Décision dans l'incertain

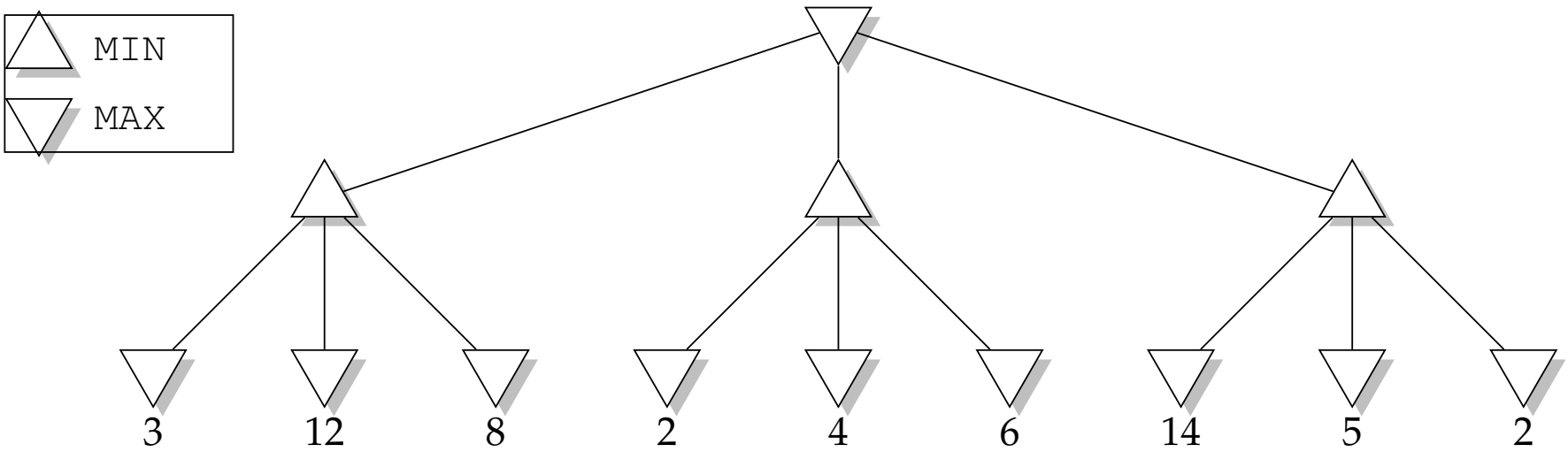
Cours 8: Décisions séquentielles

Stéphane Airiau

Université Paris-Dauphine

Retour sur jeux à deux joueurs

Dans le cours d'IA, on vous a présenté l'algorithme `minimax` pour jouer à des jeux à deux joueurs, avec information complète.

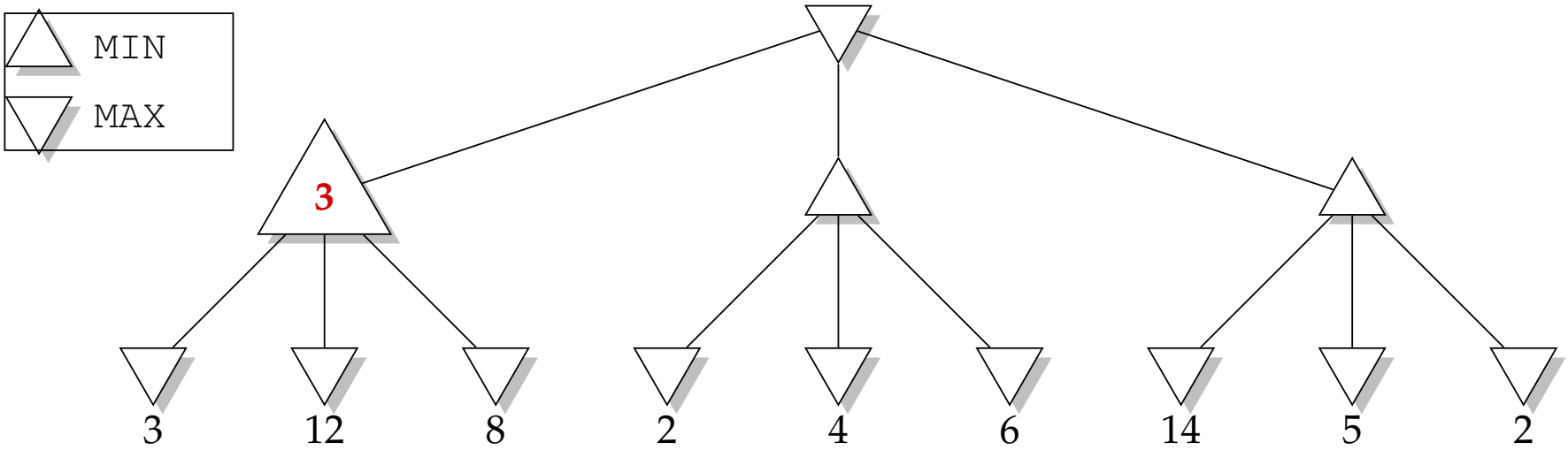


On raisonne sur les coups possibles de l'adversaire et on choisit son premier coup en fonction.

Que se passet-t-il si l'adversaire est "la nature" et qu'on ne sait pas précisemment comment elle joue ?

Retour sur jeux à deux joueurs

Dans le cours d'IA, on vous a présenté l'algorithme `minimax` pour jouer à des jeux à deux joueurs, avec information complète.

