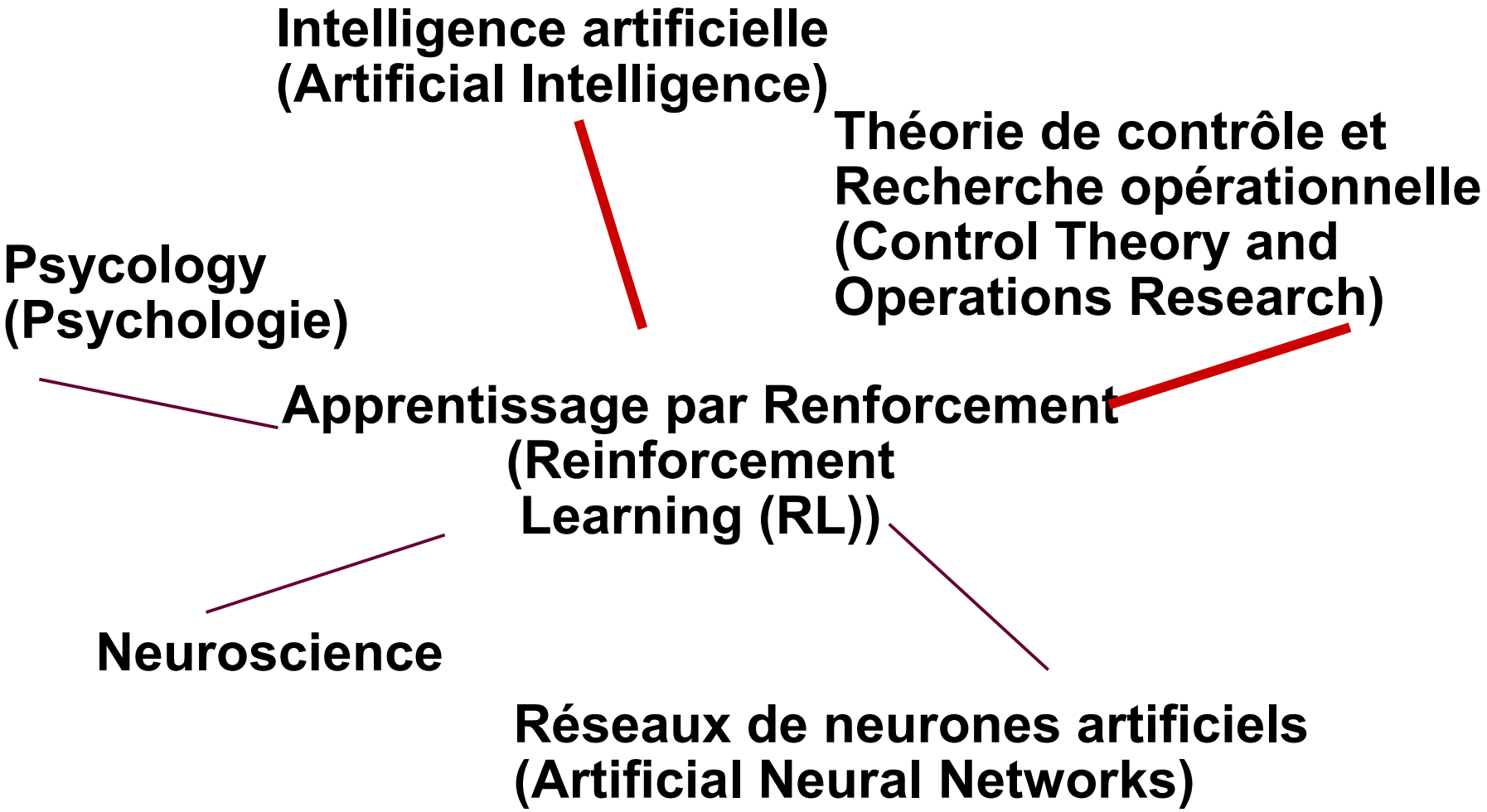
Apprentissage par renforcement (Reinforcement Learning (RL)) Chapitre 1:Introduction

T. AL-ANI
A²SI-ESIEE-PARIS

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Chapitre 1: Introduction



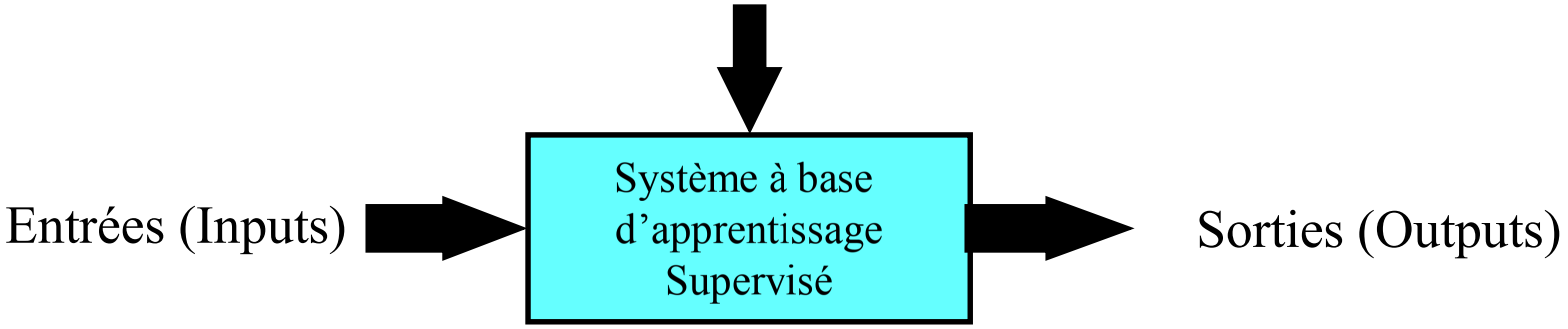
Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Que veut dire apprentissage par Renforcement?

- Apprentissage par l'**interaction;**
- Apprentissage type **orienté-objectif-** (Goal-oriented learning);
- Apprentissage basé sur l'interaction avec l'environnement externe (apprendre sur, de, et pendant l'interaction avec un environnement externe);
- Apprendre quoi faire – comment associer des situations à des actions pour qu'un signal de récompense numérique soit maximisé.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Apprentissage Supervisé (Supervised Learning (SL))



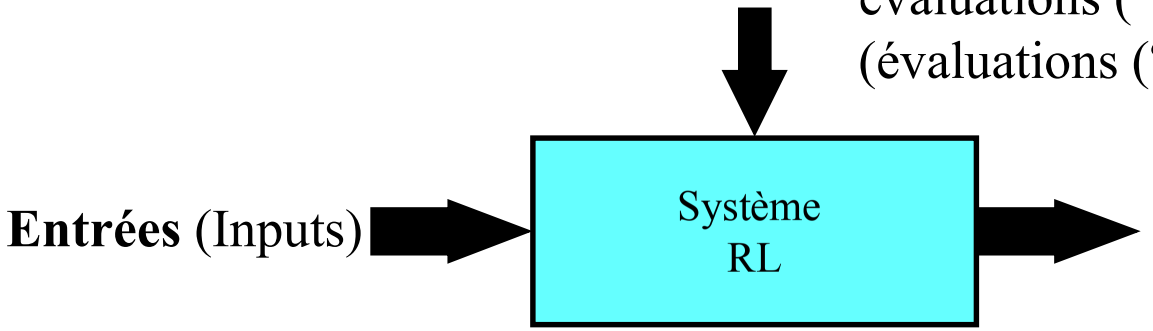
Erreur (Error) = cible ou consigne (target output) – sortie réelle (actual output)

Dans des problèmes interactifs il est souvent pas pratique d'obtenir des exemples des comportements désirés qui sont à la fois corrects et représentatifs de toutes les situations dans les quelles l'agent doit agir.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Apprentissage par Renforcement (Reinforcement Learning)

nformation d'apprentissage (Training Info) =
évaluations ("**récompenses**/"**pénalités**")
(évaluations ("rewards" / "penalties"))



Objectif: obtenir une récompense maximale.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Apprentissage par Renforcement et Apprentissage Supervisé

L'apprentissage supervisé peut être intégré dans l'apprentissage par renforcement pour des raisons très spécifiques qui détermines quels aptitudes sont critiques, et quelles aptitudes ne les sont pas.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Apprentissage par Renforcement et Planification

Quand l'apprentissage par renforcement inclut une planification, il doit s'adresser à l'interaction entre la planification et la sélection de l'action en temps réel aussi bien qu'à la question de comment les modèles de l'environnement sont acquis et améliorés.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Caractéristique clefs de RL

- L'**agent** (agent) ou le **système** (system) apprend lui même l'action nécessaire pour arriver à la cible.

- Recherche basée sur **Essai-Erreur** (Trial-and-Error).

