Raisonnement sur l'incertain

Introduction

- Le monde réel est incertain:
  - Incomplétude de la connaissance de l'état réel du monde (où se trouve la boîte bleue? la porte P est-elle ouverte ou fermée?)
  - Imprécision de la connaissance de l'état réel du monde (position du robot? position du cube rouge?)
  - Résultat des actions (l'objet manipulé a glissé, le robot a dérapé)
  - Evolution de l'état du monde par d'autres agents.

Besoin

- Nécessité de tenir compte de l'incertain en prenant les meilleures décisions possibles.
- Impossibilité de représenter explicitement toutes les contingences (combinatoire) pour produire un plan conditionnel.
- Raisonnement permettant de tenir compte de résultats différents des actions et d'états du monde partiellement connus.

Décision dans l'incertain

- Utilisation de la notion d'*utilité*
- recherche d'une *politique* maximisant une utilité, plutôt que d'un *plan* minimisant un coût.

Représentations

- Théorie des probabilités et raisonnement bayésien.
- Autres approches:
  - raisonnement flou (L. Zadeh),
  - théorie de "l'évidence" (Dempster-Shafer).
- Nous utiliserons les probabilités, formalisme et cadre rigoureux et efficace pour traiter des informations incertaines.

Rappels de probabilités

- Définitions, axiomes
- Probabilités conditionnelles
- Théorème de Bayes
- Raisonnement bayesien

Définitions

- Événement aléatoire: événement qui peut ou non se réaliser au cours d'une expérience ou une observation:
  - pile ou face d'une pièce de monnaie
  - 1 à 6 sur un dé
  - Accident d'avion
- La « probabilité » est une valeur numérique qui quantifie la possibilité de réalisation de l'événement

Définitions: probabilité

- Deux manières de définir les probabilités:
  - Fréquentiste. La probabilité exprime la fréquence avec laquelle l'événement se réalise au cours d'un nombre croissant d'observations (loi des grands nombres):

$$P(evt) = \underset{nbre\ observations \to \infty}{Lim} \frac{\text{nbre de cas événement observé}}{\text{nbre total de cas}}$$

  - Subjective. La probabilité exprime la croyance dans l'occurrence d'un événement. N'exprime pas une réalité.

Probabilités

- Probabilité de l'événement *X*: *P(X)*
  - $P(\neg X)$ ou $P(\bar{X})$: Probabilité que *X* ne se produise pas (événement complémentaire de *X*).

$$P(X) + P(\neg X) = 1$$

  - Probabilité que se réalise l'un ou l'autre de deux événements *X, Y* : $P(X \text{ ou } Y) = P(X \vee Y) = P(Y \vee X)$
  - Probabilité que se réalise l'un et l'autre de deux événements:

$$P(X \text{ et } Y) = P(X \wedge Y) = P(Y \wedge X) = P(X, Y) = P(Y, X)$$

Probabilités totales

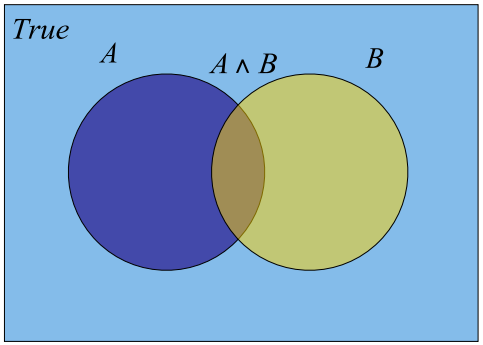
- La probabilité que se réalise l'un ou l'autre de deux événements est la somme de leurs probabilités diminuée de la probabilité qu'ils surviennent simultanément:

$$P(X \vee Y) = P(X) + P(Y) - P(X \wedge Y)$$

- Si X et Y sont disjoints (incompatibles):

$$P(X \vee Y) = P(X) + P(Y) \quad (P(X \wedge Y) = 0)$$

Probabilités totales



Variables aléatoires

- Variable (discrète ou continue) dont la variation ne suit pas une une loi déterministe.
- La valeur réelle est inaccessible. On peut en connaître la *probabilité*.
- Une loi probabiliste permet de caractériser l'évolution de la variable: probabilités de l'ensemble des valeurs possibles, ou distribution de probabilités (densité dans le cas des variables continues).