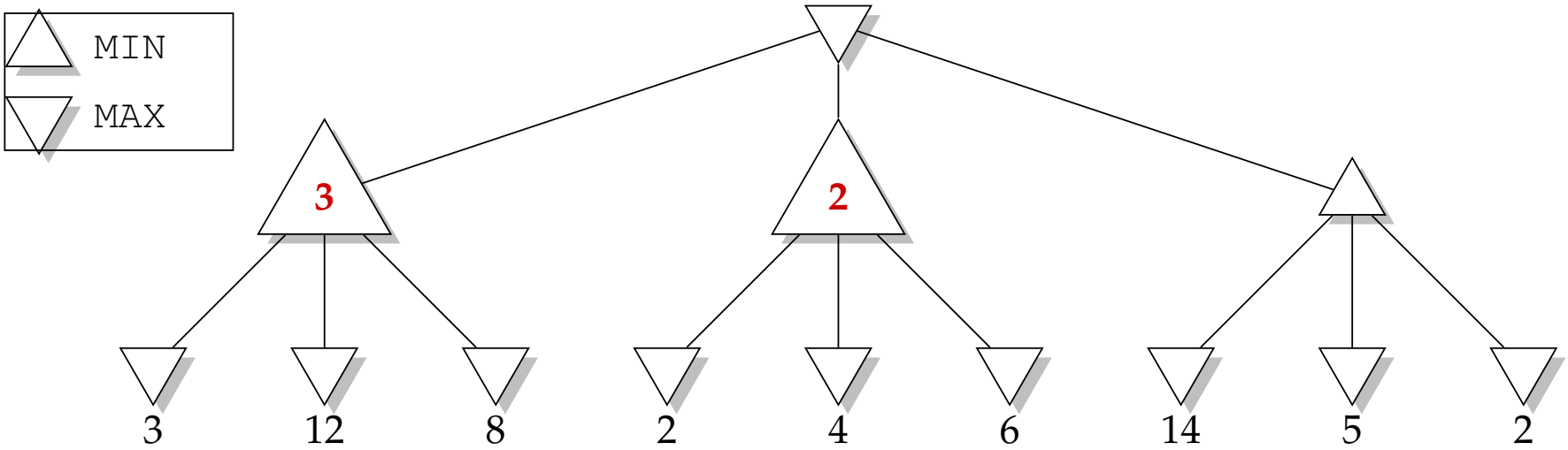


On raisonne sur les coups possibles de l'adversaire et on choisit son premier coup en fonction.

Que se passet-t-il si l'adversaire est "la nature" et qu'on ne sait pas précisemment comment elle joue ?

Retour sur jeux à deux joueurs

Dans le cours d'IA, on vous a présenté l'algorithme `minimax` pour jouer à des jeux à deux joueurs, avec information complète.

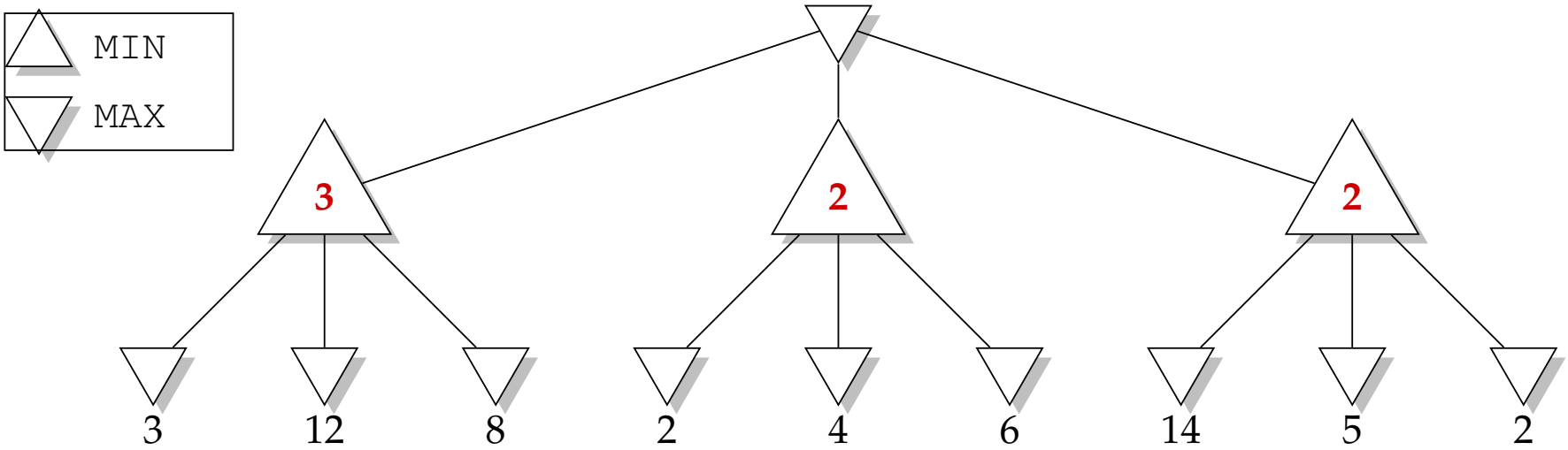


On raisonne sur les coups possibles de l'adversaire et on choisit son premier coup en fonction.

Que se passet-t-il si l'adversaire est "la nature" et qu'on ne sait pas précisemment comment elle joue ?

Retour sur jeux à deux joueurs

Dans le cours d'IA, on vous a présenté l'algorithme `minimax` pour jouer à des jeux à deux joueurs, avec information complète.