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Caractéristique clefs de RL

- Possibilité d'une **récompense retardée** (**Valeur** (value))**V** (delayed reward)
  - Sacrifier des **gains à court terme** (short-term gains) pour des **gains plus grands à long terme** (long-term gains)

La **récompense r** ressemble à notre joie (si r est grande) et notre douleur (si elle est petite) dans l'immédiat tandis que les valeurs V correspondent à un jugement plus raffiné et une vue à long terme de comment heureux ou malheureux nous sommes quand notre environnement sera dans un état particulier.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Caractéristique clefs de RL

- Besoin d'**explorer** (***explore)*** et d'**exploiter** (***exploit)***

Challenges : compromis d'**explorer/exploiter**

Pour augmenter la récompense, l'agent doit préférer des actions qui ont étés essayés dans le passé et trouvées d'être effectives pour produire une récompense. Mais pour découvrir des telles actions il doit essayer des actions qui n'ont pas encore étaient sélectionnées.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Caractéristique clefs de RL

- Besoin d'**explorer** (***explore)*** et d'**exploiter** (***exploit)*** *(suite)*

Challenges (suite)

L'agent doit **exploiter** ce qui est déjà connu pour obtenir une récompense, mais il doit aussi **explorer** pour sélectionner les meilleurs actions dans le future.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Caractéristique clefs de RL

- Besoin d'**explorer (*explore)*** et d'**exploiter (*exploit)*** *(suite)*

  Dilemme *: ni exploitation ni exploration peut être poursuivi exclusivement sans échouer dans la tâche. L'agent doit essayer une variété d'actions et favoriser progressivement celles qui apparaissent les meilleurs.*

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Caractéristique clefs de RL

- Besoin d'**explorer** (***explore)*** et d'**exploiter** (***exploit)*** *(suite)*

  Sur le plan stochastique, chaque action doit être essayée plusieurs fois pour estimer efficacement ses récompenses espérées.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Caractéristique clefs de RL

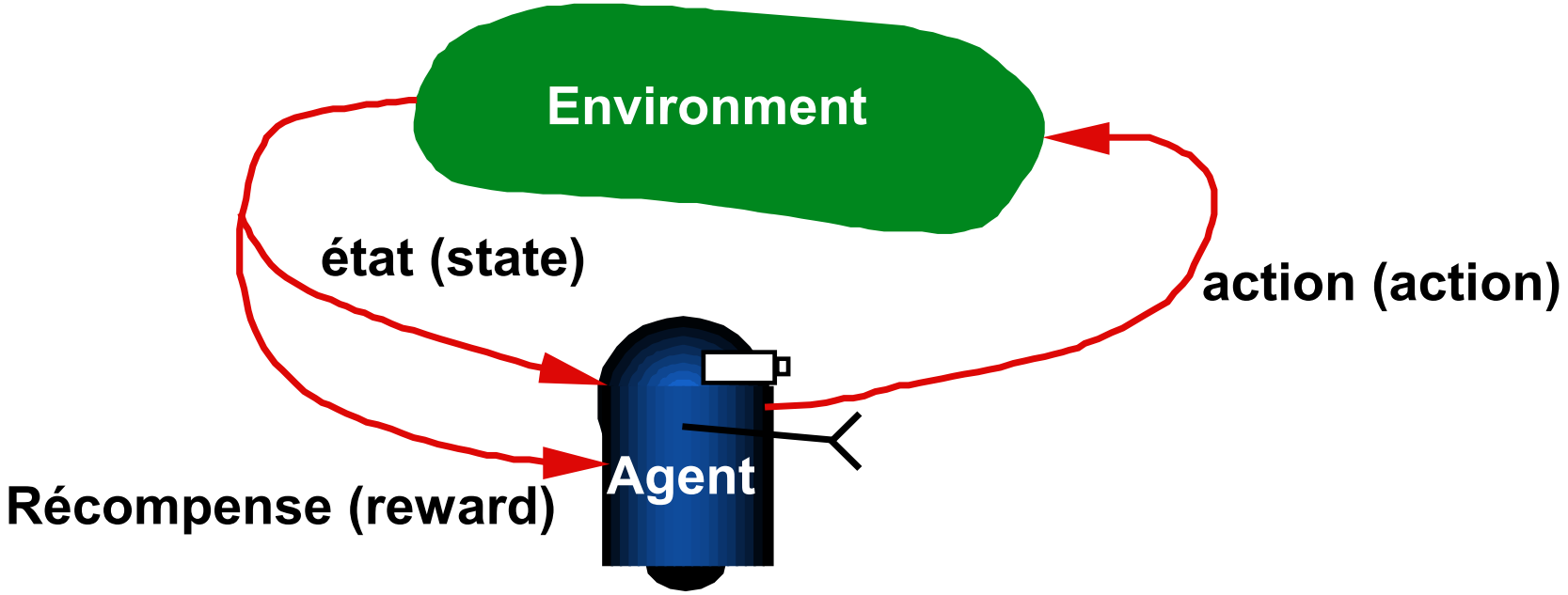
- Considère le problème global d'un agent **orienté-objectif** (goal-directed agent) qui agit avec un environnement incertain

  Ceci est en contraste avec beaucoup d'approches (e.g. apprentissage supervisé) qui s'adressent à des sous-problèmes sans s'adresser à la question comment ils pourraient correspondre à une problème plus large.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Agent Complet

- Processus temporel
- Apprentissage et planification en continue
- L'objectif est d'agire sur l'environnement
- L'environnement est stochastique et incertain



Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Exemples

- **Jeux d 'échec** : Un bon joueur d'échec effectue un mouvement. Le choix d'une pièce est guidé à la fois par la planification de ses réponses et par l'anticipation des contre-réponses possibles (de l'adversaire) et par les jugements intuitifs immédiats de l'appréciation des positions et des mouvements.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Exemples

- **Contrôleur d'une raffinerie de pétrole** : Ce Contrôleur adaptatif régule en temps réel les paramètres de la raffinerie. Il optimise le compromis produit/coût/qualité en se basant sur des coûts spécifiques secondaires sans avoir besoin d 'utiliser strictement la consigne fournie à l'origine par l'ingénieur.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Exemples

- **Un bébé apprend à marcher** :

Sans commentaires, nous avons tous vécus cet épreuve!!!

Exemples

- **Un robot mobile** : Un robot mobile qui décide s'il devrait entrer dans une autre pièce à la recherche d'un casier supplémentaire pour la collecte des canettes ou commencer d'essayer de trouver son chemin à sa station pour la recharge de ses batteries. Il prend sa décision en se basant sur son expérience dans le passé pour suivre le chemin le plus simple et le plus rapide pour rejoindre cette station.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Exemples - remarques

- Tous ces exemples partagent les caractéristiques suivantes :
  - **interaction** : agent effectuant une décision active/environnement
  - objectif incluant un **environnement incertain**
  - les actions de l'agent peuvent **influencer l'état future** de l'environnement (e. g. la position suivante d'une pièce d'échec, le niveau des réservoirs de la raffinerie, la position par rapport au sol pour le bébé, la prochaine position du robot). Par conséquent, les options et les opportunités disponibles pour l'agent seront affectées dans le future.

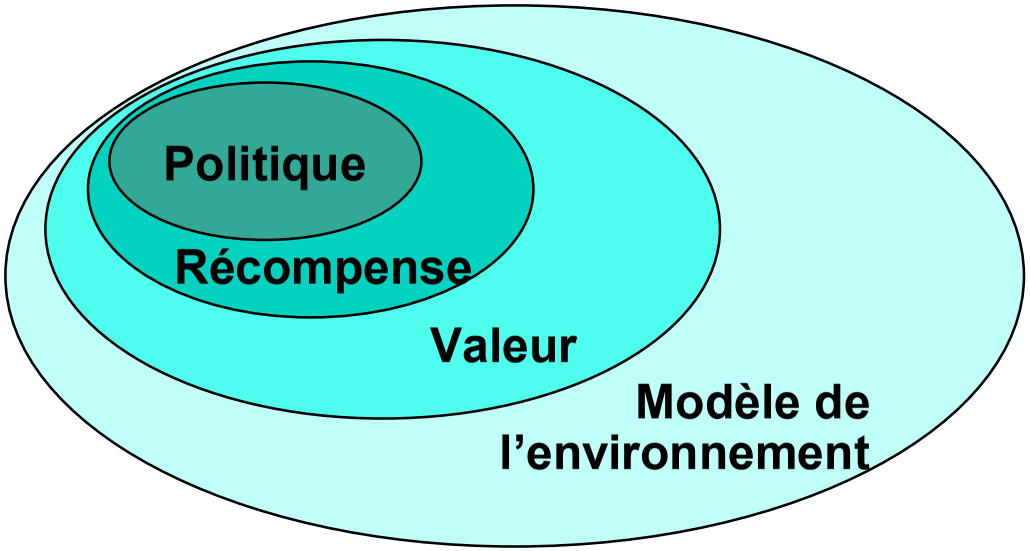
Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Exemples - remarques (suite)

- Les **effets des actions** ne peuvent pas être prédits complètement, ainsi l'agent doit fréquemment contrôler son environnement et agir d'une façon appropriée.
- Les **objectifs** sont connus par l'agent (gagner pour le joueur d'échec, la quantité de pétrole à produire pour le contrôleur, être debout pour le bébé, le début du déchargement des batteries pour le robot mobile).
- L'agent utilise son **expérience** pour améliorer .

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Éléments de RL



- **Politique (Policy) P** : Ce q'on doit faire
- **Récompense (Reward) r** : Ce qui est bien
- **Pénalité (Panalty) (-r)** : Ce qui est mouvais
- **Valeur (Value) V** : Ce qui est bien parce qu'elle *prédit* la récompense
- **Modèle de l'environnement (Model) M** : quoi suit quoi

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Éléments de RL

- **Politique P**: le comportement de l'agent à un instant donné :

  Une **association** (mapping), en général stochastique, à partir des états de l'environnement mesurés par l'agent aux actions à prendre quand l'environnement se trouve dans ces états.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Éléments de RL

- **Politique P** (suite) :

En **Psychologie** : un ensemble des règles ou associations stimulus-réponse.

Autres cas :

- Fonction simple : table de correspondance « lookup table »
- une fonction complexe nécessitant des calculs intensifs tel qu'un processus de recherche.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Éléments de RL

- **Récompense r** : une fonction qui définit une **association** (mapping), en général stochastique, à partir des états de l'environnement mesurés par l'agent ou à partir des couples état-action à un seul nombre, récompense, indiquant l'appréciation intrinsèque de l'état.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Éléments de RL

- **Récompense r** (suite) : Une valeur immédiate qui définit les caractéristiques du problème vues par l'agent. Tel qu'il est, r doit être nécessairement fixé. Elle peut, cependant, être utilisée comme une base pour modifier la politique.

