Équations de Bellman pour la valeur optimale

- Les **équations de Bellman** nous donnent une condition qui est garantie par la valeur V^* des plans optimaux

$$V^*(s) = R(s) + \max_a \; \gamma \sum_{s' \in S} P(s'|s,a)\, V^*(s') \quad \forall\, s \in S$$

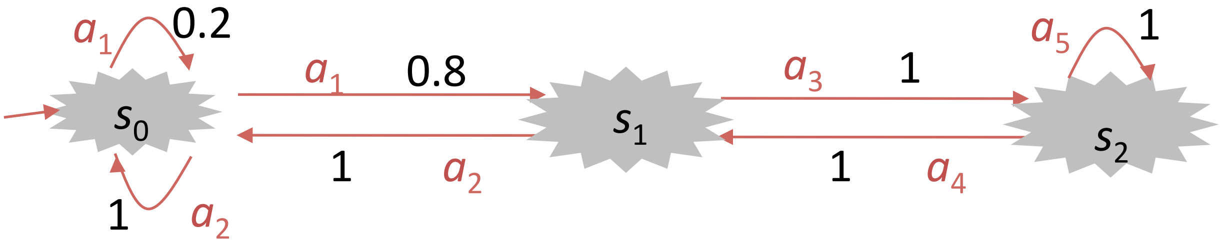
- Deux algorithmes différents pour le calcul du plan optimal:
  - **itération par valeurs** (***value iteration***)
  - **itération par politiques** (***policy iteration***)

Algorithme *value iteration*

1. Initialiser $V(s)$ à 0 pour chaque état s
2. Répéter (jusqu'à ce que le changement en V soit négligeable)
   I. pour chaque état s calculer:
      $$V'(s) \leftarrow R(s) + \max_a \gamma \, \Sigma_{s' \in S} P(s'|s,a) \; V(s')$$
   II. si $\Sigma_{s \in S} |V(s) - V'(s)| \leq$ tolérance, quitter
   III. $V \leftarrow V'$
3. Dériver le plan optimal en choisissant l'**action *a* ayant la meilleure récompense future espérée**, pour chaque état s
   I. $\pi(s) = \operatorname{argmax}_a \Sigma_{s' \in S} P(s'|s,a) \; V(s')$

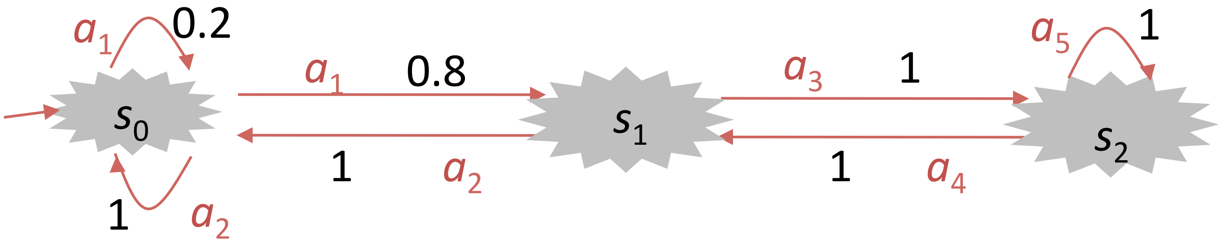
- En mots, on choisit l'action qui **maximise l'espérance** des sommes de récompenses futures

Exemple de MDP



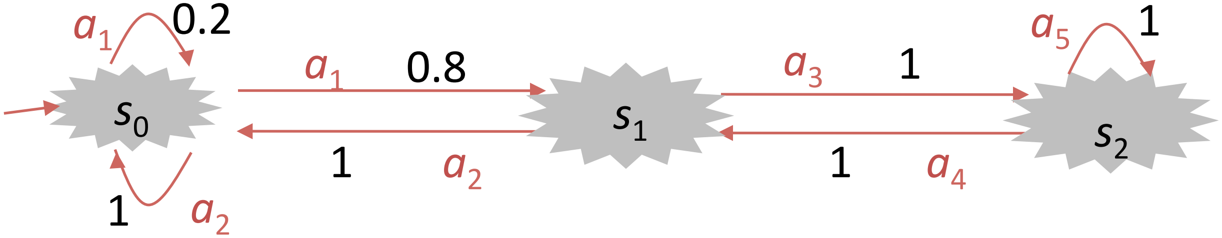
- MDP à 3 états: $S = \{s_0, s_1, s_2\}$
- But: s_2