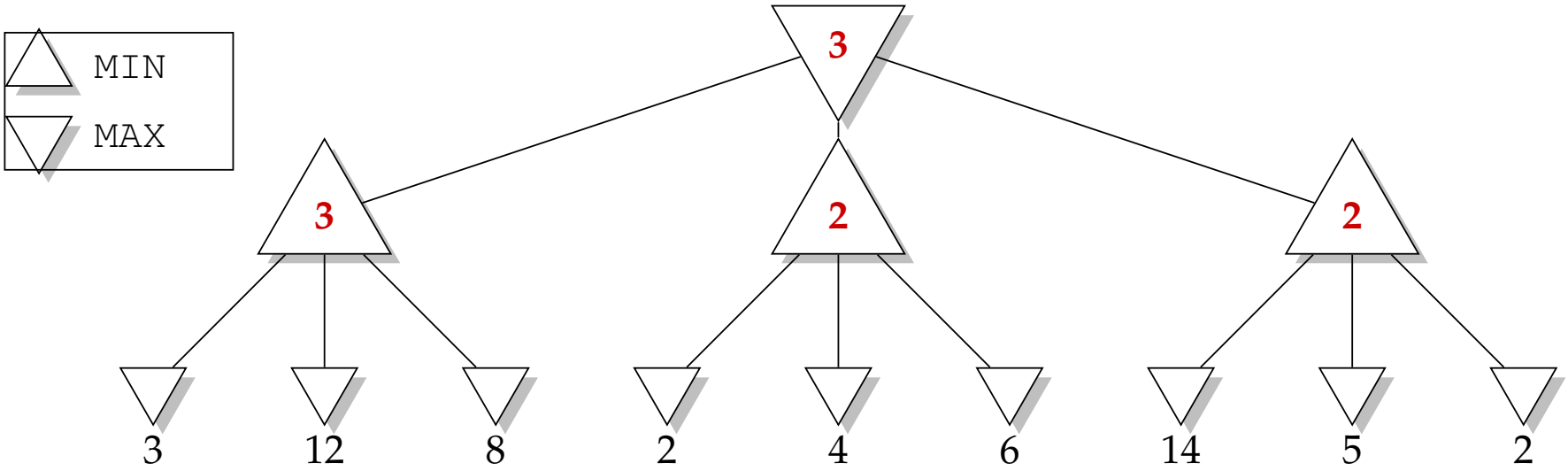


On raisonne sur les coups possibles de l'adversaire et on choisit son premier coup en fonction.

Que se passet-t-il si l'adversaire est "la nature" et qu'on ne sait pas précisemment comment elle joue ?

Retour sur jeux à deux joueurs

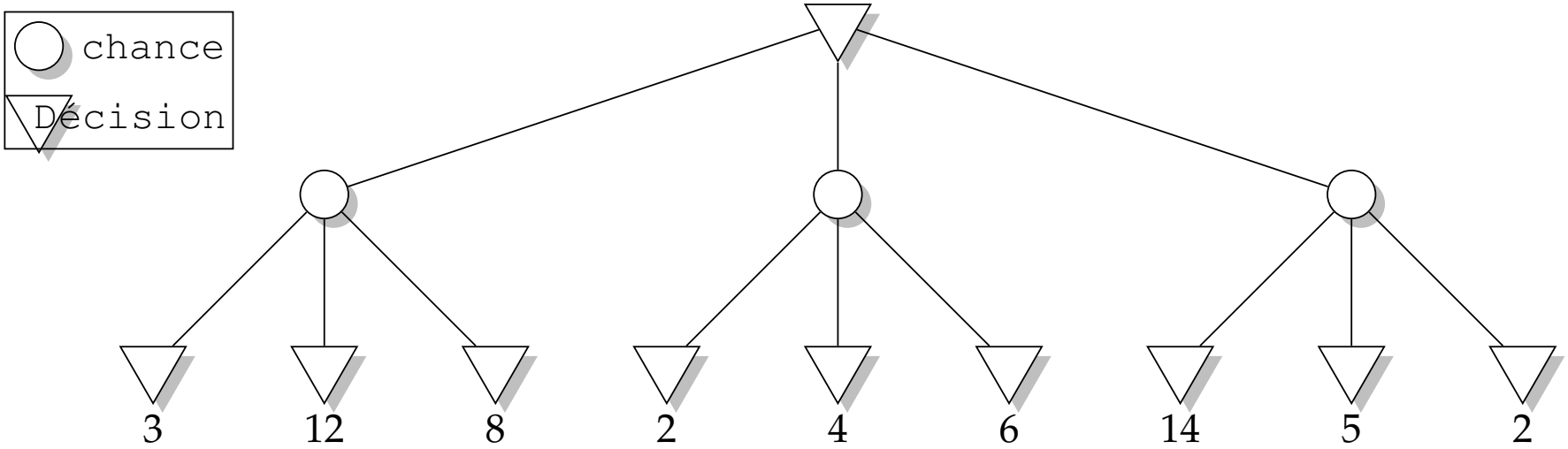
Dans le cours d'IA, on vous a présenté l'algorithme `minimax` pour jouer à des jeux à deux joueurs, avec information complète.



On raisonne sur les coups possibles de l'adversaire et on choisit son premier coup en fonction.

Que se passet-t-il si l'adversaire est "la nature" et qu'on ne sait pas précisemment comment elle joue ?

