Apprentissage par renforcement actif

- Dans le cas **passif**, le plan à suivre est pré-déterminé
  - peu utile si on ne connaît pas le plan optimal à suivre
- Dans le cas **actif**, l'agent doit aussi chercher le plan optimal
  - l'agent doit **simultanément** chercher le plan optimal et sa fonction de valeur
  - ***V*(*s*) est maintenant une estimation de la fonction de valeur du plan optimal**
- Dans le cas PDA, trois changements sont à faire
  - on va estimer $P(s'|s,a)$ pour plus d'une action a (toujours à partir des fréquences)
  - on applique ***value iteration*** au MDP estimé (c.-à-d. on résout les équations pour la **politique optimale**)
  - l'action choisie par l'agent devient

$$\pi(s) = \underset{a \in A(s)}{\text{argmax}} \sum_{s' \in S} P(s'|s,a)\, V(s')$$

Appentissage actif avec PDA

function ~~PASSIVE~~ ACTIVE-ADP-AGENT(*percept*) returns an action

inputs: *percept*, a percept indicating the current state s' and reward signal r'

persistent: π, a fixed policy
mdp, an MDP with model P, rewards R, discount γ
U, a table of utilities, initially empty
N_{sa}, a table of frequencies for state–action pairs, initially zero
$N_{s'|sa}$, a table of outcome frequencies given state–action pairs, initially zero
s, a, the previous state and action, initially null

if s' is new **then** $U[s'] \leftarrow r'$; $R[s'] \leftarrow r'$
if s is not null **then**
  increment $N_{sa}[s,a]$ and $N_{s'|sa}[s',s,a]$
  for each t such that $N_{s'|sa}[t,s,a]$ is nonzero **do**
    $P(t|s,a) \leftarrow N_{s'|sa}[t,s,a] \,/\, N_{sa}[s,a]$
$U \leftarrow$ POLICY-EVALUATION(π, U, mdp)
if s'.TERMINAL? **then** $s, a \leftarrow$ **null else** $s, a \leftarrow s', \pi[s']$
return a

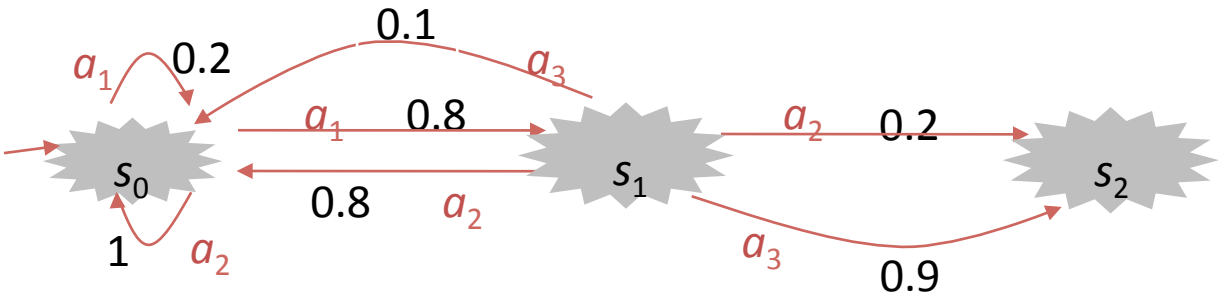
Value iteration

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

$$\text{argmax}_{a \in A(s)} \sum_{s \in S} P(s|s',a)\, V(s)$$

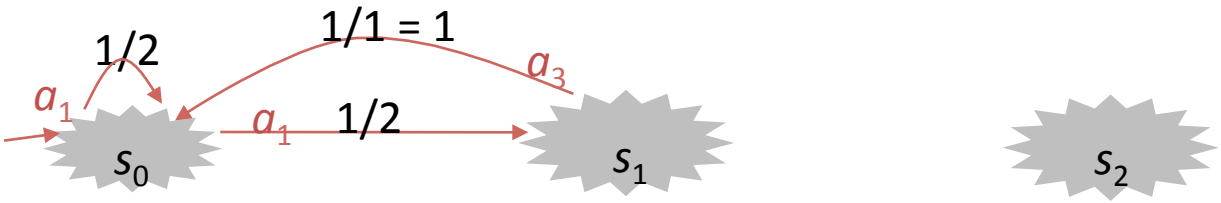
Appentissage actif avec PDA

- Rappel de l'exemple :



- On a des actions possibles différentes, pour chaque état
  - $A(s_0) = \{a_1, a_2\}$
  - $A(s_1) = \{a_2, a_3\}$
  - $A(s_2) = \{\}$

Appentissage actif avec PDA



- Observations: $(s_0)\ _{-0.1} \xrightarrow{a_1} (s_0)\ _{-0.1} \xrightarrow{a_1} (s_1)\ _{-0.1} \xrightarrow{a_3} (s_0)\ _{-0.1}$

$V(s_0) = -0.1 + 0.5 \max\{ 0.5\ V(s_0) + 0.5\ V(s_1), 0 \}$
$V(s_1) = -0.1 + 0.5 \max\{ 0, V(s_0) \}$
$V(s_2) = 1$

value iteration →

$V(s_0) = -0.1$
$V(s_1) = -0.1$
$V(s_2) = 1$

- Pour choisir quelle action prendre, on compare
  - $\Sigma_{s \in S} P(s|s',a_2)\ V(s) = 0$ (puisque $P(s|s',a_2)$ pas appris encore pour a_2)
  - $\Sigma_{s \in S} P(s|s',a_1)\ V(s) = 0.5\ V(s_0) + 0.5\ V(s_1) = -0.1$
- **L'action choisie par l'agent est donc a_2**