Variables aléatoires

- Moyenne ou espérance mathématique ou moment d'ordre 1:
  - V.a. X discrète prenant n valeurs, de probabilités $p(x_i)$:

$$E(X) = \sum_{i=1}^{n} p(x_i)x_i$$

  - V.a. X continue sur un intervalle [a, b], de densité de probabilité $p(x)$:

$$E(X) = \int_a^b xp(x)dx$$

Variables aléatoires

- Dispersion: distribution autour de la moyenne
  - Variance de X : moyenne des carrés des écarts à la moyenne (moment d'ordre 2):

$$var(X) = E[(X - E[X])^2] = E[X^2] - E^2[X]$$

  - Ecart-type: $\sigma_X = \sqrt{var(X)}$

Probabilités conditionnelles

- Probabilité d'un événement selon qu'un autre événement s'est déjà produit ou en fonction d'un contexte (ex: probabilité qu'il pleuve s'il y a des nuages).
- Notation: *P(X|Y)* = probabilité que l'événement *X* se produise sachant que *Y* s'est produit:

$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

- Indépendance en probabilité: deux événements sont indépendants (en probabilité) si $P(X \mid Y) = P(X)$
- d'où dans ce cas:

$$P(X,Y) = P(X \mid Y)P(Y) = P(X)P(Y)$$

Proba totales et conditionnement

- Probabilités totales: si Y et $\neg Y$ sont deux événements incompatibles alors:

$$P(X) = P(X \mid Y)P(Y) + P(X \mid \neg Y)P(\neg Y)$$

- Plus généralement, si les Y_i sont une partition (événements incompatibles les uns avec les autres):

$$P(X) = \sum_i P(X \mid Y_i)P(Y_i)$$

Théorème de Bayes

- Probabilités conditionnelles: $P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$
- Symétriquement: $P(Y \mid X) = \frac{P(Y,X)}{P(X)}$
- Or: $P(X,Y) = P(Y,X)$
- D'où: $P(X,Y) = P(Y,X) = P(X \mid Y)P(Y) = P(Y \mid X)P(X)$

Théorème de Bayes: $$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

Le théorème de Bayes permet de mettre à jour la probabilité d'un événement, d'une hypothèse, compte tenu d'indices observés.

Théorème de Bayes

- On peut écrire que la probabilité de Y est la somme des probabilités de deux événements disjoints: $(\text{Y et } \overline{\text{X}})$ et (Y et X)
- D'où:
$$P(X \mid Y) = \frac{P(X,Y)}{P(Y)} = \frac{P(Y \mid X)P(X)}{P(Y,X) + P(Y,\overline{X})}$$
- Et:
$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y \mid X)P(X) + P(Y \mid \overline{X})P(\overline{X})}$$
- Le dénominateur de la dernière formule est une constante de normalisation.

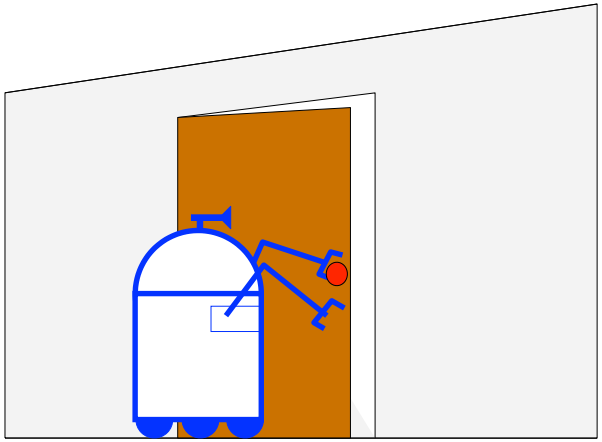
Probabilités, IA et Robotique

- Représentation et traitement des incertitudes par les outils probabilistes
  - L'état du monde partiellement observé par les capteurs est *estimé*.
  - L'exécution des actions peut produire des incertitudes. Traitement du *non-déterminisme*.
  - La production de plans conditionnels est fastidieuse, voire impossible. Elaboration d'une *politique* basée sur la maximisation d'une *utilité*.