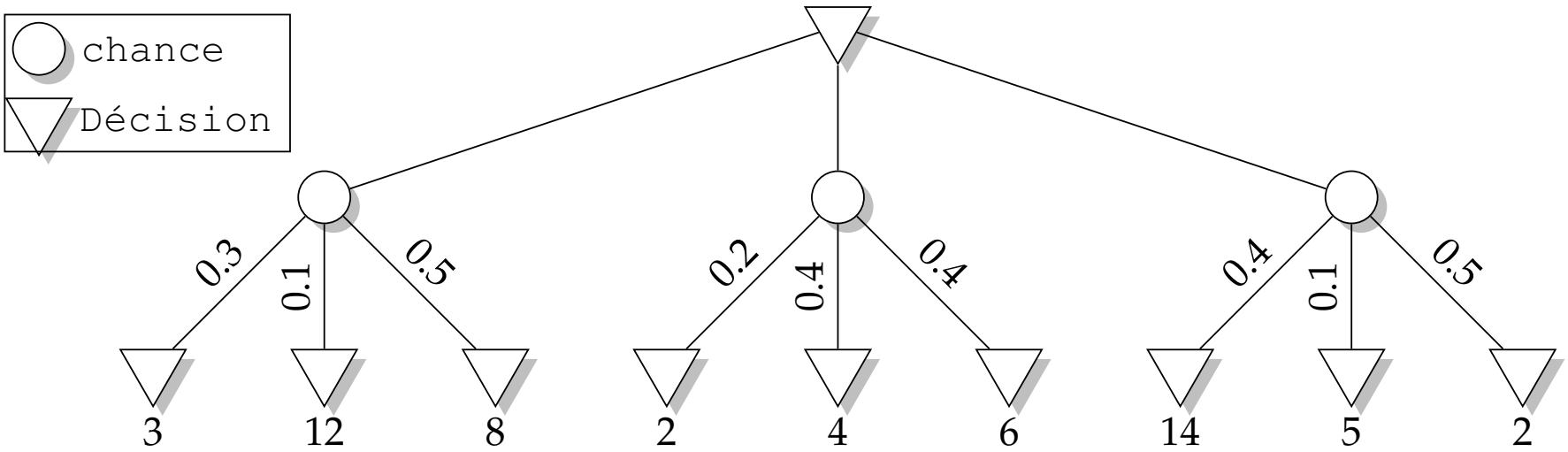
Arbres de décision



Comment jouer maintenant ?

- on considère que la nature est un joueur MIN ⇨ approche dans le pire des cas. ⇨ on choisit la meilleure action dans le pire des cas. on reverra ce problème dans avec des aspects orientés "théorie de la décision" plus tard.
- si on a des probabilités, on peut choisir de prendre l'action qui donne la meilleure espérance.

Arbres de décision



On va utiliser EXPECTIMAX

- pour les noeuds `décision` ⇨ comme les noeuds MAX dans maxmin
- pour les noeuds `chance` ⇨ on calcule la valeur moyenne des valeurs des enfants.

Algorithmes

```
1 function EXPECTIMAX(s) returns an action
2 arg max_{a∈actions(s)} EXPECTEDVAL(result(s,a))
```

```
1 function EXPECTEDVAL (s) returns a utility value
2     if terminal?(s) then return utility(s)
3     v ← 0
4     for each n ∈ next(s) do
5         v ← v + ℙ(n) × EXPECTIMAX(n)
6     return v
```

élagage pour la recherche expectimax ?

A part les valeurs des feuilles, on n'observera pas (peu) les valeurs des noeuds pour une partie donnée !

élagage pour la recherche expectimax ?

A part les valeurs des feuilles, on n'observera pas (peu) les valeurs des noeuds pour une partie donnée !

Si on une borne sur les valeurs des feuilles, on peut peut-être faire de l'élagage, mais sinon NON !

élagage pour la recherche expectimax ?

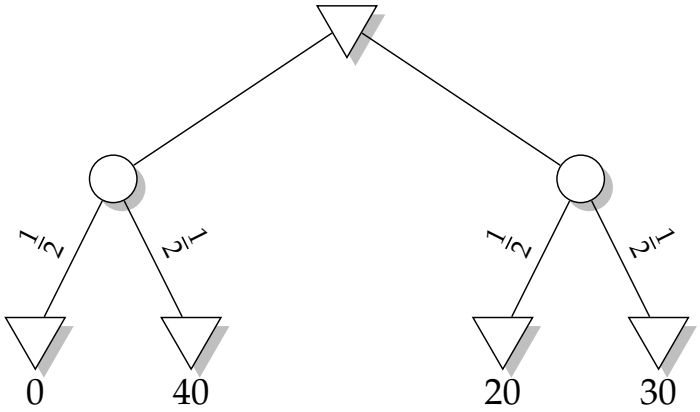
A part les valeurs des feuilles, on n'observera pas (peu) les valeurs des noeuds pour une partie donnée !

Si on une borne sur les valeurs des feuilles, on peut peut-être faire de l'élagage, mais sinon NON !

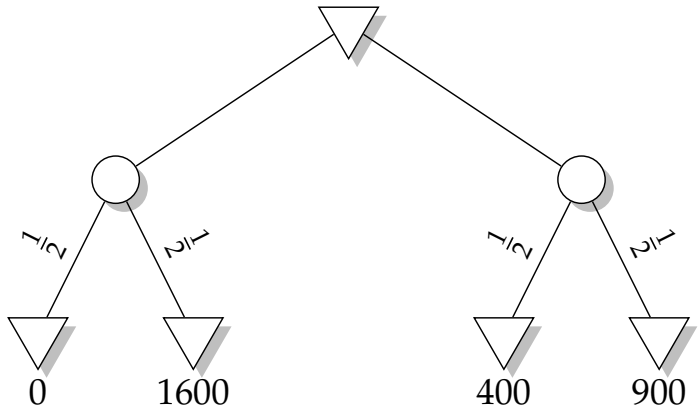
On peut utiliser depth limited expectimax avec des fonctions d'évaluation si on n'a pas le temps d'étudier tout l'arbre

Danger des échelles d'utilités

Pour minimax, l'échelle des valeurs des feuilles n'était pas importante, tant que l'ordre entre les valeurs est conservé, le raisonnement sera correct.
pour expectimax, ce n'est pas le cas !



⇨ on prend le carré des valeurs



Récapitulatif

- si on ne connait pas les probabilités : pour le moment, pas mieux que MINIMAX ! approche pessimiste !
- si on connait les probabilités
  - l'approche pessimiste est toujours disponible
  - approche en moyenne
  - ⇨ cela dépend comment on modélise la situation !

Pourquoi est-ce la bonne chose d'utiliser la moyenne des utilités ?

- d'où viennent les utilités ?
- est-ce que ça existe vraiment ?
- calculer la moyenne a-t-il un sens ?
- est-ce qu'on peut définir la notion d'utilité de manière "scientifique"

Préférences

Soit A l'ensemble des actions potentielles (ou alternatives)
Une loterie décrit les actions potentielles et une probabilité associée
exemple :

- $A = \{a, b, c\}$
  - a : un cornet avec une boule de glace
  - b : un cornet avec deux boules de glace
  - c : un cornet avec zéro boule de glace
- loterie $L = [p : b, (1-p) : c]$: si je prend un cornet avec deux (grosses) boules de glace, avec une probabilité $p = 90\%$ je mange deux boules, mais avec une probabilité $1-p = 10\%$, les boules sont trop instables et tombent, je me retrouve sans glace !

