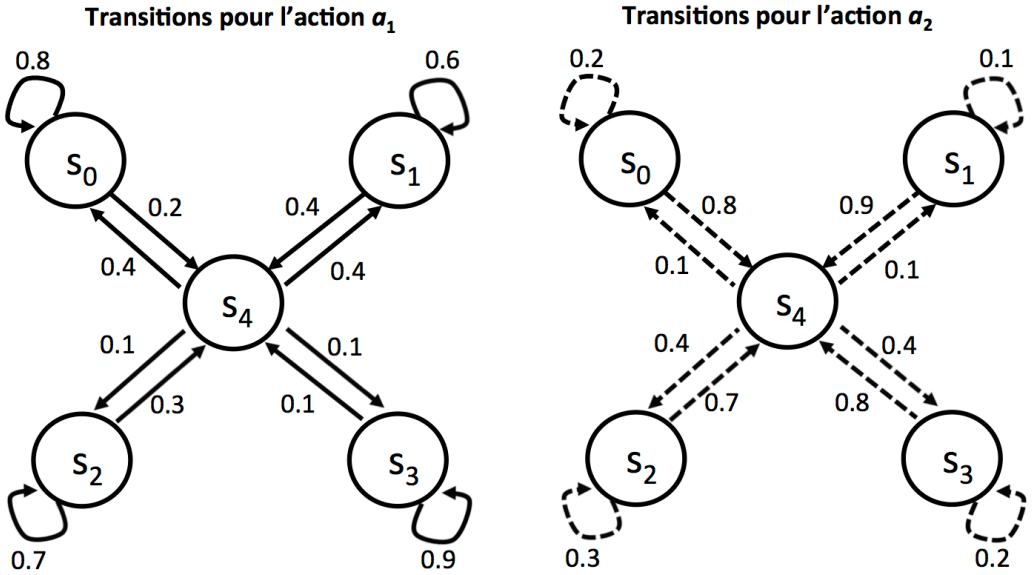
Soit le processus de décision markovien (PDM) ayant l'ensemble d'état $S = \{s_0, s_1, s_2, s_3, s_4\}$, l'ensemble d'actions $A = \{a_1, a_2\}$, la fonction de récompense $R(s_0) = -2$, $R(s_1) = -6$, $R(s_2) = 2$, $R(s_3) = 9$ et $R(s_4) = 0$, un facteur d'escompte $\gamma = 0.5$, ainsi que les distributions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



b) (2 points) Supposons que vous n'ayez pas accès aux distributions de transition, mais que vous ayez accès à un simulateur pour ce PDM, vous permettant de faire de l'apprentissage par renforcement par *Q-learning*. Supposons que vous connaissiez la fonction de récompense. Soit l'initialisation de la fonction action-valeur suivante :

$Q(s_0,a_1) = -1$	$Q(s_0,a_2) = 1$
$Q(s_1,a_1) = -7$	$Q(s_1,a_2) = -4$
$Q(s_2,a_1) = 6$	$Q(s_2,a_2) = 4$
$Q(s_3,a_1) = 15$	$Q(s_3,a_2) = 11$
$Q(s_4,a_1) = 1$	$Q(s_4,a_2) = 2$

Supposons que vous observiez dans une simulation une **transition de l'état s_2 à l'état s_4 après avoir exécuté l'action a_2**. Exécutez la mise à jour de la fonction action-valeur à faire dans le cadre du *Q-learning*.