Probabilités, IA et Robotique

- Approche générale: utilisation du formalisme et du raisonnement bayésien.
  - Espace d'états S: $\{s_i\}$
  - Ensemble d'actions A : $\{a_j\}$
  - Non déterminisme des actions: Probabilités de transition $P(s' \mid a,s)$

Exemple: fermer la porte

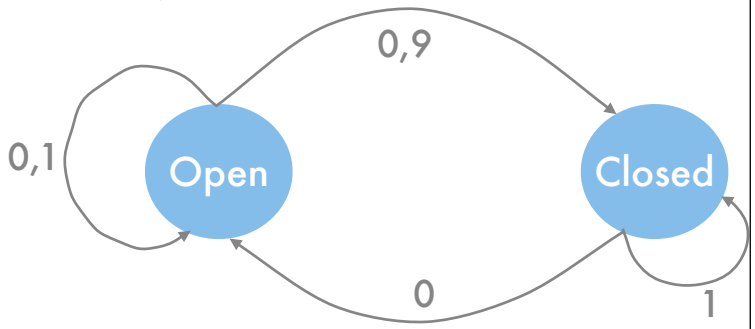


Fermer la porte: Transitions d'état

$P(s' \mid a,s)$ pour a = "fermer_porte":

Si la porte est ouverte, l'action réussit dans 90% des cas.

Si la porte est fermée, l'action réussit dans 100% des cas



On donne la connaissance *a priori*: probabilité porte ouverte 5/8

Intégration des effets des actions

Cas continu:

$$P(s' \mid a) = \int P(s' \mid a, s) P(s) ds$$

Cas discret:

$$P(s' \mid a) = \sum_s P(s' \mid a, s) P(s)$$

Croyance après l'action "fermer"

$$\begin{aligned} P(\text{Closed} \mid close) &= \sum_s P(\text{Closed} \mid close, s) P(s) \\ &= P(\text{Closed} \mid close, \text{Open}) P(\text{Open}) \\ &\quad + P(\text{Closed} \mid close, \text{Closed}) P(\text{Closed}) \\ &= \frac{9}{10}\frac{5}{8} + \frac{1}{1}\frac{3}{8} = \frac{15}{16} \end{aligned}$$

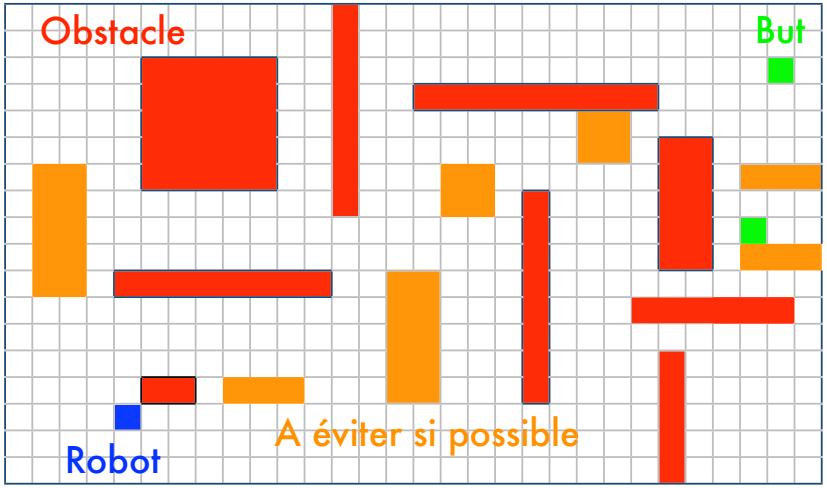
$$P(\text{Open} \mid close) = 1 - P(\text{Closed} \mid close)$$

Processus Décisionnels de Markov (PDM-MDP)

Exemple: navigation

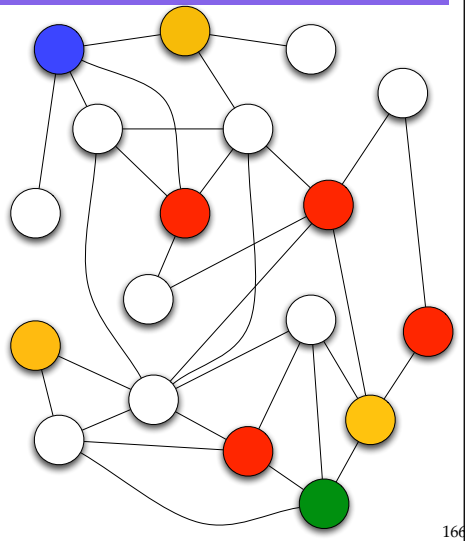
- Le robot doit atteindre un but dans un environnement comportant des zones libres, des zones infranchissables et des zones franchissables mais dangereuses.
- On suppose l'environnement et les positions du robot et du but parfaitement connus (observabilité).