On va mettre une notion d'ordre sur de telles loteries :
$L_1 \succ L_2$ je préfère strictement L_1 à L_2
$L_1 \sim L_2$ je suis indifférent entre L_1 et L_2

- je ne sais pas faire la différence entre L_1 et L_2

Rationalité

On peut essayer de formaliser le concept d'une préférence rationnelle.

méthode axiomatique : lister les axiomes qui semblent être souhaitables (i.e. si une préférence ne satisfait pas un axiome, on pourrait trouver cela étrange !)

Definition (axiome de transitivité)

Une préférence est transitive ssi $L_1 \succ L_2$ et $L_2 \succ L_3$ implique que $L_1 \succ L_3$.

Supposons qu'on ait trois alternatives a, b, et c et l'agent a les préférences qui ne sont pas transitives $a \succ b$, $b \succ c$, et $c \succ a$

- supposons que l'agent possède a
- puisque $c \succ a$, l'agent est content d'échanger c avec a et supposons qu'il soit même prêt à payer 1centime pour cet échange.
- puisque $b \succ c$, idem : échange b avec c et paiement 1 centime
- puisque $a \succ b$, idem : échange a avec b et paiement 1 centime

L'agent risque donc de dépenser une fortune !

Rationalité

- ordre complet, asymmétrique, transitif

vNM0 $A \succ A \Rightarrow B \not\succ A$ soit je préfère strictement A à B, soit l'inverse, mais évidemment pas les deux en même temps !

vNM1 $A \succ B$ ou $A \sim B$ ou $B \succ A$ on peut toujours dire quelque chose entre chaque paires (on ne peut pas dire je ne sais pas !)

vNM2 transitivité

vNM3 continuité pour chaque $A \succ B \succ C$, il existe $(p,q) \in]0,1[^2$ tels que $pA + (1-p)C \succ B \succ qA + (1-q)C$

Si A est gagner 2€pour sûr, B un €pour sûr et C ne rien gagner, on peut trouver p et q tels que je préfère gagner 2€avec une probabilité p à gagner 1€pour sûr et je préfère un €pour sûr à gagner 2 €avec une probabilité q

vNM4 indépendance $A \succ B$ ssi $pA + (1-p)C \succ pB + (1-p)C$ ajouter l'alternative C avec la même probabilité ne devrait rien changer à votre préférence

Théorème de von Neumann et Morgenstern

Theorem (vNM)

Une relation de préférence satisfait les axiomes vNM1-4 ssi il existe une fonction u qui prend une loterie et qui retourne un réel entre 0 et 1 avec les propriétés suivantes :

1– $A > B$ ssi $u(A) > u(B)$

2– $u(pA + (1-p)B) = p \cdot u(A) + (1-p) \cdot u(B)$

3– Pour toute autre fonction v qui satisfait (1) et (2), il existe $c > 0$ et $d \in \mathbb{R}$ tels que $v = c \cdot u + d$

les préférences "rationnelles" impliquent un comportement que l'on peut décrire comme une maximisation de l'utilité espérée.

On peut utiliser des axiomes légèrement différents mais équivalents aux axiomes vNM1-4.

utilités en pratique (pas en théorie)

Mettre en place des échelles d'utilité

- normalisation ? on peut toujours utiliser une transformation linéaire
- Un micromort (mot formé de « micro » et de « mortalité ») est une unité de risque égale au millionième d'une probabilité de décès.
- QALYS : Quality Adjusted Life Years : équilibrer le faire de vivre plus longtemps et le fait de bien vivre (ex ma grand mère a 98 ans mais vie sur un lit d'hopital depuis 5 ans)
- thème de recherche ex au lamsade : échelle des mesures des prisons dans les différents pays.

Rationalité humaine

a– 80% de chance d'avoir 4k€et 20% de chance d'avoir 0€.

b– avoir 3k€garanti (100%)

c– 20% de chance d'avoir 4k€et 80% de chance d'avoir 0€.

d– 25% de chance d'avoir 3k€et 75% de chance d'avoir 0€.

La plupart des personnes préfères $b \succ a$ et $c \succ d$
Cependant si $u(0e) = 0$

- $b \succ a \Rightarrow u(3k€) > 0.8u(4k€)$
- $c \succ d \Rightarrow 0.8u(4k€) > u(3k€)$

ceci est une contradiction ! ! ! ! !
Paradoxe de Allais.

Elicitation des préférences

demander à un agent des loteries comme "êtes vous prêt à parier k €ou participer à la loterie p la meilleure alternative, $1-p$ la pire alternative

⇨on peut faire varier p jusqu'à ce que l'agent soit indifférent entre les deux loteries
⇨ p est l'utilité normalisée.

en pratique, on s'aperçoit souvent que les humains n'utilisent pas toujours des préférences rationelles.

