Approche par programmation dynamique adaptative

- Un problème avec l'approche par PDA est qu'on **doit mettre à jour toutes les valeurs de *V*(*s*), pour tout s, après chaque observation**
  - très coûteux en pratique si le nombre d'états est grand (ex.: exponentiel)
  - inutile pour un état *s* qui n'est pas atteignable via l'état de la nouvelle observation

- On doit résoudre les équations de $V(s)$ parce qu'on estime $V(s)$ seulement indirectement, via notre estimation de $P(s'|s, a)$

- Serait-il possible d'estimer directement $V(s)$ et tenir compte des interactions entre les valeurs, sans avoir à passer par $P(s'|s, a)$?

Apprentissage par différence temporelle

- Observation: si la transition de s vers s' a une probabilité de 1, on a que

$$\begin{aligned} V(s) &= R(s) + \gamma \ \Sigma_{s' \in S} P(s'|s, \pi(s)) \ V(s') \\ &= R(s) + \gamma \ V(s') \end{aligned}$$

- Plutôt que d'attendre la fin de l'essai pour mettre à jour notre estimation de $V(s)$, on pourrait la rapprocher de $R(s) + \gamma \ V(s')$:

$$V(s) \leftarrow (1\text{-}\alpha) \ V(s) + \alpha \ (R(s) + \gamma \ V(s'))$$

  où α est un **taux d'apprentissage**, entre 0 et 1

- On obtient la règle d'apprentissage **par différence temporelle** (***temporal difference***) ou **TD**

$$V(s) \leftarrow V(s) + \alpha \ (\ R(s) + \gamma \ V(s') - V(s) \ )$$

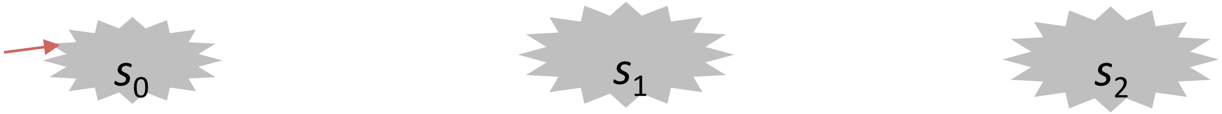
Apprentissage par différence temporelle

function PASSIVE-TD-AGENT(*percept*) **returns** an action
  inputs: *percept*, a percept indicating the current state s' and reward signal r'
  persistent: π, a fixed policy
    U, a table of utilities, initially empty
    N_s, a table of frequencies for states, initially zero
    s, a, r, the previous state, action, and reward, initially null

  if s' is new **then** $U[s'] \leftarrow r'$
  if s is not null **then**
    increment $N_s[s]$
    $U[s] \leftarrow U[s] + \alpha(N_s[s])(r + \gamma\, U[s'] - U[s])$
  if s'.TERMINAL? **then** $s, a, r \leftarrow$ null **else** $s, a, r \leftarrow s', \pi[s'], r'$
  return a

utile si on veut varier le taux d'apprentissage

Apprentissage par différence temporelle



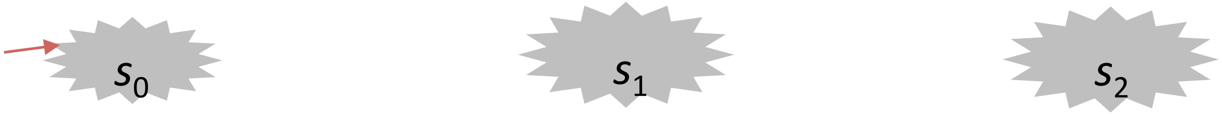
- Initialisation

$$\left.\begin{array}{l} V(s_0) = 0 \\ V(s_1) = 0 \\ V(s_2) = 0 \end{array}\right\} \text{ si on connaît } R(s)\text{, on peut tout initialiser } V(s) \text{ à } R(s)$$

- On va utiliser $\alpha = 0.1$

Hugo Larochelle et Froduald Kabanza

Apprentissage par différence temporelle

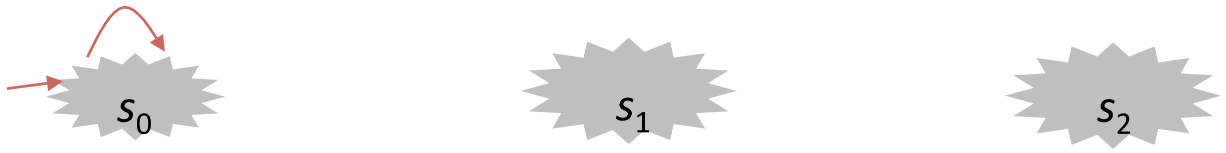


- Observations: $(s_0)\ _{-0.1}$

  $V(s_0) \leftarrow -0.1$ parce que s_0 est visité pour la première fois
  $V(s_1) = 0$
  $V(s_2) = 0$

- On va utiliser $\alpha = 0.1$

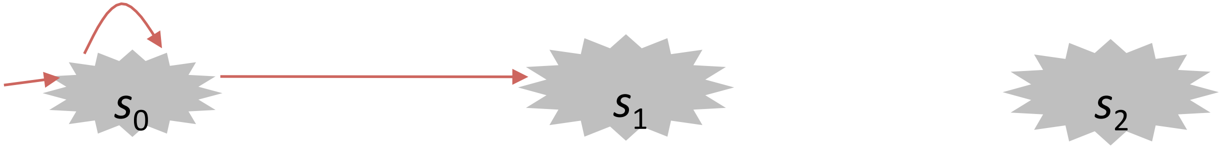
Apprentissage par différence temporelle



- Observations: $(s_0)_{\ -0.1} \rightarrow (s_0)_{\ -0.1}$

$$V(s_0) \leftarrow V(s_0) + 0.1\ (R(s_0) + 0.5\ V(s_0) - V(s_0)\ )$$
$$= -0.1 + 0.1\ (-0.1 - 0.05 + 0.1)$$
$$= -0.105$$
$$V(s_1) = 0$$
$$V(s_2) = 0$$