Exemple: navigation



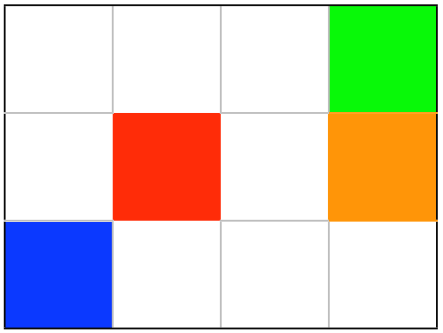
Exemple

- Exemple généralisable: états indifférents, interdits, dangereux, souhaitables, ...
- Espace d'états et transitions



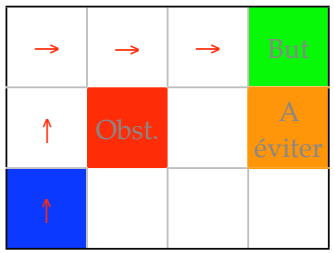
Exemple: navigation

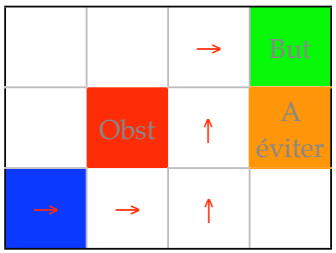
Zoom



Modèle des Actions

- Actions possibles dans chaque état: Nord (N,↑); Sud (S,↓), Est (E, →), Ouest (W, ←)
- Actions déterministes: donnent les effets attendus avec certitude. Deux chemins possibles pour atteindre le but.

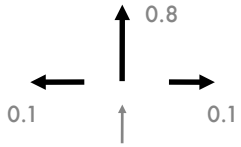




- Actions non déterministes: donnent les effets avec une probabilité donnée.

Actions non déterministes

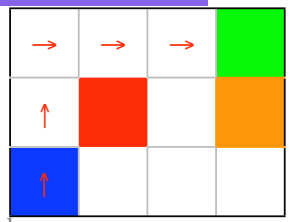
- **Non déterminisme**: les actions ne sont pas fiables.
  - L'effet désiré n'est réalisé qu'avec une certaine probabilité. Exemple:
    - P(N|N)= 0.8
    - P(W|N)= 0.1
    - P(E|N)= 0.1



Action: Nord

Exemple

- En exécutant la séquence d'actions [N,N,E,E,E]:

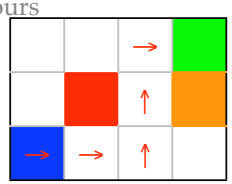


  - Probabilité d'effectuer les effets désirés:

    P_1= 0.8x0.8x0.8x0.8x0.8 = 0.8^5 =0.32768

  - Probabilité d'atteindre le but accidentellement avec les mêmes actions, mais l'exécution ne produisant pas toujours l'effet désiré (ici, le résultat est [E,E,N,N,E])

    P_2 = 0.1x0.1x0.8x 0.1x0.1x0.8 = 0.1^4 x 0.8 = 0.00008

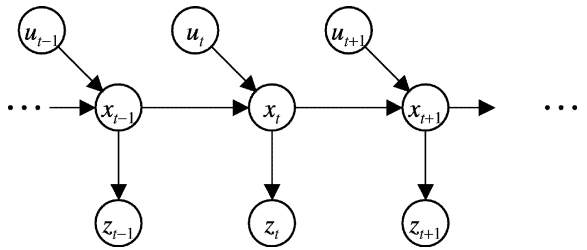


  - Probabilité totale d'atteindre le but avec cette séquence d'actions: P_1+P_2= **0.32776**

Modèle de Transition

- Spécification des probabilités du résultat de chaque action, dans chaque état possible.
- Tableau *T(S, A, S')* contenant les probabilités $P(s' \mid s, a)$ d'atteindre un état *s'* si une action *a* est exécutée dans l'état *s*.
- Hypothèse de Markov: la probabilité conditionnelle d'un état ne dépend que de l'état qui le précède et de l'action exécutée.

