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 $Actions(s)$ 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** π est un ensemble de décisions, spécifiant à chaque état s une action $a = \pi(s)$ à exécuter

Valeur d'un plan

- La **valeur** $V(\pi,s)$ d'un plan π à l'état s
  - récompenses accumulées en moyenne si l'on suit le plan π et qu'on débute à l'état s
  - peut définir de façon récursive :

$$V(\pi,s) = \underbrace{R(s)}_{\text{récompense actuelle}} + \gamma \underbrace{\sum_{s' \in S} P(s'|s, \pi(s))\, V(\pi, s')}_{\text{somme des récompenses futures espérée}}$$

  - γ: **facteur d'escompte** ($0 < \gamma < 1$), soit l'importance relative des récompenses futures

Plan optimal

- Un plan π **domine** un plan π' si les deux conditions suivantes sont réunies:
  - $V(\pi,s) >= V(\pi',s)$ pour tout état *s*
  - $V(\pi,s) > V(\pi',s)$ pour au moins un *s*

- Un plan est **optimal** s'il n'est pas dominé par un autre
  - il peut y avoir plusieurs plans optimaux, mais ils ont tous la même valeur
  - on peut avoir deux plans **incomparables** (aucun ne domine l'autre)
    - » la dominance induit une fonction d'ordre partiel sur les plans

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

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

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

- Si nous pouvons calculer V^*, nous pourrons calculer un plan optimal aisément:
  - il suffit de choisir dans chaque état s l'action qui maximise $V^*(s)$ (c.-à-d. le argmax)