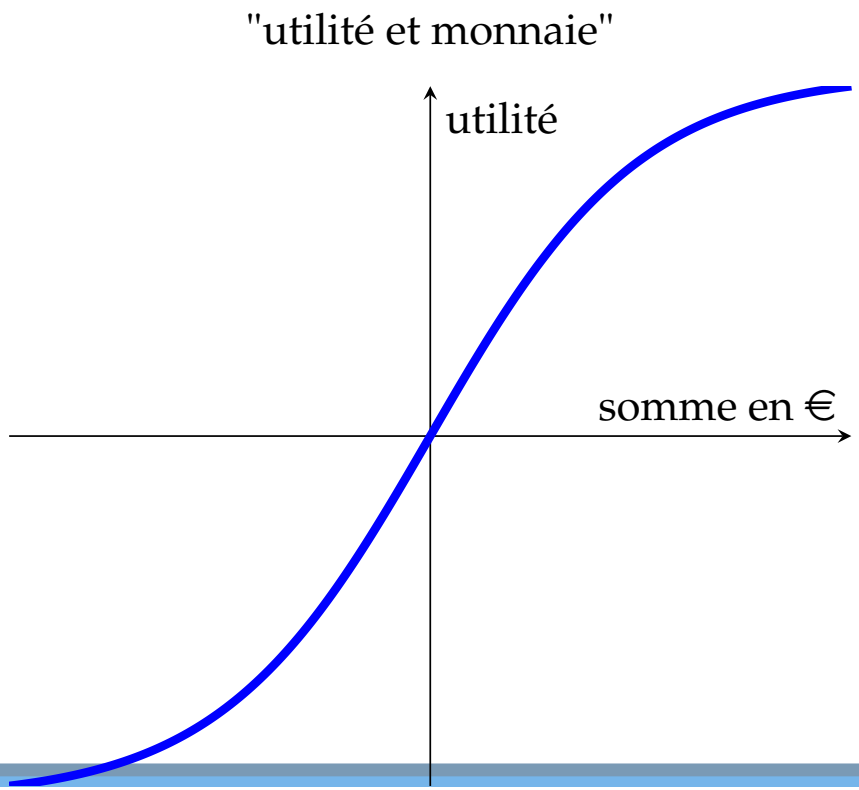
Si le domaine est combinatoire, pas facile d'obtenir des utilités.

- dans le sens combinaison de critères
- dans le sens où quelqu'un veut connaître la valeur d'un ensemble d'alternatives (ex configuration d'une voiture, d'un voyage, etc.)
- en plus, on ne veut pas poser trop de questions à l'agent...

Utilité et monnaie

La monnaie ne se comporte pas comme une utilité.

- les personnes sont souvent averses au risque : on préfère plus souvent la garantie d'avoir quelque chose plutôt que le risque de ne rien avoir.
- si une personne est vraiment endettée, on observe souvent que le comportement est plus risqué.



Processus de décision markovien

Cours 8: Décisions séquentielles

Stéphane Airiau

Université Paris-Dauphine

Généralisation de expectimax

Avec les arbres de décisions, on a vu un premier type de décision séquentielle dans l'incertain.

On va maintenant voir une généralisation.

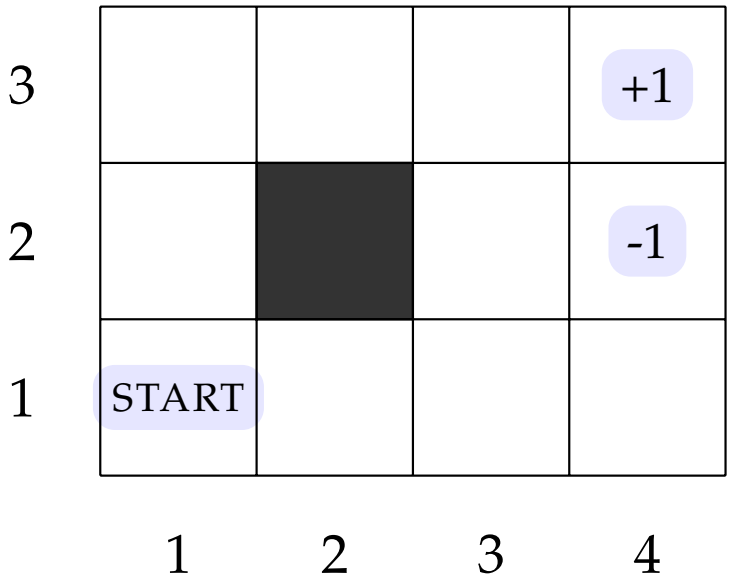
Exemple canonique : un robot se déplace dans une grille.

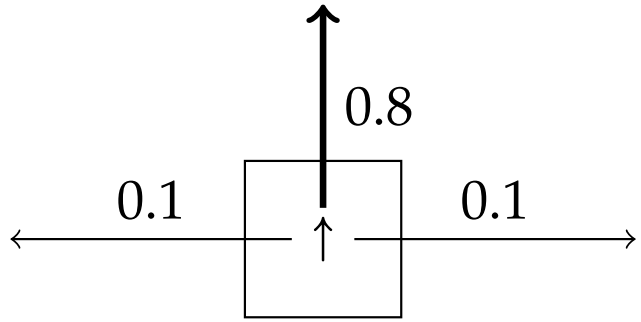
- Les murs bloquent le déplacement de l'agent. Il ne connait pas a priori la position de tous les murs
- Il peut tomber dans des trous, il ne connait pas la position de chaque trou ou s'ils existent
- le robot est un peu plus réaliste : ses actions peuvent échouer (les roues peuvent patiner, ou non)
  - on estime que l'action D mène bien le robot dans la case au D,
  - mais dans 10% des cas, il dévie vers la gauche
  - mais dans 10% des cas, il dévie vers la droite
  - Si $D = Nord$, il a 80% de chance d'aller dans la case au nord, 10% d'aller dans la case à l'est, 10% d'aller dans la case à l'est.

Exemple : gridworld

- l'agent paie une pénalité pour chaque déplacement (il dépense de l'energie)
- certaines cases peuvent contenir une grosse récompense

cf Opportunity et Curiosity sur Mars...





exemple action vers le haut

Definition (Processus décisionnel de Markov)

Un *Processus décisionnel de Markov* est un tuple $\langle S, A, T, R, \gamma \rangle$ où

- S est un ensemble fini d'états
- A est un ensemble fini d'actions
- T est une matrice de transition
  $T^a_{ss'} = \mathbb{P}\left[S_{t+1} = s' \mid S_t = s, A_t = a\right]$ probabilité d'arriver dans l'état s' à l'instant +1 quand on a pris l'action a dans l'état s à l'instant t
- R est le vecteur de récompenses
  $R^a_s = \mathbb{E}\left[R_{t+1} \mid S_t = s, A_t = a\right]$ valeur moyenne obtenue après avoir pris l'action a dans l'état s
- un ensemble d'état initial
- parfois un ensemble d'états terminaux

Hypothèse de Markov

Si on exécute l'action a dans l'état s :

- On obtient une récompense r
- On arrive dans un état s'

En principe, r et s' peuvent dépendre de tout l'historique !