Hypothèse de Markov

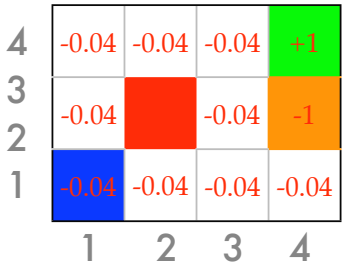


$$p(z_t \mid x_{0:t}, z_{1:t}, u_{1:t}) = p(z_t \mid x_t)$$
$$p(x_t \mid x_{1:t-1}, z_{1:t}, u_{1:t}) = p(x_t \mid x_{t-1}, u_t)$$

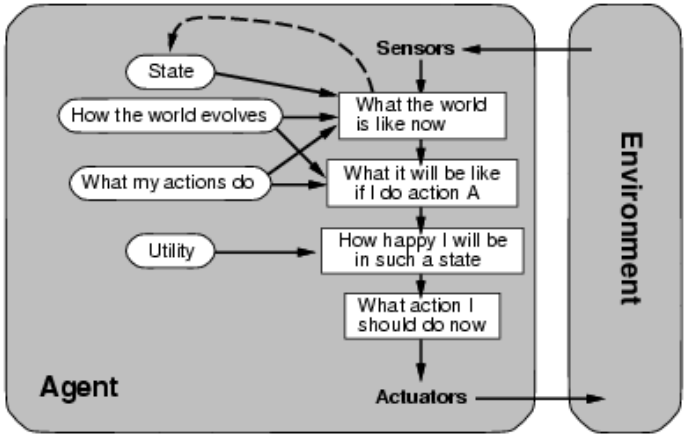
L'état x_t ne dépend que de l'état x_{t-1} et de la dernière transition u_t

Notion de Récompense

- Nombre réel
  - Associée à l'état atteint uniquement : R(S)
  - Associée à l'état et l'action: R(S,A)
  - Associée à l'état, l'action et l'état but final: R(S,A,J)
- Exemple:
  - *R(s) = -0.04* pour tout *s* sauf:
  - *R(4,3) = +1*
  - *R(4,2) = -1*

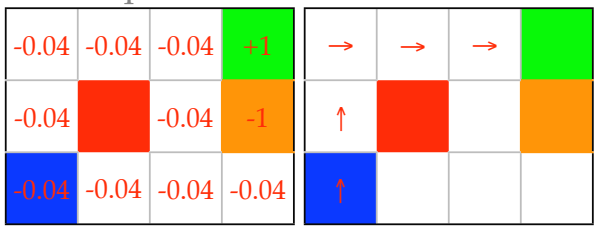


Agent orienté utilité



Notion d'Utilité

- Atteindre le but est un problème de décision *séquentiel*.
- La *fonction d'utilité* est la **somme des récompenses** reçues. Dépend de la séquence d'états visités.
- Dans l'exemple:



  - Si l'état but (4,3) est atteint après 5 actions [N,N,E,E,E] l'utilité totale U=4x(-0.04)+1 = 0,84

Utilité et Politique

- Utilité évaluée pour chaque état. Exprime la contribution d'un état donné pour l'exécution de la tâche globale.
- Une *politique* établit un lien entre actions et états. Guide le choix de l'action à exécuter dans un état donné:

Politique: Etat →Action dans le but de maximiser l'utilité

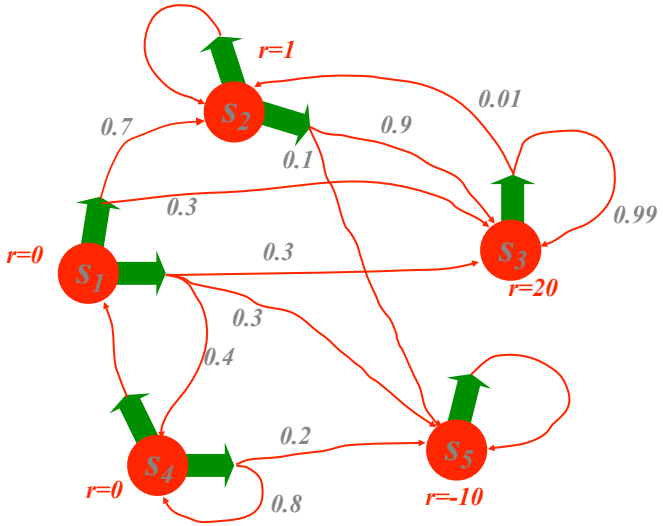
Processus Décisionnel de Markov

- Spécification d'un problème de décision séquentiel dans un environnement entièrement observable qui satisfait l'hypothèse de Markov et dans lequel les récompenses sont additives.
- Formellement défini comme un tuple: $<S, A, T, R>$
  - S : ensemble fini d'états du monde.
  - A : ensemble fini d'actions.
  - T *ou* P: $S \times A \rightarrow S$: fonction de transition d'état. Probabilité qu'une action change l'état: $T_a(s, s') = Pr(s_{t+1} = s' \mid s_t = s, a_t = a)$
  - T fournie sous la forme d'une table de probabilités de transition.
  - R: $S \times A \rightarrow \Re$: récompense reçue après exécution d'une action donnée a aboutissant à l'état s : $R(s, a)$.

Rq: Problème général

- Si la transition entre états ne comporte pas un choix *d'action* mais uniquement une distribution de probabilités, et pas de notion de récompense: *chaîne de Markov.*
- Si $<S, A, T, R>$ sont connus: *MDP*
- Si l'état est partiellement observable (probabilité de se trouver dans un état donné): *POMDP*
- Si les probabilités de transition sont inconnues: *apprentissage par renforcement.*

Représentation graphique



Dans notre exemple…

- S: Position de l'agent sur la grille
  - Cellule (x,y) ; P. ex. (4,3)
- A: Actions de l'agent
  - N,W,S,E
- P: Fonction de Transition. Table P(s'| s, a), proba de s' sachant a et s
  - Ex: **P((4,3) | (3,3), N) = 0.1**
  - Ex: **P((3,2) | (3,3), N) = 0.8**
- R: Récompense. Ex: R(3, 3) = -0.04; R (4, 3) = +1

Résolution d'un MDP

- Dans un processus déterministe, la solution est un **plan**.
- Dans un processus observable stochastique, la solution est une **politique**
- **Politique:** fonction de S dans A. L'agent doit décider d'une politique
- La qualité d'une politique est mesurée par son utilité totale attendue.

Notation:

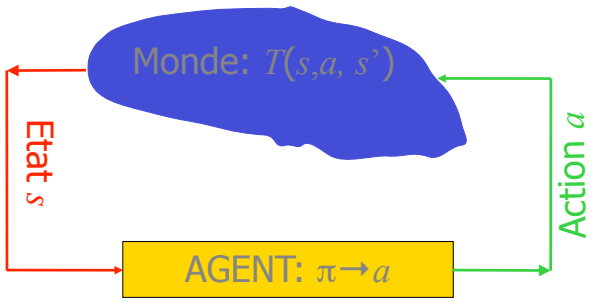
π Politique

$\pi(s)$ Action recommandée par la politique dans l'état s

π^* Politique optimale (Utilité attendue maximale)

Politique

- MDP:



Politique

- Une Politique est un choix a priori d'une séquence d'actions.
- C'est une indication de l'action à exécuter dans chaque état pour atteindre un état final.
- Le choix des actions est guidé par la maximisation d'une récompense globale, et non par la simple atteinte du but.
- Une politique **n'est pas** un plan d'action qui doit être exécuté strictement; elle peut être poursuivie malgré des actions non réussies.

Objectif d'un MDP

- Trouver la politique optimale π liant les états S aux actions A pour maximiser une fonction de valeur (récompense totale, ou utilité) $U(s)$.

Politique Optimale

$$\pi^*(s) = argmax_a \sum_{s'} T(s, a, s')U(s')$$

$T(s, a, s')$ = Probabilité d'atteindre un état s' à partir de l'état s

$U(s')$ = Utilité de l'état s'.

- Si l'utilité est connue, la politique optimale peut être calculée.
- Comment calculer l'utilité de chaque état sachant que la récompense dépend des voisins?