Exemple de MDP



- MDP à 3 états: $S = \{s_0, s_1, s_2\}$
- Le but (atteindre s_2) est exprimé par une fonction de récompense:
  - $R(s_0) = 0$, $R(s_1)= 0$, $R(s_2) = 1$
- Le facteur d'escompte est γ=0.5
- Notons $r_i=R(s_i)$ et $v_i = V(\pi,s_i)$

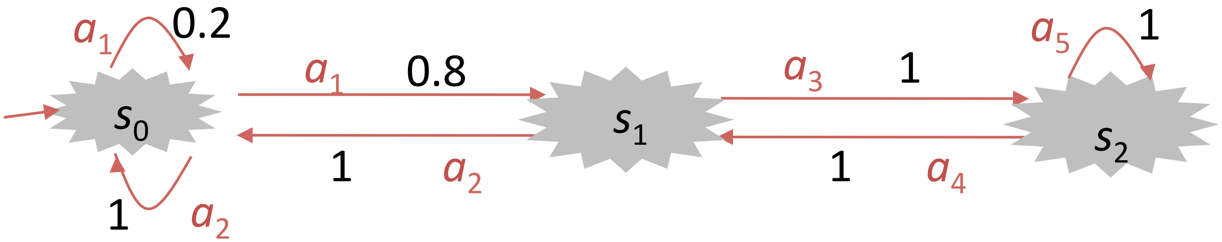
***Value iteration*: initialisation**



- Valeurs initiales fixées à 0:

$v_0 = 0$
$v_1 = 0$
$v_2 = 0$

Value iteration: itération #1

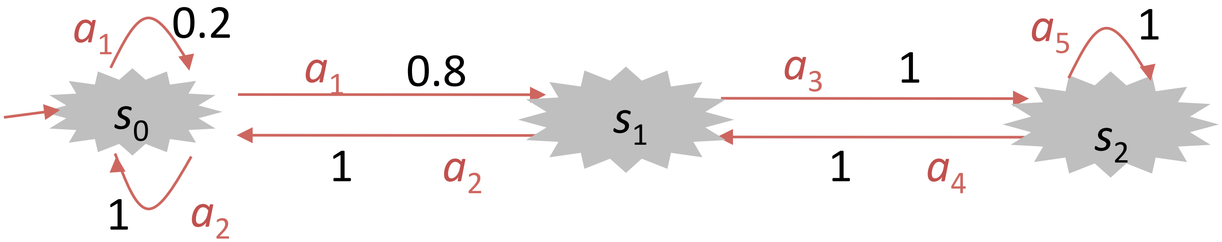


- Mise à jour droite-gauche des valeurs

  $v_0 \leftarrow 0 + 0.5 \max\{\, 0.2\, v_0 + 0.8\, v_1, v_0 \,\} = 0 + 0.5 \max\{\, 0, 0 \,\} = 0$
  $v_1 \leftarrow 0 + 0.5 \max\{\, v_0, v_2 \,\} = 0 + 0.5 \max\{\, 0, 0 \,\} = 0$
  $v_2 \leftarrow 1 + 0.5 \max\{\, v_1, v_2 \,\} = 1 + 0.5 \max\{\, 0, 0 \,\} = 1$