  Exemple : si une action sélectionnée par une politique est suivie par une récompense faible alors cette politique peut être modifiée pour sélectionner une autre politique quand l'agent se trouve dans la même situation dans le future.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Éléments de RL

- **Pénalité (-r)** : Une valeur négative immédiate qui définit les caractéristiques du problème vues par l'agent.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Éléments de RL

- **Valeur V** : Tandis qu'une fonction r indique ce qui est bien dans l'immédiat, une fonction valeur spécifie ce qui est bien dans le future.

  Une Valeur de l'état est la quantité totale des récompenses accumulées par l'agent dans le future en partant de cet état.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Éléments de RL

- **Valeur V** (suite) : Malheureusement, il est plus difficile de déterminer V que de déterminer r.

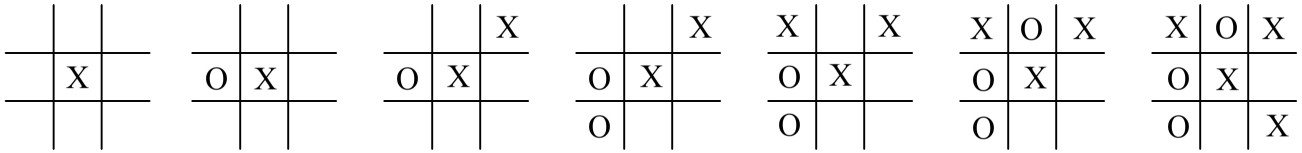
  r est, en principe, donnée directement par l'environnement tandis que V doit être **estimée** et **ré-estimée** à partir des séquences d'observations (séquences d'états de l'environnement) que l'agent fait sur la totalité de sa vie (ou totalité du processus).

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Éléments de RL

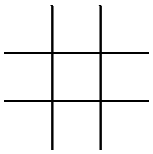
- **Modèle de l'environnement M** : Quelques chose qui imite l'environnement. Par exemple, étant donné un couple état-action un modèle permet de prédire le couple état-action future.

Exemple: Jeu de morpion (Tic-Tac-Toe)



- Supposer un adversaire imparfait : il/elle fait quelques fautes
- Si l'un des deux joueurs réussit le premier à aligner sur une ligne horizontale ou verticale ou diagonale alors ce jeu est gagné (win) par ce joueur
- Si aucun joueur n'a gagné alors fin du jeu (draw)
- Comment pourrions nous apprendre à un joueur, qui découvrira des imperfections dans les mouvements de son adversaire, à maximiser sa chance de gagner?

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

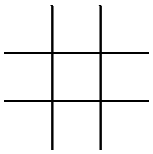


Exemple: Jeu de morpion (Tic-Tac-Toe)

- Bien que ce soit un problème très simple, il ne peut pas aisément être résolu d'une manière satisfaisante par des techniques classiques.
- Exemple : Technique minimax suppose que le joueur adversaire joue toujours d'une manière particulière.

  Par exemple, un joueur de minimax n'atteindrait jamais un état de jeu dont il pourrait perdre, même si en fait il gagnait toujours de cet état en raison du jeu incorrect par l'adversaire.

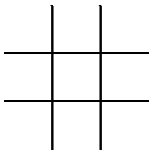
Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



Exemple: Jeu de morpion (Tic-Tac-Toe)

- Méthodes d'optimisation classiques pour les problèmes à décisions séquentielles comme la programmation dynamique peuvent calculer une solution optimale pour n'importe quel adversaire mais elles nécessitent une spécification complète de cet adversaire (modèle) dont les probabilités avec lesquelles il effectue chaque mouvement dans chaque états du jeu. Supposons que ce modèle n'existe pas a priori ce qui est le cas pour ce problèmes et beaucoup d'autres.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



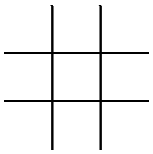
Exemple: Jeu de morpion

- En pratique un tel modèle n'est pas disponible.

Ce modèle peut alors être estimé à partir de l'expérience en jouant dans ce cas plusieurs fois contre le même adversaire.

- Étudier d'abord son modèle comportement jusqu'à un certain limite de confidence;
- En se basant sur ce modèle, appliquer la programmation dynamique pour calculer une solution optimale.

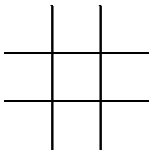
Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



Exemple: Jeu de morpion (Tic-Tac-Toe)

- Une méthode évolutionnaire pourrais chercher directement dans l'espace des politiques possibles celle qui possède une probabilité élevée de gagner contre l'adversaire. Ici, une politique est une règle qui indique au joueur quel mouvement doit effectuer pour chaque état du jeu : chaque configuration possible des X et des O sur le tableau 3X3. Pour chaque politique considérée, une estimation de sa probabilité de gagner serait obtenue en jouant un certain nombre de jeux contre l'adversaire. Cette évaluation déterminerait alors quelle politique (ou politiques) peut être considérée(s).

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

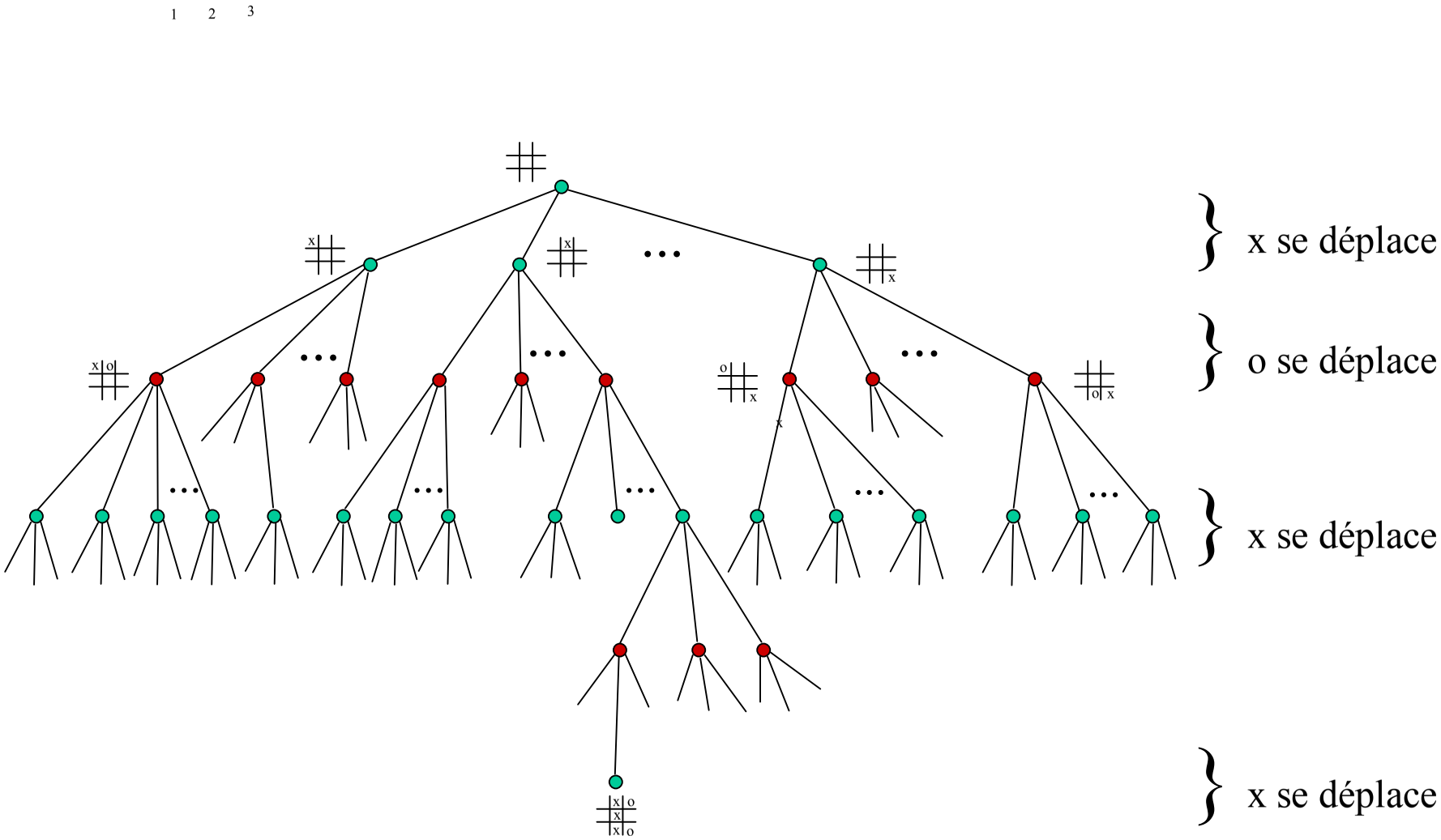


Exemple: Jeu de morpion (Tic-Tac-Toe)

- Une méthode évolutionnaire typique monte en grade « hillclimb » dans l'espace de politique, successivement produisant et évaluant des politiques afin d'essayer d'obtenir des améliorations incrémentales. Ou, peut-être, on pourrait employer un algorithme de type génétique qui maintiendrait et évaluerait une population des politiques. Littéralement des centaines de différents algorithmes d'optimisation ont pu être appliquées. Rechercher directement dans l'espace de politique signifie que toutes les politiques sont proposées et comparées sur la base des évaluations scalaires.

