Rappel: processus de décision markovien

- Un **processus de décision markovien** (***Markov decision process***, ou **MDP**) est défini par:
  - un **ensemble d'états** S (incluant un étant initial s_0)
  - un **ensemble d'actions** possibles *Actions*(s) (ou $A(s)$) lorsque je me trouve à l'état s
  - un **modèle de transition** $P(s'|s, a)$, où a $\in A(s)$
  - une **fonction de récompense** $R(s)$ (utilité d'être dans l'état s)

- Un **plan** (**politique**) π est un ensemble de décisions qui associe un état s à une action $a = \pi(s)$

Hugo Larochelle et Froduald Kabanza

Rappel: processus de décision markovien

- La **fonction de valeur** $V(s)$ d'un plan est donnée par les équations

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

  où γ est un facteur d'escompte donné

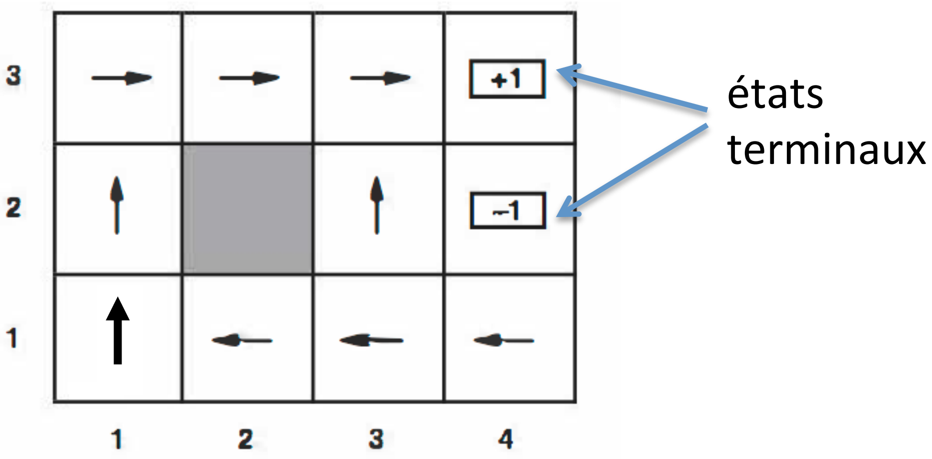
- Plutôt que $V(s)$, on note parfois $V(\pi,s)$, ou $U^{\pi}(s)$ dans le livre

- **NOUVEAU**: on va supposer l'existence d'**états terminaux**
  - lorsque l'agent atteint cet état, la simulation est arrêtée
  - on s'intéresse à la somme des récompenses jusqu'à l'atteinte d'un état terminal
    - ex.: au tic-tac-toe, l'état terminal est une grille de fin de partie (c.-à-d. une grille complète ou une grille où un des joueurs a gagné)

Apprentissage par renforcement passif

- **Définition**: soit un plan π donné, apprendre la fonction de valeur sans connaître $P(s'|s, a)$

- **Exemple illustratif**: déplacement sur une grille 3 x 4
  - plan π illustré parles flèches
  - $R(s)$ = -0.04 partout sauf aux états terminaux
  - l'environnement est stochastique
  - l'agent arrête aux états terminaux
  - on utilise γ=1



Apprentissage par renforcement passif

- **Définition**: soit un plan π donné, apprendre la fonction de valeur sans connaître $P(s'|s, \text{a})$

- Puisqu'on ne connaît pas $P(s'|s, \text{a})$ on doit apprendre à partir d'**essais** (***trials***)
  1. $(1,1)_{-.04} \rightarrow (1,2)_{-.04} \rightarrow (1,3)_{-.04} \rightarrow (1,2)_{-.04} \rightarrow (1,3)_{-.04} \rightarrow (2,3)_{-.04} \rightarrow (3,3)_{-.04} \rightarrow (4,3)_{+1}$
  2. $(1,1)_{-.04} \rightarrow (1,2)_{-.04} \rightarrow (1,3)_{-.04} \rightarrow (2,3)_{-.04} \rightarrow (3,3)_{-.04} \rightarrow (3,2)_{-.04} \rightarrow (3,3)_{-.04} \rightarrow (4,3)_{+1}$
  3. $(1,1)_{-.04} \rightarrow (2,1)_{-.04} \rightarrow (3,1)_{-.04} \rightarrow (3,2)_{-.04} \rightarrow (4,2)_{-1}$

- Comment estimer la fonction de valeurs $V(s)$ à partir de ces essais?

Approche par estimation directe

- Approche la plus simple: **calculer la moyenne de ce qui est observé** dans les essais

- Essais
  1. $(1,1)_{-.04} \rightarrow (1,2)_{-.04} \rightarrow (1,3)_{-.04} \rightarrow (1,2)_{-.04} \rightarrow (1,3)_{-.04} \rightarrow (2,3)_{-.04} \rightarrow (3,3)_{-.04} \rightarrow (4,3)_{+1}$
  2. $(1,1)_{-.04} \rightarrow (1,2)_{-.04} \rightarrow (1,3)_{-.04} \rightarrow (2,3)_{-.04} \rightarrow (3,3)_{-.04} \rightarrow (3,2)_{-.04} \rightarrow (3,3)_{-.04} \rightarrow (4,3)_{+1}$
  3. $(1,1)_{-.04} \rightarrow (2,1)_{-.04} \rightarrow (3,1)_{-.04} \rightarrow (3,2)_{-.04} \rightarrow (4,2)_{-1}$

- Estimation de V((1,1))
  - dans l'essai 1, la somme des récompenses à partir de (1,1) est 0.72
  - dans l'essai 2, on observe également 0.72
  - dans l'essai 3, on observe plutôt -1.16
  - l'estimation directe de V((1,1)) est donc (0.72+0.72-1.16)/3 = 0.09333

Approche par estimation directe

- Approche la plus simple: **calculer la moyenne de ce qui est observé** dans les essais

- Essais
  1. $(1,1)_{-.04} \rightarrow (1,2)_{-.04} \rightarrow (1,3)_{-.04} \rightarrow (1,2)_{-.04} \rightarrow (1,3)_{-.04} \rightarrow (2,3)_{-.04} \rightarrow (3,3)_{-.04} \rightarrow (4,3)_{+1}$
  2. $(1,1)_{-.04} \rightarrow (1,2)_{-.04} \rightarrow (1,3)_{-.04} \rightarrow (2,3)_{-.04} \rightarrow (3,3)_{-.04} \rightarrow (3,2)_{-.04} \rightarrow (3,3)_{-.04} \rightarrow (4,3)_{+1}$
  3. $(1,1)_{-.04} \rightarrow (2,1)_{-.04} \rightarrow (3,1)_{-.04} \rightarrow (3,2)_{-.04} \rightarrow (4,2)_{-1}$

- Estimation de *V*((1,2))
  - dans l'essai 1, l'état (1,2) est visité deux fois, avec des sommes de récompenses à partir de (1,2) de 0.76 et 0.84
  - dans l'essai 2, on observe 0.76
  - l'essai 3 ne visite pas (1,2)
  - l'estimation directe de *V*((1,2)) est donc (0.76+0.84+0.76)/3 = 0.78666