- Les nouvelles valeurs sont $v_0 = 0$, $v_1 = 0$, $v_2 = 1$

Value iteration: itération #2

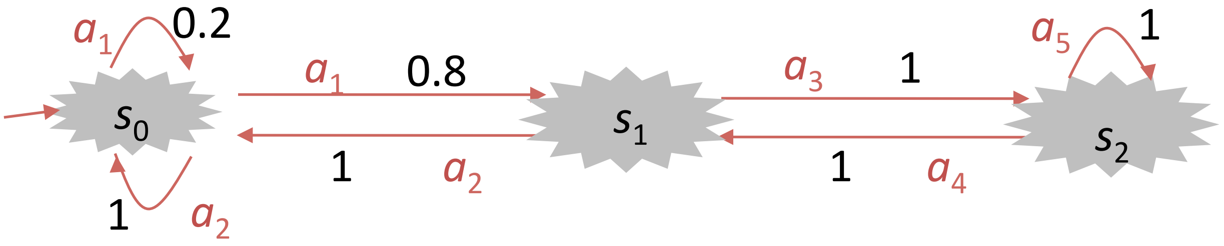


- Mise à jour droite-gauche des valeurs

  $v_0 \leftarrow 0 + 0.5 \max\{ 0.2\, v_0 + 0.8\, v_1, v_0 \} = 0 + 0.5 \max\{ 0, 0 \} = 0$
  $v_1 \leftarrow 0 + 0.5 \max\{ v_0, v_2 \} = 0 + 0.5 \max\{ 0, 1 \} = 0.5$
  $v_2 \leftarrow 1 + 0.5 \max\{ v_1, v_2 \} = 1 + 0.5 \max\{ 0, 1 \} = 1.5$

- Les nouvelles valeurs sont $v_0 = 0$, $v_1 = 0.5$, $v_2 = 1.5$

Value iteration: itération #3

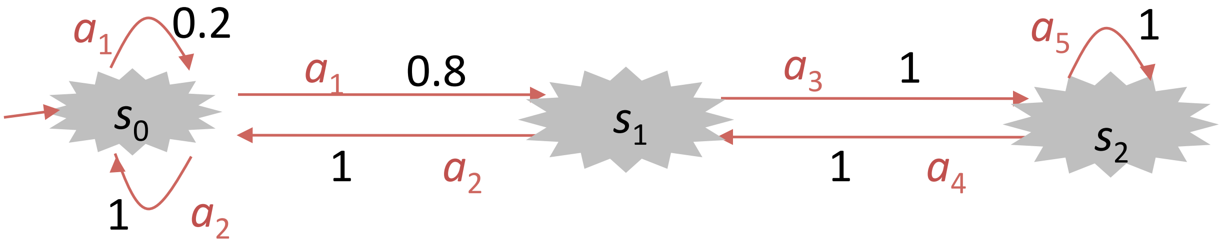


- Mise à jour droite-gauche des valeurs

  $v_0 \leftarrow 0 + 0.5 \max\{ 0.2\ v_0 + 0.8\ v_1, v_0 \} = 0 + 0.5 \max\{ 0.8 * 0.5, 0 \} = 0.2$
  $v_1 \leftarrow 0 + 0.5 \max\{ v_0, v_2 \} = 0 + 0.5 \max\{ 0, 1.5 \} = 0.75$
  $v_2 \leftarrow 1 + 0.5 \max\{ v_1, v_2 \} = 1 + 0.5 \max\{ 0.5, 1.5 \} = 1.75$

- Les nouvelles valeurs sont $v_0 = 0.2$, $v_1 = 0.75$, $v_2 = 1.75$

Value iteration: itération #3



- Si on arrêtait à la 3[e] itération, le plan retourné serait

$\pi(s_0)$ = argmax{ 0.2 v_0 + 0.8 v_1, v_0 } = argmax{ 0.2*0.2+0.8*0.75, 0.2} = a_1
$\pi(s_1)$ = argmax{ v_0, v_2 } = argmax{ 0.2, 1.75 } = a_3
$\pi(s_2)$ = argmax{ v_1, v_2 } = argmax{ 0.75, 1.75 } = a_5

- On a déjà le plan optimal!
  - ça aurait pu ne pas être le cas, seulement garanti si on boucle jusqu'à convergence

Démonstration de *value iteration*

http://planiart.usherbrooke.ca/~eric/ift615/demos/vi/