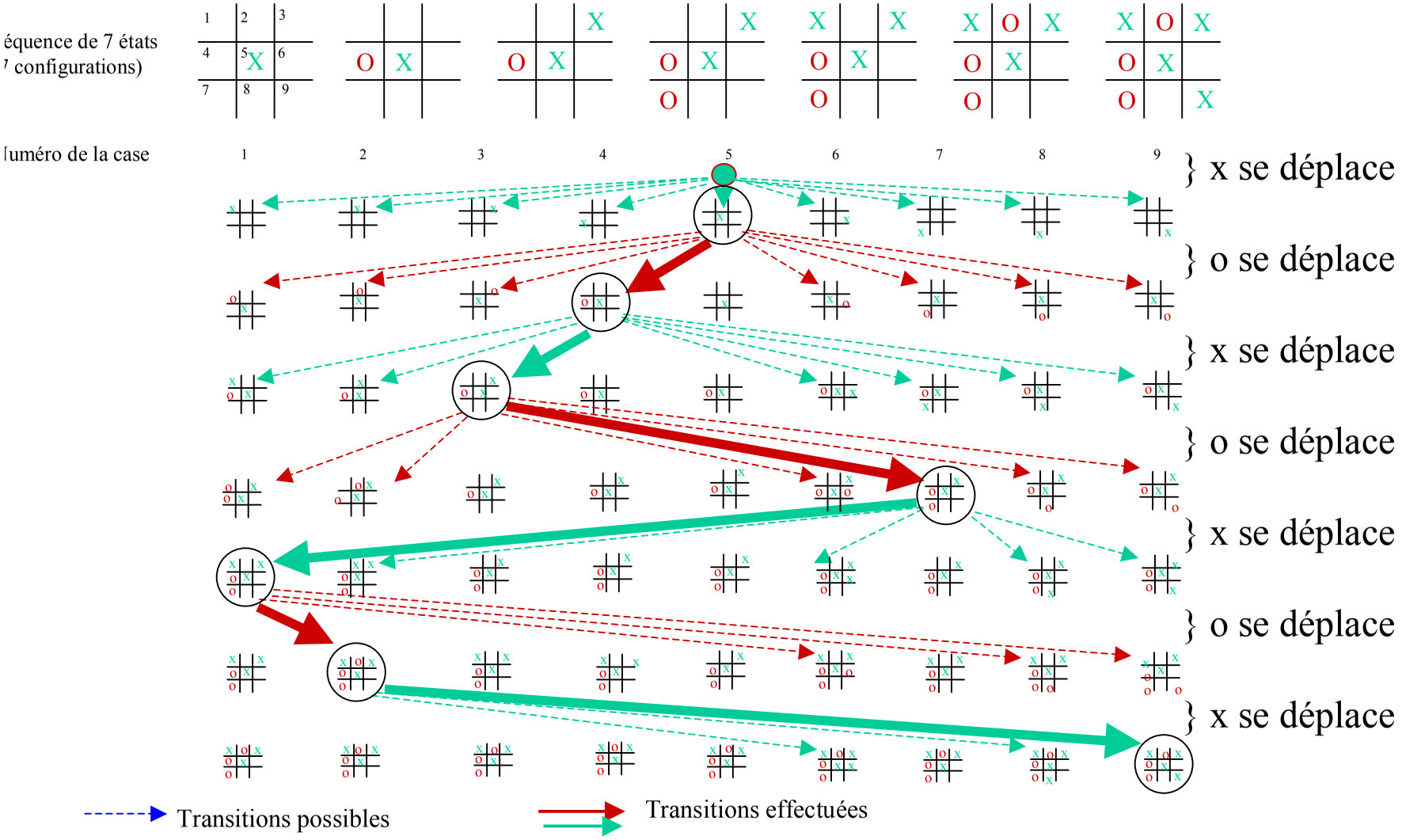
Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Exemple: Jeu de morpion

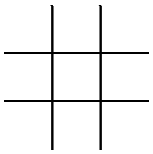


Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Exemple: Jeu de morpion



Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



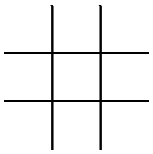
Une approche RL au jeu de morpion

1. Construire un tableau (le tableau complet est appelé **Fonction Apprise de Valeur d'état** (**Learned State Value function**) V(s)) avec l'état comme une entrée :

s-état	*V*(*s*) –probabilité estimée de gagner à partir de l'état s
	.5
	?
⋮	⋮
	1 gagner
⋮	⋮
	0 perdre
⋮	⋮
	0 fin de jeu

État s_t est mieux que état s_{t-1} si la probabilité courante estimée de gagner à partir de s_t est plus grande que celle obtenue à partir de s_{t-1}.

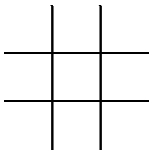
Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



Règle d'apprentissage RL pour le jeu de morpion

2. Jouer plusieurs jeux avec l'adversaire. Pour prédire le mouvement de cet adversaire, il faut chercher un pas vers l'avant.

Pour choisir nos mouvements nous examinons les états qui résulteraient de chacun de nos mouvements possibles (un pour chaque case vide sur les 9 cases) et rechercheraient leurs valeurs courantes dans la table.

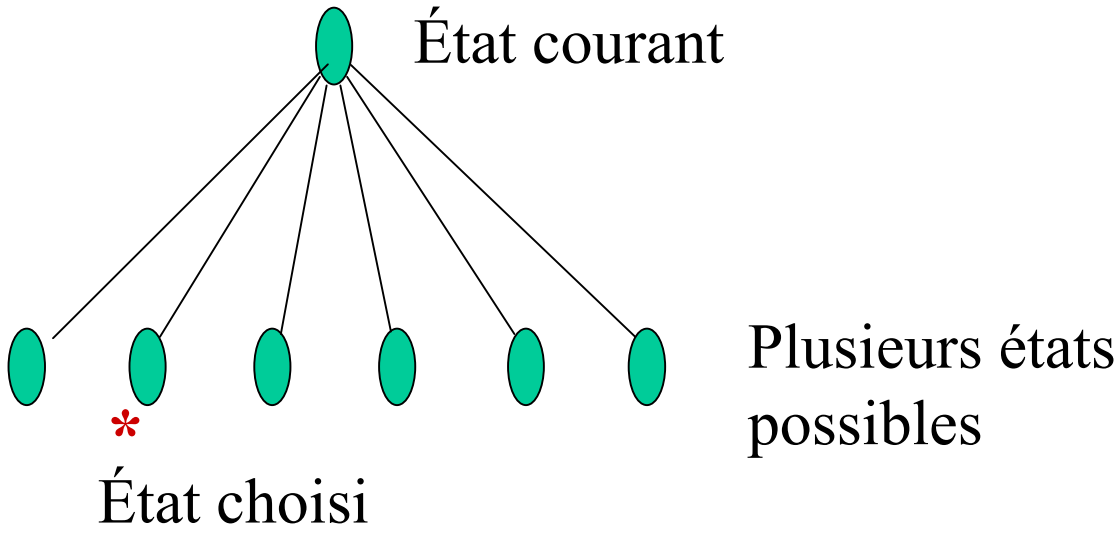


Règle d'apprentissage RL pour le jeu de morpion

La majeure partie du temps nous nous déplaçons avec un **mouvement glouton** (**greedy *move)***, choisissant le mouvement qui mène à l'état avec la plus grande valeur, c.-à-d., avec la probabilité estimée la plus élevée du gain. De temps en temps, cependant, nous choisissons aléatoirement un mouvement parmi d'autres mouvements possibles; ceux-ci s'appellent les **mouvements exploratoires (*exploratory move)*** parce qu'ils nous permettent de tester les états que nous ne pourrions jamais visiter autrement.

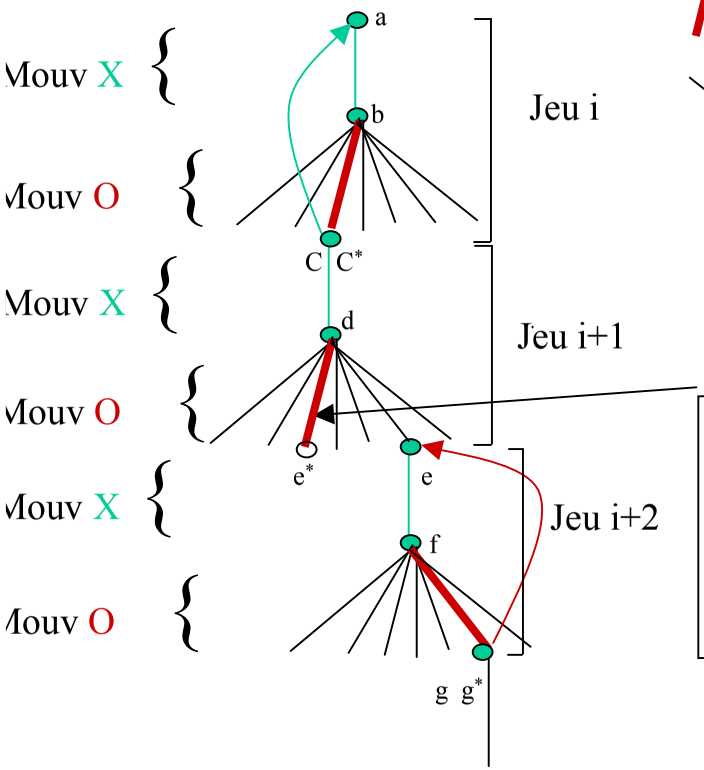
Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Règle d'apprentissage RL pour le jeu de morpion



Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Règle d'apprentissage RL pour le jeu de morpion



// Mouvement effectué pendant un jeu

Mouvements considérés (mais non effectués par le joueur X (ou par son algorithme).