Definition (Etat de Markov)

Un état S_t est dit de **Markov** ssi

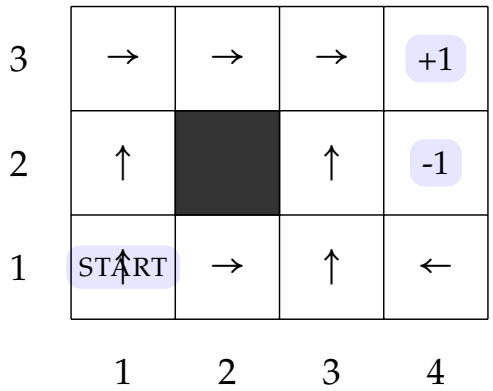
$$\mathbb{P}[S_{t+1}|S_t] = \mathbb{P}[S_{t+1}|S_1, \ldots, S_t]$$

- "Etant donné le présent, le futur est indépendant du passé"
- Une fois que l'état est connu, on peut effacer son historique !
- *ex :* aux échecs, l'état du jeu ne dépend pas de l'historique des coups !

Les composants d'un agent : politique

C'est ce qui gouverne le comportement de l'agent

- **politique déterministe** : La fonction associe à chaque état **une action** $\pi : S \mapsto A$



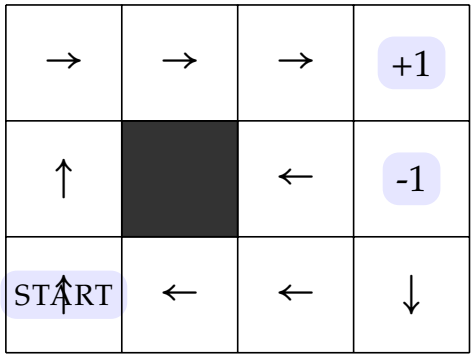
politique optimale pour un pénalité de 0.03 par déplacement

- **politique stochastique** : une distribution de probabilité sur les actions possibles

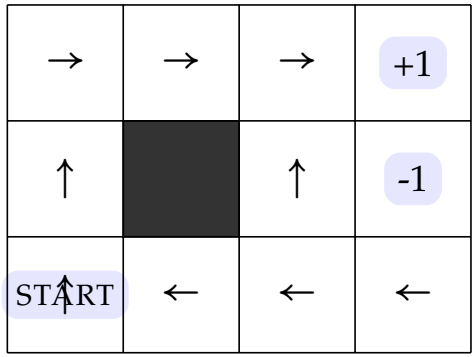
$$\pi : S \mapsto \Delta(A)$$

où $\Delta(N)$ désigne une distribution de probabilité sur l'ensemble (fini) N.
$p \in \Delta(N)$ ssi $\forall i \in N, p(i) \in [0,1]$ et $\sum_{i \in N} p(i) = 1$

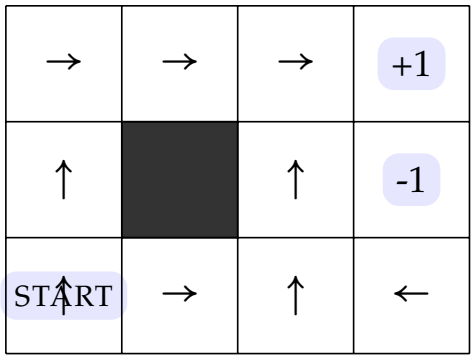
Politiques optimale



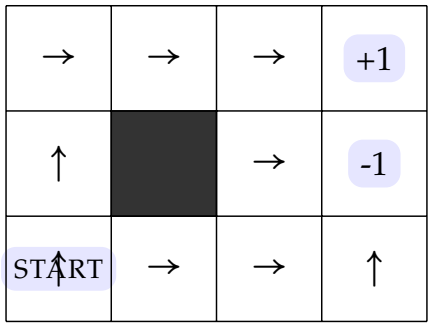
pénalité de 0.01 par déplacement
aucun risque : on fait le tour!