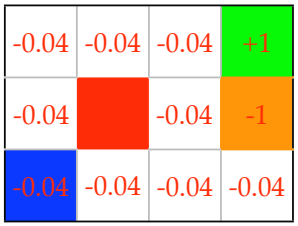
Valeur d'Utilité des Séquences

- Récompense Additive: l'utilité est la somme des récompenses.
  - $U_h([s_0, s_1, s_2, \ldots]) = R(s_0) + R(s_1) + R(s_2) + \ldots$
- Facteur d'escompte: Les récompenses à venir comptent moins.
  - $U_h([s_0, s_1, s_2, \ldots]) = R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \ldots$
  - $0 \leq \gamma \leq 1$
  - L'effet de l'utilité d'une action décroît avec l'horizon

Récompense additive

- Horizon infini: l'utilité d'une séquence infinie devient infinie
- Avec une récompense escomptée, l'utilité d'une séquence infinie reste finie.
  - $U_h([s_0, s_1, s_2, \ldots]) = \sum_{t=0}^{\infty} \gamma^t R(s_t) \leq \sum_{t=0}^{\infty} \gamma^t R_{\max} = R_{\max}/(1-\gamma)$

Exemple



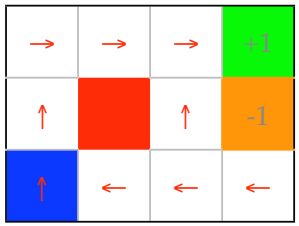
- Si l'agent est dans (1,3), quelle action effectuer?
  - Horizon fini T=3 : action N
  - Horizon infini: dépend des autres paramètres

Politique Optimale (exemple)

R(4,3) = +1

R(4,2) = -1

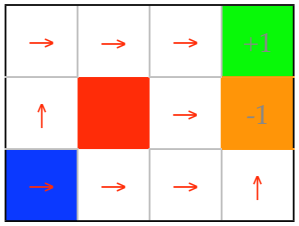
R(S) = -0.04



La récompense de chaque état est plus élevée que la case "-1".

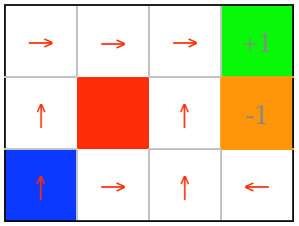
Préférence de contourner la case "-1"pour éviter d'y tomber.

Politiques Optimales (exemple)



R(S) < -1.6284

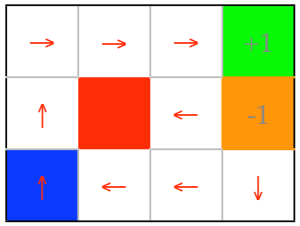
La récompense est très faible dans les états. L'agent préfère prendre l'action vers la "sortie" la plus proche.

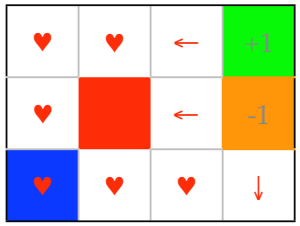


-0.4278 < R(S) < -1.6284

La récompense est comparable à celle de l'état à éviter. L'agent prend le risque de tomber dans -1 en voulant atteindre +1.

Politiques Optimales (exemple)





-0.0221 < R(S) < 0

Récompense plus importante que -1. L'agent ne prend pas le risque d'aller vers -1.

R(s) > 0

Récompense dans tous les états. L'agent veut "rester dans le jeu".

Calcul de π^*

- Deux méthodes:
  - Itération de valeur
  - Itération de politique

1. Itération de Valeur

- Démarche:
  - Calculer itérativement l'utilité de chaque état
  - Utiliser les valeurs pour sélectionner une action optimale

Itération de valeur

- Valeurs initiales arbitraires U_0 (p. ex. celles des récompenses).
- Mise à jour de l'utilité de chaque état à partir de ses voisins:
  - $U_{i+1}(s) \leftarrow R(s) + \gamma \max_a \Sigma (T(s, a, s') U_i(s'))$
  - itération jusqu'à convergence