Retour vers un état précédent optimal c^* ou g^* à partir du tableau

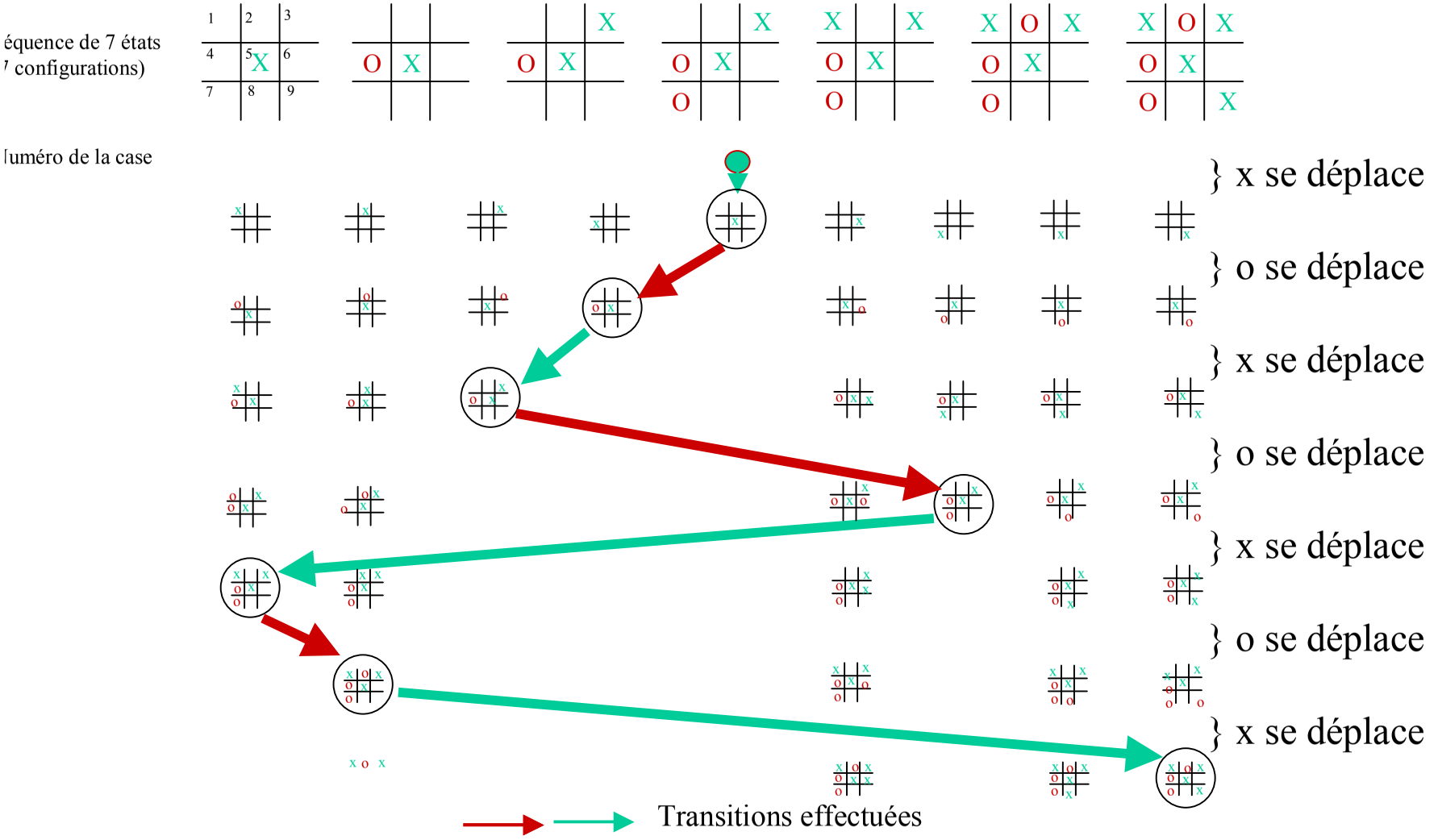
mouvement exploratoire (e* a un score plus élevé mais X a choisi quand même un autre état nouveau). Ce mouvement ne produit aucun apprentissage, contrairement à ceux réalisant un retour (mouvements gloutons).

s- l'état avant mon mouvement glouton (état précédent)

s'- l'état après mon mouvement glouton (état courant)

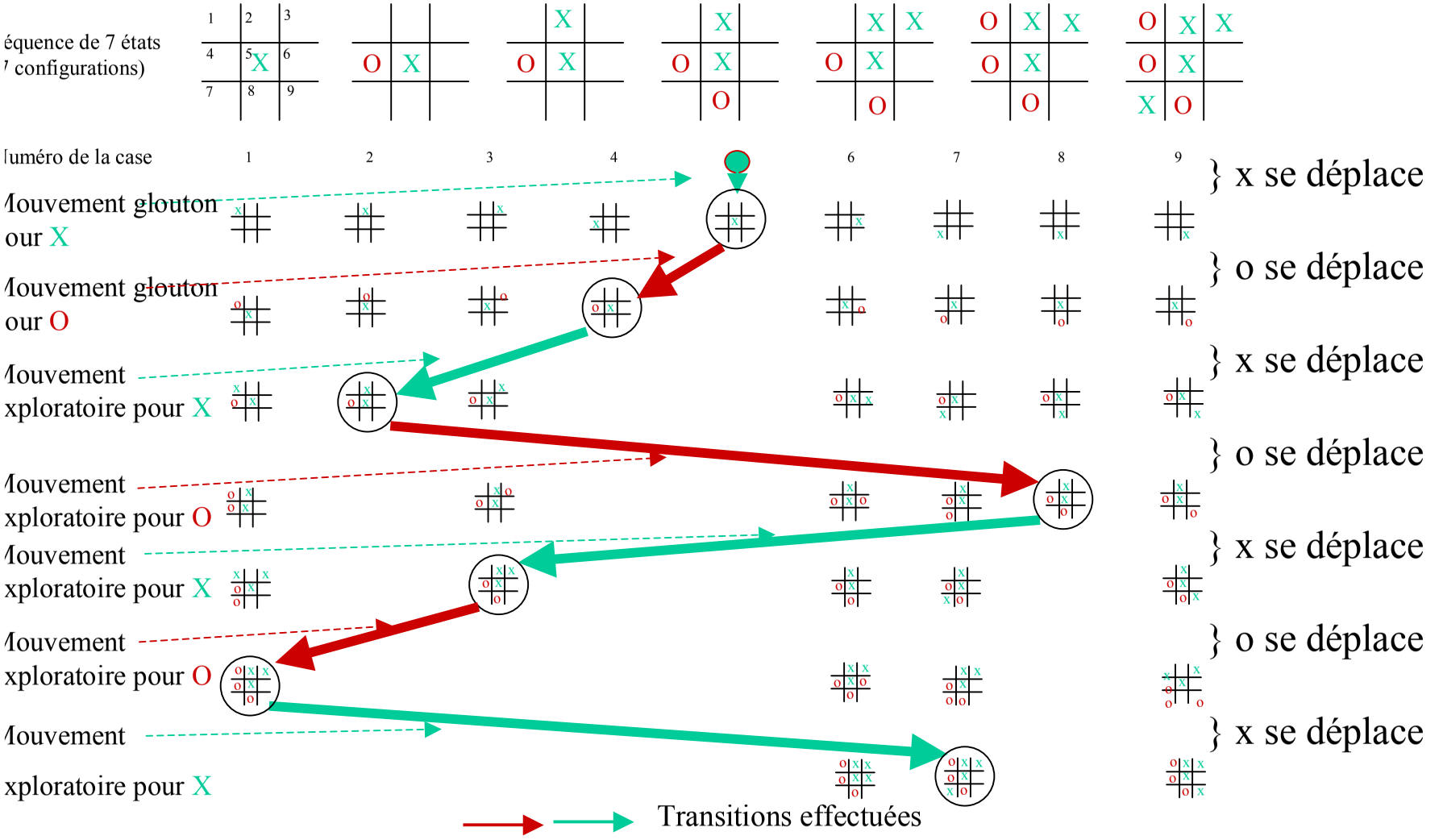
Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Exemple: Jeu de morpion-une réalisation à t=1



