.1 \times (1 + 0.5 \times 1 - 1.05) = 1.095 \\
V(s_1) &\leftarrow 1 + 0.1 \times (1 + 0.5 \times 1 - 1) = 1.05 \\
V(s_2) &\leftarrow 10 \text{ (initialisation)} \\
V(s_1) &\leftarrow 1.05 + 0.1 \times (1 + 0.5 \times 10 - 1.05) = 1.545 \\
V(s_0) &\leftarrow 1.095 + 0.1 \times (1 + 0.5 \times 1.095 - 1.095) = 1.14025 \\
V(s_3) &\leftarrow 2 \text{ (initialisation)} \\
V(s_0) &\leftarrow 1.14025 + 0.1 \times (1 + 0.5 \times 2 - 1.14025) = 1.226225 \\
V(s_3) &\leftarrow 2 + 0.1 \times (2 + 0.5 \times 1.545 - 2) = 2.07725 \\
V(s_1) &\leftarrow 1.545 + 0.1 \times (1 + 0.5 \times 10 - 1.545) = 1.9905
\end{aligned}
$$

Supposez maintenant que les essais suivants aient été générés par un agent faisant de l'apprentissage par renforcement à l'aide du *Q-learning*, suivant une certaine politique d'exploration et avec un taux d'apprentissage $\alpha = 0.1$.

$(s_0)_1 \xrightarrow{a_1} (s_1)_1 \xrightarrow{a_2} (s_1)_1 \xrightarrow{a_2} (s_2)_{10}$
$(s_0)_1 \xrightarrow{a_3} (s_0)_1 \xrightarrow{a_3} (s_1)_1 \xrightarrow{a_1} (s_1)_1 \xrightarrow{a_1} (s_0)_1 \xrightarrow{a_3} (s_2)_{10}$
$(s_0)_1 \xrightarrow{a_2} (s_3)_2 \xrightarrow{a_1} (s_2)_{10}$
$(s_0)_1 \xrightarrow{a_1} (s_0)_1 \xrightarrow{a_1} (s_3)_2 \xrightarrow{a_1} (s_1)_1 \xrightarrow{a_2} (s_2)_{10}$

1. Donnez la liste des mises à jour de la fonction action-valeur. Supposez une initialisation de $Q(s,a)$ à 0 et utilisez un taux d'apprentissage $\alpha = 0.1$.

$$
\begin{aligned}
Q(s_0,a_1) &\leftarrow 0 + 0.1 \times (1 + 0.5 \times \max(0,0,0) - 0.0) = 0.1 \\
Q(s_1,a_2) &\leftarrow 0 + 0.1 \times (1 + 0.5 \times \max(0,0,0) - 0.0) = 0.1 \\
Q(s_2,\text{None}) &\leftarrow 10 \text{ (initialization)} \\
Q(s_1,a_2) &\leftarrow 0.1 + 0.1 \times (1 + 0.5 \times 10 - 0.1) = 0.69 \\
Q(s_0,a_3) &\leftarrow 0 + 0.1 \times (1 + 0.5 \times \max(0.1,0,0) - 0.0) = 0.105 \\
Q(s_0,a_3) &\leftarrow 0.105 + 0.1 \times (1 + 0.5 \times \max(0,0.69,0) - 0.105) = 0.229 \\
Q(s_1,a_1) &\leftarrow 0 + 0.1 \times (1 + 0.5 \times \max(0,0.69,0) - 0.0) = 0.1345 \\
Q(s_1,a_1) &\leftarrow 0.1345 + 0.1 \times (1 + 0.5 \times \max(0.1,0,0.229) - 0.1345) = 0.2325 \\
Q(s_0,a_3) &\leftarrow 0.229 + 0.1 \times (1 + 0.5 \times 10 - 0.229) = 0.8061 \\
Q(s_0,a_2) &\leftarrow 0 + 0.1 \times (1 + 0.5 \times \max(0,0,0) - 0.0) = 0.1 \\
Q(s_3,a_1) &\leftarrow 0 + 0.1 \times (2 + 0.5 \times 10 - 0.0) = 0.7 \\
Q(s_0,a_1) &\leftarrow 0.1 + 0.1 \times (1 + 0.5 \times \max(0.1,0.1,0.8061) - 0.1) = 0.230305 \\
Q(s_0,a_1) &\leftarrow 0.230305 + 0.1 \times (1 + 0.5 \times \max(0.7,0,0) - 0.230305) = 0.3422745 \\
Q(s_3,a_1) &\leftarrow 0.7 + 0.1 \times (2 + 0.5 \times \max(0.2325,0.69,0) - 0.7) = 0.8645 \\
Q(s_1,a_2) &\leftarrow 0.69 + 0.1 \times (1 + 0.5 \times 10 - 0.69) = 1.221
\end{aligned}
$$

2. Quelle serait la politique apprise à la fin ? Donnez l'action choisie par cette politique pour chaque état, excepté l'état terminal.

$$
\begin{aligned}
\pi(s_0) &= \text{argmax}(Q(s_0,a_1),Q(s_0,a_2),Q(s_0,a_3)) = \text{argmax}(0.3422745, 0.1, 0.8061) = a_3 \\
\pi(s_1) &= \text{argmax}(Q(s_1,a_1),Q(s_1,a_2),Q(s_1,a_3)) = \text{argmax}(0.2325, 1.221, 0) = a_2 \\
\pi(s_3) &= \text{argmax}(Q(s_3,a_1),Q(s_3,a_2),Q(s_3,a_3)) = \text{argmax}(0.8645, 0, 0) = a_1
\end{aligned}
$$