pénalité de 0.02 par déplacement
petit risque



pénalité de 0.04 par déplacement
on prend le chemin le plus court

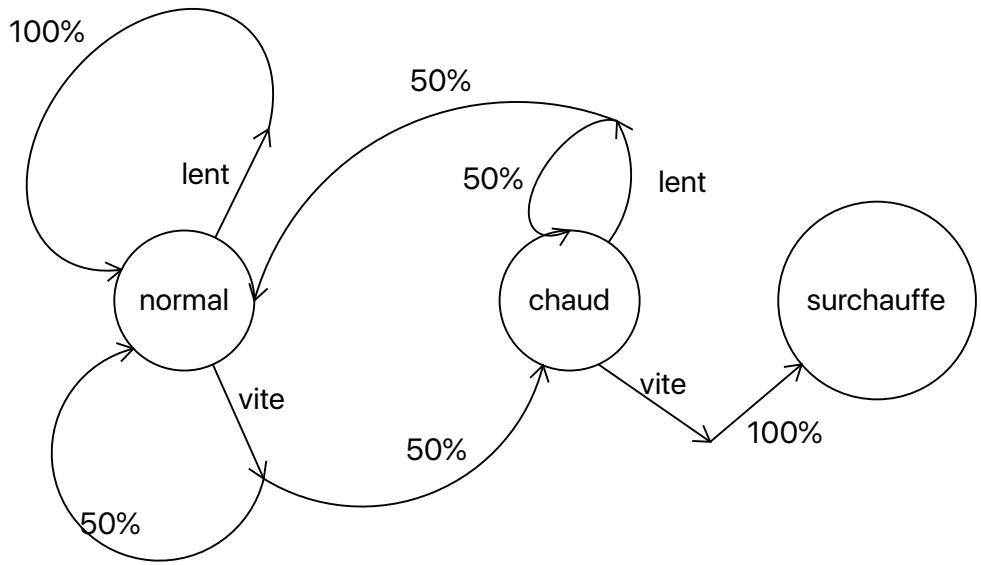


pénalité de 2 par déplacement
on prend des risques pour terminer au plus vite

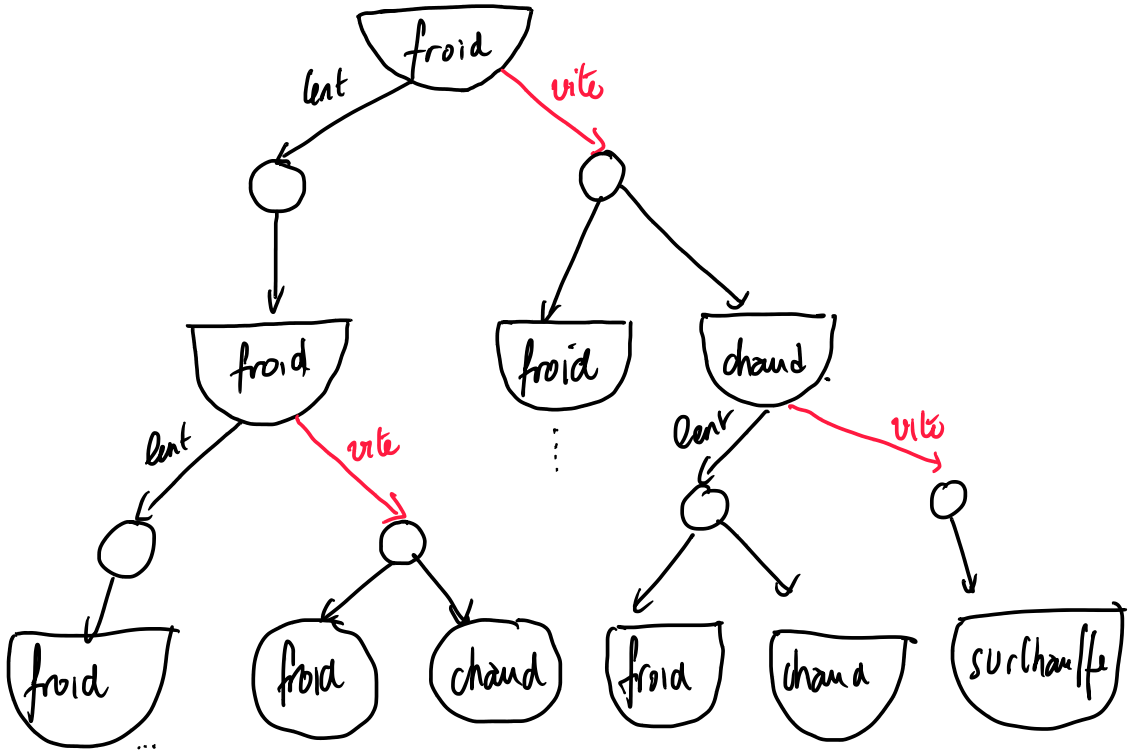
Exemple de la voiture de course

un robot voiture veut aller vite et loin !

- trois état du moteur : normal, chaud, surchauffe
- deux actions : lent, vite
- en allant plus vite, on double la récompense



Exemple de la voiture de course & expectimax



Quel est notre but

On va obtenir une séquence de valeurs d'utilité. Que préfère-t-on ?

- maintenant, plus tard ? (ex $\langle 0,0,10 \rangle$ ou $\langle 8,2,0 \rangle$)
  vous préférez 10 dans 3 jours, ou 8 aujourd'hui, 2 demain, et 0 dans 3 jours ?
- généralement, on préfère avoir plus d'utilité en tout, mais une distribution plus régulière peut être appréciable !

Dans certains problèmes où il y a une **fin**, on pourra chercher à maximiser la somme des utilités

Dans les autres cas, on peut utiliser un taux d'escompte γ :

- avoir 1 aujourd'hui vaut 1 aujourd'hui !
- avoir 0 aujourd'hui et 1 demain vaut aujourd'hui γ aujourd'hui
- avoir 0 aujourd'hui, 0 demain et 1 dans deux jours vaut γ^2 aujourd'hui
- $\gamma = 0$ l'agent est "myope" : il n'est intéressé que par la récompense immédiate
- $0 < \gamma < 1$ l'agent cherche un équilibre entre la récompense immédiate et celle qu'il obtiendra dans le futur

Quel est notre but : version formelle

- pour des tâches épisodiques
  - il y a des états terminaux et initiaux
  - on repart dans un état initial une fois qu'on atteint un état terminal

  ⇨ maximise le cumul des récompenses sur *un épisode* (ici de longueur T)

$$G_T = r_1 + r_2 + \cdots + r_T$$

- pour des tâches en continue

  ⇨maximise une récompense "escomptée" $G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$

  γ est le taux d'escompte
  - $\gamma = 0$ l'agent est "myope" : il n'est intéressé que par la récompense immédiate
  - $0 < \gamma < 1$ quand $\{r_t, t \in \mathbb{N}\}$ est bornée, alors R_T est bien définie.
    ⇨l'agent cherche un équilibre entre la récompense immédiate et celle qu'il obtiendra dans le futur

- maximiser une récompense "en moyenne" $R_t = \frac{1}{t} \cdot \sum_{k=0}^{\infty} r_{t+k+1}$

Focus pour le cours

- problèmes itératifs en continue.
- objectif : maximiser la somme "avec dévaluation" $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$
  - Pour éviter une récompense infinie si on tombe dans des cycles
  - Le futur reste uncertain ! Bon compromis entre court et long terme
  - Tendance naturelle vers le court terme
  - Mathématiquement, c'est quand même pratique !
- la fonction de transition est stochastique
- la fonction de récompense est connue

Comment trouver la meilleure politique ? / Comment calculer la fonction de valeur ?