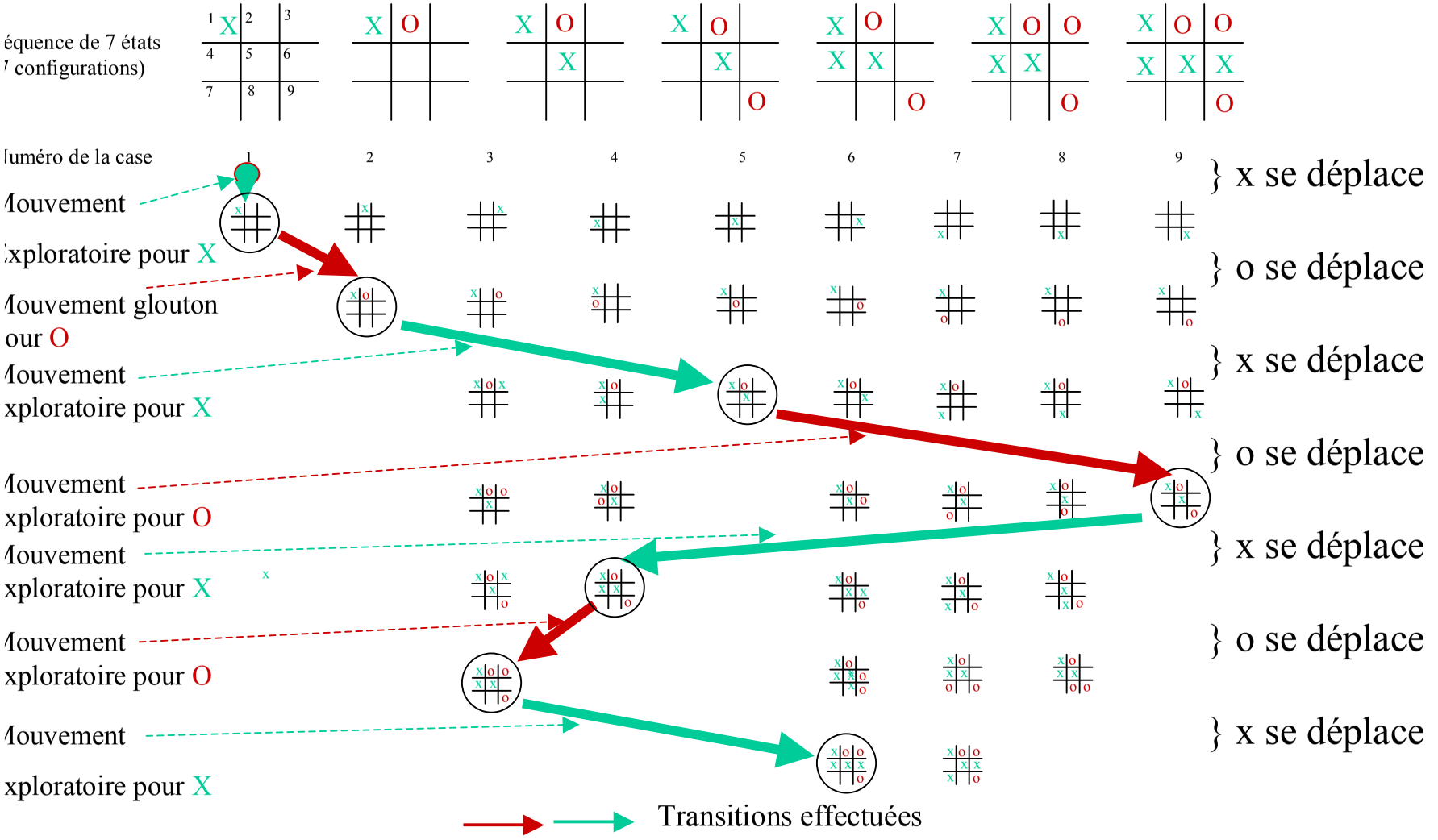
Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Exemple: Jeu de morpion -un autre jeu à t=2

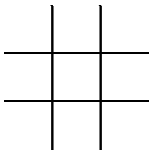


Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Exemple: Jeu de morpion -un autre jeuà t=3



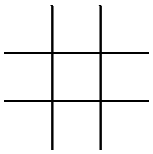
Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



Règle d'apprentissage RL pour le jeu de morpion

L'idée est la suivante :
Supposons que la probabilité exacte de gagner à partir d'un état particulier, qu'un joueur peut rencontrer après chaque mouvement, est connue pour toutes les situations possibles, il pourrait toujours dans ce cas faire une décision parfaite (en jouant de tel sorte que l'état suivant ait la probabilité maximale par rapport à tous les états suivants possibles). Mais comment peut-on obtenir ces probabilités?

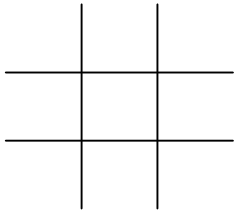
Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



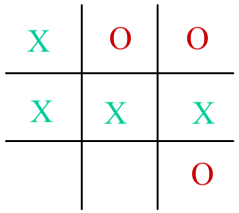
Règle d'apprentissage RL pour le jeu de morpion

Le processus d'apprentissage (estimer les probabilités de gagner à partir de chaque état) est basé sur la récompense obtenue après chaque jeu gagnant. Ainsi, nous connaissons a priori la probabilité de l'état final : elle sera soit 0 pour perdre soit 1 pour gagner (ou fin jeu sans aucun gagnant (draw)).

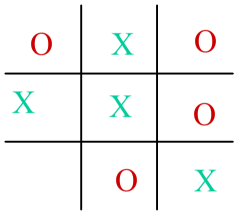
Exemple:
b : état final gagnant pour X
c : état final « draw » pour les deux joueurs.



a. Début

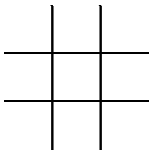


b. Fin : X gagnat



c. Fin :draw

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



Règle d'apprentissage RL pour le jeu de morpion

Initialement, tous les états de gagner sont attribués une probabilité ($V_0(s)$) de 0.5.

Après chaque jeu, tous les états intermédiaires sont mis à jour pour chaque joueur.

A fur et à mesure que nous jouons, nous changeons les valeurs des états dans lesquels nous nous trouvons. Nous tentons de choisir les bons états pour obtenir les meilleures estimations des probabilités de gagner à partir de ces états.

Pour cela, nous effectuons après chaque mouvement glouton (gagnant) un retour à l'arrière de la valeur de ces états vers les états avant le mouvement.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Règle d'apprentissage RL pour le jeu de morpion

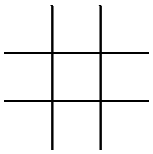
Plus précisément, la valeur de l'état précédent s est ajustée pour être plus proche à la valeur de l'état courant s'.
Nous incrémentons V(s) vers V(s') – un **recule** (**backup**) :

$$V(s) \leftarrow V(s) + \alpha(V(s') - V(s)), (0 \leq \alpha \leq 1).$$

α est un coefficient positif appelé **pas d'apprentissage** ou **paramètre de mise à jour à un pas** (**step-size parameter**).

Cette règle est un exemple de ce qui est appelé **Apprentissage par une Différence Temporelle** (**Temporal-Difference Learning**)) parce que sa mise à jour est basée sur la différence (V(s')-V(s)) entre deux estimées correspondantes à deux instants de temps.

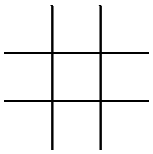
Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



Règle d'apprentissage RL pour le jeu de morpion

- En effectuant une réduction convenable de α dans le temps (le réduire par une valeur constante ou en le multipliant par une valeur positive inférieur à 1), et en jouant d'une façon optimale, cette méthode converge, pour un adversaire donné, vers les vrais probabilités de gagner à partir de chaque état.

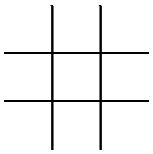
Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



Règle d'apprentissage RL pour le jeu de morpion

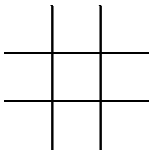
- Les mouvements ainsi effectués (sauf les cas exploratoires) sont les mouvements optimales contre l'adversaire : la méthode converge vers une politique optimale pour ce jeu.
- Si α n'est pas réduit entièrement à zéro au cours de temps, alors le joueur jouera aussi bien contre des adversaires qui modifient lentement leurs comportements au cours du temps.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



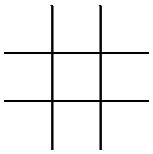
Règle d'apprentissage RL pour le jeu de morpion

- Cet exemple illustre les différences entre les méthodes évolutionnaires et les méthodes qui apprennent des fonctions de valeur. Pour évaluer une politique, une méthode évolutionnaire doit la considérer comme étant fixe et jouer beaucoup de jeux contre l'adversaire, ou simulez beaucoup de jeux en utilisant un modèle de l'adversaire.



Règle d'apprentissage RL pour le jeu de morpion

La fréquence des gains donne une évaluation impartiale de la probabilité du gain avec cette politique, et peut être employée pour diriger le prochain choix de la politique. Mais chaque modification de la politique est faite seulement après beaucoup de jeux, et seulement les résultats finals de chaque jeu sont employés : ce qui se produit pendant les jeux est ignoré.



Règle d'apprentissage RL pour le jeu de morpion

Par exemple, si le joueur gagne, alors on donne un crédit à tout son comportement dans le jeu, indépendamment du comment des mouvements spécifiques qui pourraient avoir été critiques à la victoire. Le crédit est donné même aux mouvements qui ne se sont jamais produits! Les méthodes de la fonction valeur, en revanche, permettent à différents états d'être évalués. En fin de compte, méthodes évolutionnaires et celles de la fonction valeur recherchent dans l'espace des politiques, mais l'apprentissage d'une fonction de valeur tire profit de l'information disponible pendant le jeu.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Exemple:Deux joueurs X et O [4]

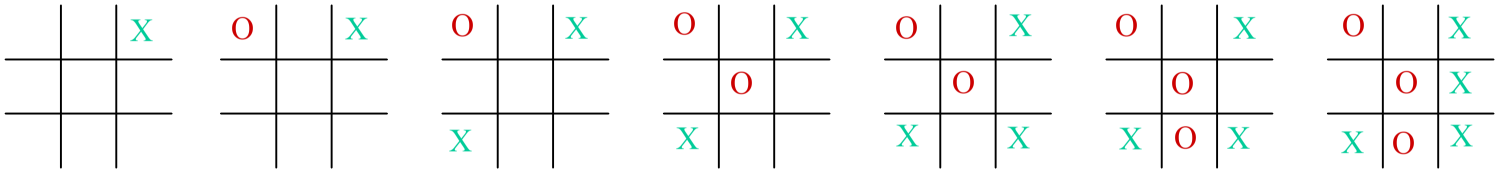
Pour tous les états :

Probabilités initiales de gagner =1- Probabilités initiales de perdre :

