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.
2. Donnez le système d'équations des valeurs $V(s)$ pour π tel qu'estimé par apprentissage par programmation dynamique adaptative.
3. Estimez les valeurs $V(s)$ de la politique π par apprentissage par différence temporelle à l'aide d'un taux d'apprentissage $\alpha = 0.1$.

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 $\alpha = 0.1$.
2. Quelle serait la politique apprise à la fin ? Donnez l'action choisie par cette politique pour chaque état, excepté l'état terminal.