Apprentissage par différence temporelle



- Observations: $(s_0)_{\,-0.1} \rightarrow (s_0)_{\,-0.1} \rightarrow (s_1)_{\,-0.1}$

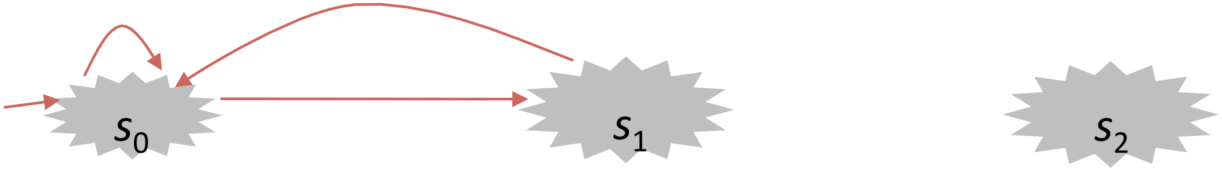
  $V(s_1) \leftarrow -0.1$ parce que s_1 est visité pour la première fois

  ---

  $V(s_0) \leftarrow V(s_0) + 0.1\,(R(s_0) + 0.5\,V(s_1) - V(s_0))$
  $= -0.105 + 0.1\,(-0.1 - 0.05 + 0.105)$
  $= -0.1095$

  $V(s_1) = -0.1$
  $V(s_2) = 0$

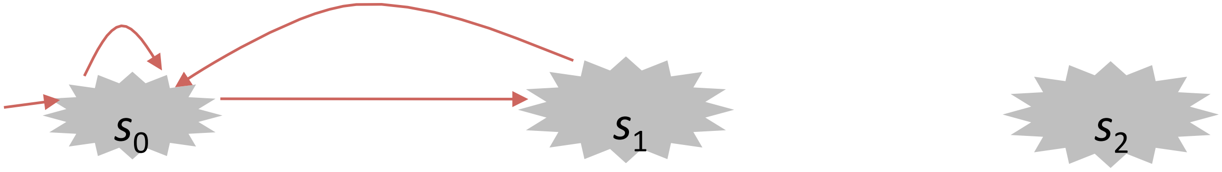
Apprentissage par différence temporelle



- Observations: $(s_0)_{-0.1} \rightarrow (s_0)_{-0.1} \rightarrow (s_1)_{-0.1} \rightarrow (s_0)_{-0.1}$

  $V(s_0) = -0.1095$
  $V(s_1) \leftarrow V(s_1) + 0.1\ (R(s_1) + 0.5\ V(s_0) - V(s_1)\ )$
  $\qquad = -0.1 + 0.1\ (-0.1 - 0.05475 + 0.1)$
  $\qquad = -0.105475$
  $V(s_2) = 0$

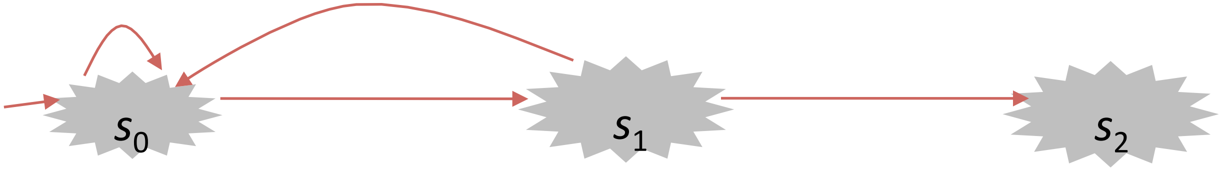
Apprentissage par différence temporelle



- Observations: $(s_0)_{\,-0.1} \rightarrow (s_0)_{\,-0.1} \rightarrow (s_1)_{\,-0.1} \rightarrow (s_0)_{\,-0.1} \rightarrow (s_1)_{\,-0.1}$

  $V(s_0) \leftarrow V(s_0) + 0.1\ (R(s_0) + 0.5\ V(s_1) - V(s_0)\ )$
  $= -0.1095 + 0.1\ (-0.1 - 0.0527375 + 0.1095)$
  $= -0.11382375$
  $V(s_1) = -0.105475$
  $V(s_2) = 0$

Apprentissage par différence temporelle



- Observations: $(s_0)\ _{-0.1} \rightarrow (s_0)\ _{-0.1} \rightarrow (s_1)\ _{-0.1} \rightarrow (s_0)\ _{-0.1} \rightarrow (s_1)\ _{-0.1} \rightarrow (s_2)\ _1$

fin de l'essai

$V(s_2) \leftarrow 1$ parce que s_2 est visité pour la première fois

---

$V(s_0) = -0.11382375$

$V(s_1) \leftarrow V(s_1) + 0.1\ (R(s_1) + 0.5\ V(s_2) - V(s_1))$

$= -0.105475 + 0.1\ (1 + 0.5 + 0.105475\ )$

$= 0.0550725$

$V(s_2) = 1$