Probabilité (X_0) = Probabilité (O _0)=0.5,

α=0.5, i: indice de jeux

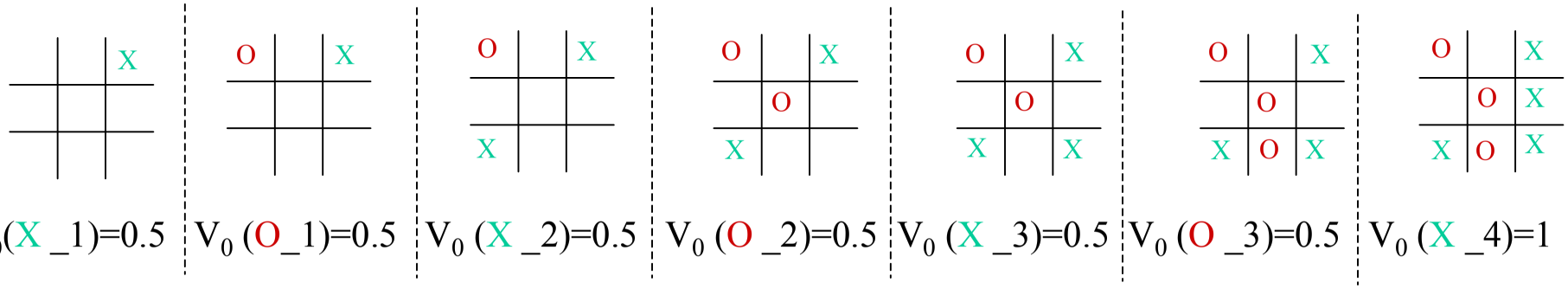
Supposons que le joueur X qui commence, voici une possibilité (jeu 1) :



Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Exemple:Deux joueurs X et O

1. Initialement, tous les états de gagner sont attribués une probabilité ($V_0(s)$) de 0.5.

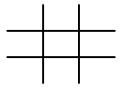
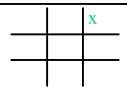
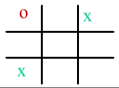
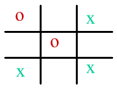
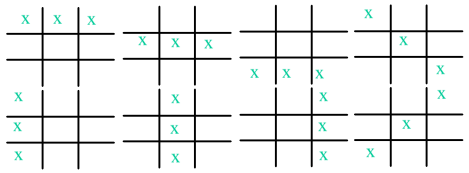
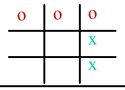
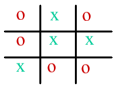


Construire pour les joueurs X et O un tableau initial ($V_0(s)$_X et $V_0(s)$_O).

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Exemple:Deux joueurs X et O

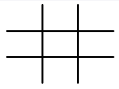
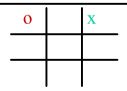
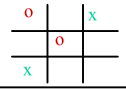
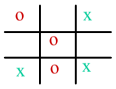
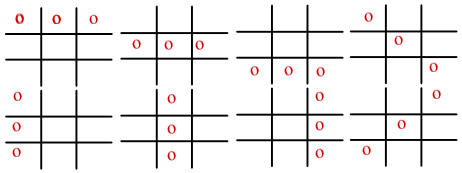
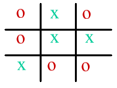
Table de $V_0(s)$_X pour le jeu 1

s-état	*V*(*s*) –probabilité estimée de gagner à partir de l'état s
	.5
	.5
	.5
	.5
	1 gagné
	0 perdu
⋮	? ? ?
	0 draw

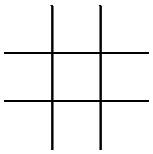
Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Exemple:Deux joueurs X et O (suite)

Table de $V_0(s)$_O pour le jeu 1

s-état	*V*(*s*) –probabilité estimée de gagner à partir de l'état s
	.5
	.5
	.5
	0
	1 gagné
	0 perdu
⋮	? ? ?
	0 draw

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



Exemple:Deux joueurs X et O (suite)

2. Mise à jour des états : La procédure de mise à jour après chaque jeu ramène le jeu vers l'arrière à partir de l'état final.

$$V(s) \leftarrow V(s) + \alpha(V(s') - V(s)), \quad \alpha = 0.5.$$

Puisque le joueur X gagne à l'état X_4, alors cet état sera récompensé avec p(X _4)=V(X _4)=1. L'état précédent serait X _3. Selon la règle de mise à jour,

V(X _3) ← V(X _3)+α(V(X _4)- V(X _3))

V(X _3) ← 0.5+0.5x(1 - 0.5)=0.75

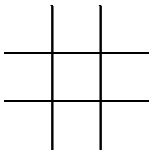
V(X _2) ← V(X _2)+α(V(X _3)- V(X _2))

V(X _2) ← 0.5+0.5x(0.75 - 0.5)=0.625

V(X _1) ← V(X _1)+α(V(X _2)- V(X _1))

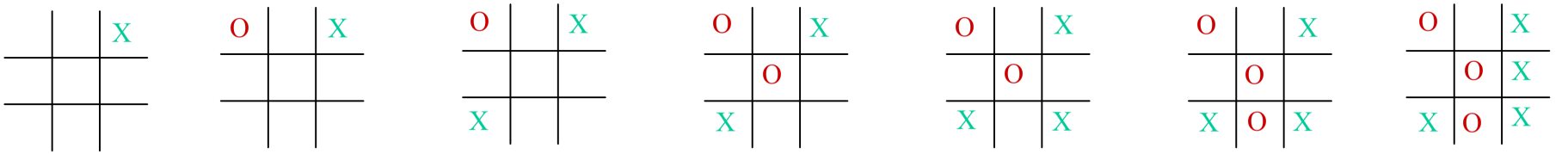
V(X _1) ← 0.5+0.5x(0.625 - 0.5)=0.5625

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



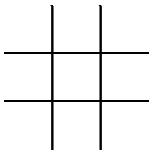
Exemple:Deux joueurs X et O (suite)

Nous faisons la mise à jour pour le joueur O de la même façon.



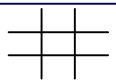
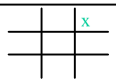
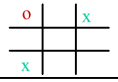
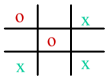
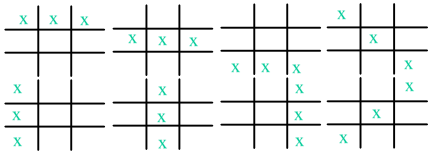
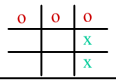
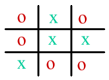
/(X _1)=0.5625 V(O _1)=0.125 V(X _2)=0.625 V(O _2)=0.25 V(X _3)=0.75 V(O _3)=0 V(X _4)=1

Une table de correspondance « lookup table » sera donc construit pour chaque joueur.

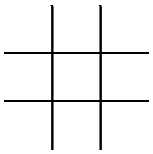


Exemple:Deux joueurs X et O (suite)

Construire la table de V(s) du joueur X pour le jeu 1

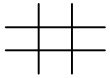
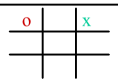
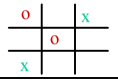
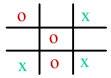
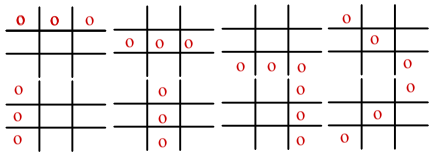
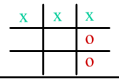
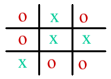
s-état	*V*(*s*) –probabilité estimée de gagner à partir de l'état s
	.5
	.5625
	.625
	.75
	1 gagné
	0 perdu
⋮	? ? ?
	0 draw

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

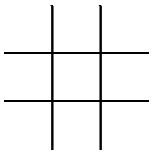


Exemple:Deux joueurs X et O (suite)

Construire la table de V(s) du joueur O pour le jeu 1

s-état	*V*(*s*) –probabilité estimée de gagner à partir de l'état s
	.5
	.125
	.25
	0
	1 gagné
	0 perdu
	? ? ?
	0 draw

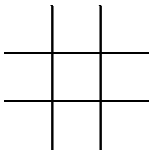
Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



Exemple:Deux joueurs X et O (suite)

Si le joueur X commence de nouveau un autre jeu, tous les états pour tous les cases vides auraient une chance de gagner de 0.5 sauf pour l'état X _1 (**0.5625**). Son premier mouvement serait semblable. De la même façon, tous les états pour le joueur O dans ce cas ont une chance de gagner de 0.5 (ils n'ont été encore jamais rencontrés) sauf pour l'état O_1 (**0.125**). En ce moment, ce joueur pourrait faire un choix d'un état différent.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



Exemple:Deux joueurs X et O (suite)

Objectif: Calculer la probabilité $V(s_i)$ de gagner à partir de l'état s_i

Idée:Après chaque jeu, faire le mise à jour de $V(s_i)$ pour tout s_i revisité

Algorithm:

(1) Initialiser tous les états avec $V(s_i)=0.5$

(2) Jouer un jeu (essayer tout $V(s_j)$ « choix glouton » (exploitation) ou une « choix aléatoire » (exploration))

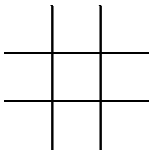
(3) V(état final) = $\begin{cases} 0 & \textit{pérdu} \\ 1 & \textit{gagné / match nul} \end{cases}$

(4) Faire la mise à jour des états intermédiaire : $V(s_i):=V(s_i)+\alpha[V(s_j)-V(s_i)]$

s_i : état avant s_j; α : taux d'apprentissage (learning rate)

(5) Revenir à (2)

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



Exemple:Deux joueurs X et O (suite) Convergence

Si α diminue avec le temps, tous les $V(s_i)$ convergent.