Un problème d'optimisation

- Problème d'optimisation séquentiel similaire à celui de la Programmation Dynamique
- Principe d'optimalité de Bellman:
  - Si une fonction f s'écrit $f(x, y) = g(x, h(y))$ alors:

$$\text{Optimum}_{x,y}(f(x,y)) = \text{Optimum}_x(g(x, \text{Optimum}_y h(y)))$$

  - Ici on cherche à maximiser l'utilité U(s) qui s'écrit comme fonction d'un état et de la transition vers ses voisins:

$$\mathbf{U(s) = R(s) + \gamma \max_a \sum_s (T(s, a, s') U(s'))}$$

Optimisation

- Equations non linéaires
- Approche itérative

Itération de valeur

ITERATION_Valeur (mdp)

Entrées: *mdp(S,A,T,R),γ,ε*

Initialiser $U_0(s)$ à $R(s)$ pour tout s

répéter

Pour chaque état $s \neq$ but **faire**

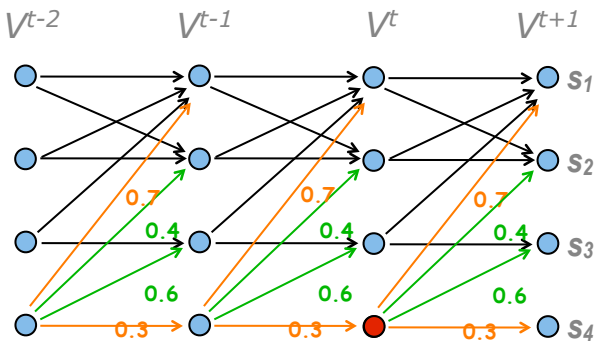
$U_{i+1}(s) \leftarrow R(s) + \gamma \max_a\{\Sigma_{s'} T(s,a,s')U_i(s')\}$

jusqu'à ($U_{i+1} \approx U_i$ "à ε près") [cf. convergence]

(ou jusqu'à N itérations - horizon fini).

Sortie: Utilités $U(s)$

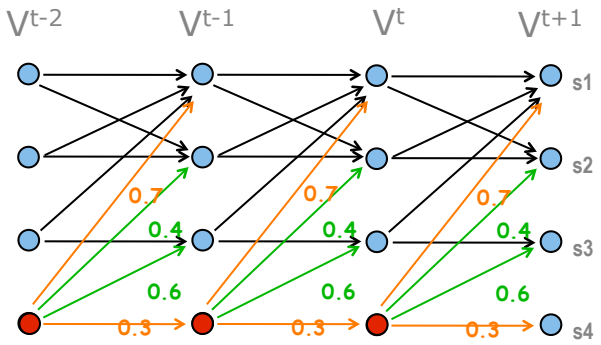
Itération de valeur



$V^t(s_4) = R(s_4) + \max\{\ 0.7\ V^{t+1}(s_1) + 0.3\ V^{t+1}(s_4)$

$0.4\ V^{t+1}(s_2) + 0.6\ V^{t+1}(s_3)\ \}$

Itération de valeur



$\Pi^t(s4)$ = max { ■ ■ }

Epoque ou horizon de Décision

- Horizon fini
  - Terminaison après un temps donné T.
  - $U_h([s_0, s_1, \ldots, s_{T+k}]) = U_h([s_0, s_1, \ldots, s_T])$, pour tout k>0
  - L'action optimale pour un état donné peut être différente.
- Horizon infini
  - Pas de fin du calcul a priori (ε).
  - Pas de raison de choisir des actions différentes dans un état donné à des instants différents.

Convergence de l'itération de valeur

- Convergence démontrée
- Solution unique à la convergence.

- Rq: Les valeurs exactes ne sont pas nécessaires. Il suffit d'être assez près.

Exemple

Environnement simplifié

2	R=-0.04	R=+1
1	R=-0.04	R=-1
	1	2

- Etat initial (1,1)
- Etats "buts": (2,1), (2,2)
- $\gamma = 1$
- Actions N,W,S,E.
- $U(1,1) = R(1,1) + \gamma \max_a \{0.8U(1,2) + 0.1U(1,1) + 0.1U(2,1), 0.9U(1,1) + 0.1U(1,2), 0.9U(1,1) + 0.1U(2,1), 0.8U(2,1) + 0.1U(1,1) + 0.1U(1,2)\}$