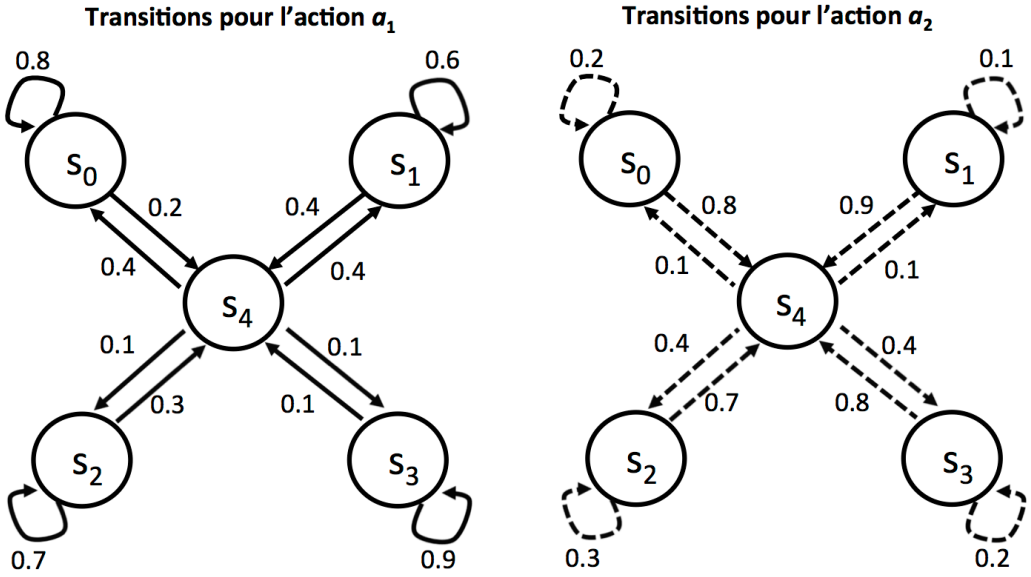
Question 3 (7 points) – *Processus de décision markovien (PDM)* et apprentissage par renforcement

Soit le processus de décision markovien (PDM) ayant l'ensemble d'état $S = \{s_0, s_1, s_2, s_3, s_4\}$, l'ensemble d'actions $A = \{a_1, a_2\}$, la fonction de récompense $R(s_0) = -2$, $R(s_1) = -6$, $R(s_2) = 2$, $R(s_3) = 9$ et $R(s_4) = 0$, un facteur d'escompte $\gamma = 0.5$, ainsi que les distributions de transition (environnement) suivantes:



a) (4 points) Simulez une itération de l'algorithme d'itération par valeurs (*value iteration*) appliqué à ce PDM. Utilisez l'initialisation $V(s_0) = -1$, $V(s_1) = -5$, $V(s_2) = 3$, $V(s_3) = 10$ et $V(s_4) = 1$. À la fin de l'itération, **donnez également la politique qui serait alors retournée par l'algorithme**.