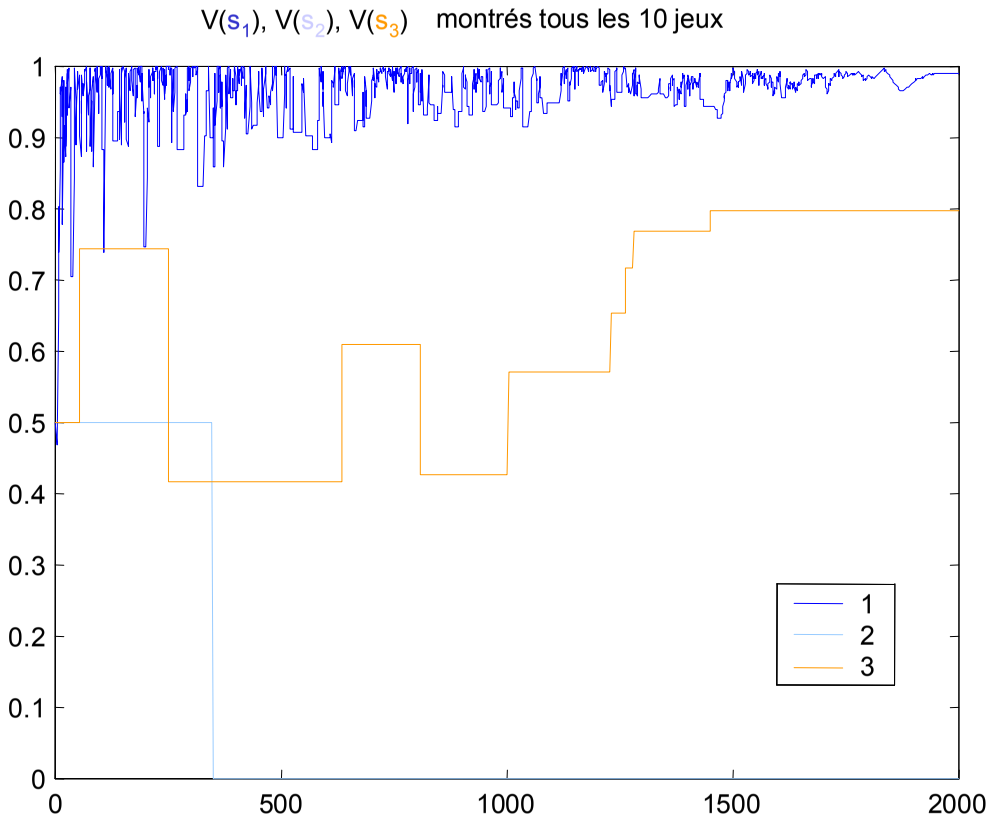
Simulation :

- 2000 jeux entre 2 tableaux de correspondance.
- En moyenne, 15 mouvements exploratoires sur 100 ont été choisis aléatoirement.
- Tous les 10 jeux, les V valeurs des trois exemples d'états futures choisis pour le joueur X sont enregistrées et tracées.

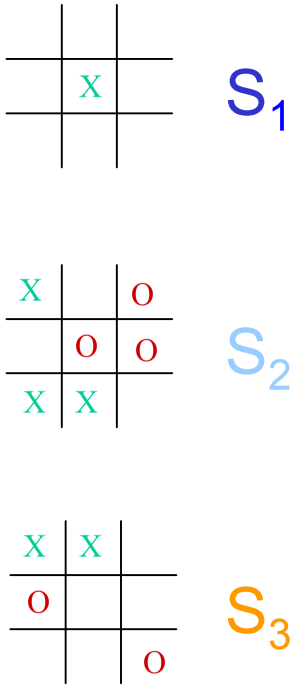
Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Exemple:Deux joueurs X et O (suite)
Convergence (suite)

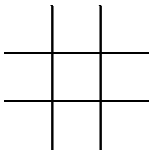
Si α diminue avec le temps, tous les $V(s_i)$ convergent.



exemples : état suivant



Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



Exemple:Deux joueurs X et O (suite)
Convergence (suite)

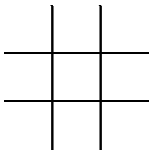
Remarques sur le choix des états suivants :

• État 2 est un état non désirable puisqu'il permet au joueur O de gagner facilement. Ainsi sa valeur initiale de 0.5 a chutée brusquement à zéro après environs 400 jeux et y restera pour toujours.

• État 1 est hautement désirable. Ainsi sa valeur tend vers un.

• La valeur de l'état 3 tend vers 0.8. Cependant, cet état est non désirable puisqu'il permet au joueur O de gagner facilement. Cependant, cet état est l'un des six états suivants possibles que le joueur X peut choisir. Tous les 5 autres états possèdent une valeur élevée. Ceci signifie que ce joueur évitera de choisir cet état sauf s'il effectue une choix aléatoire pour explorer.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

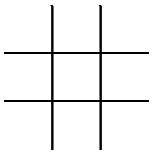


Exemple:Deux joueurs X et O (suite)
Génération des tables de correspondance

Une approche simple consiste à générer tous les états futures possibles pour chaque joueur et initialiser leurs V-valeurs (probabilités) correspondantes à 0.5.

Une autre approche simple consiste à commencer par une table vide, puisque chaque états future qui n'existe pas encore dans cette table posséderais une valeur initiale de 0.5, ensuite construire cette table en rajoutant les états visité pour la première fois ou mettre à jour les états qui existent déjà dans cette table.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



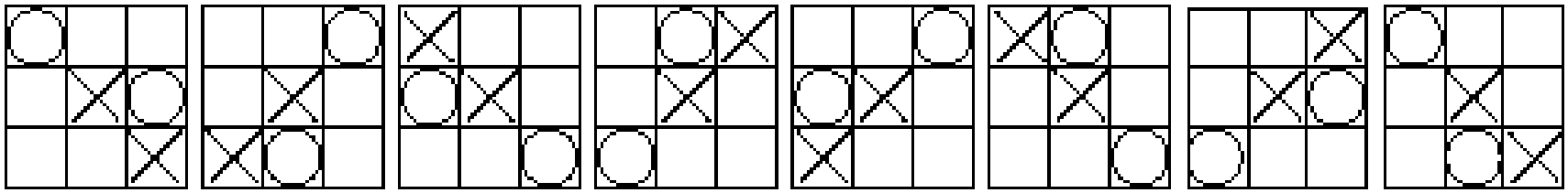
Exemple:Deux joueurs X et O (suite)
Génération des tables de correspondance
-Symétrie-

En simulant tous les états futures possibles pour un jeu de morpion 3x3, le nombre total d'états est égale alors à 5890 états et si nous augmentons la taille de la grille, ce nombre d'état augmente drastiquement : beaucoup de mémoire et temps CPU (apprentissage et convergence lents) . Il est alors nécessaire de réduire ce nombre (réduire la dimension de la table) en exploitant la propriété de symétrie entre certains états.

Exemple:Deux joueurs X et O (suite)
Génération des tables de correspondance

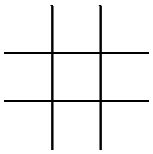
-Symétrie-

Pour exploiter la symétrie entre certains états, il est nécessaire de faire une correspondance unique qui met en correspondance tous les états équivalents (huit états) de point de vue leur valeur V à un et uniquement un seul état.



Exemples d'états équivalents par une rotation et/ou une symétrie horizontale ou verticale.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

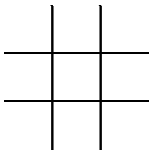


Exercices

Exercice 1.1 Jeu d'Individu (Self-Play)

Supposez, au lieu du jouer contre un certain adversaire, l'algorithme RL décrit ci-dessus a joué contre lui-même. Que pensez-vous vous de ce que pourrais produire dans ce cas-ci ? Apprendrait-il une manière différente du jeu ?

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

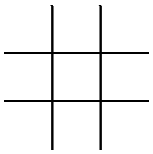


Exercices

Exercice 1.2 Symétries(Symmetries)

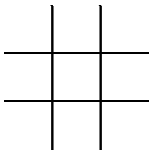
Beaucoup de positions du jeu de morpion semblent différentes mais elles sont vraiment identiques en raison des symétries. Comment pourrions-nous modifier l'algorithme RL décrit ci-dessus pour tirer profit de ceci ? De quelles manières est-ce que ceci l'améliorerait ? Réfléchissez encore, supposez que l'adversaire n'a pas tiré profit des symétries. Dans ce cas, devrions nous le faire? Puis est-il vrai que les positions symétriquement équivalentes devraient nécessairement avoir la même valeur ?

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



Exercices

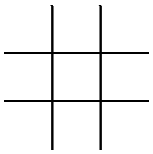
Exercice 1.3 Jouer gloutonnement (Greedy play) Supposez que le joueur RL était glouton, c.-à-d., il a toujours joué le mouvement qui lui a apportée à la position qui a été évaluée la meilleur. Apprendrait-il à jouer mieux, ou plus mauvais, qu'un joueur non-glouton ? Quels problèmes pourraient se produire ?



Exercices

Exercice 1.4 Apprentissage par l'exploration (learning from Exploration) Supposez que des mises à jour de l'apprentissage se sont produites après tout les mouvements, y compris les mouvements exploratoires. Si le paramètre α est réduit avec le temps convenablement, alors les valeurs d'état convergeraient à un ensemble de probabilités. Quels sont les deux ensembles de probabilités calculées quand nous faisons, et quand nous faisons pas, l'apprentissage des mouvements exploratoires? Supposons que nous continuons à faire des mouvements exploratoires, Quel est l'ensemble des probabilités qui pourrait être meilleurs à apprendre? Quel ensemble aurait produire plus de gains ?

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS



Exercices

Exercice 1.5 Autres améliorations Pouvez-vous réfléchir à d'autres manières d'améliorer le joueur de RL? Y a t-il une meilleure manière de résoudre le problème de morpion posé précédemment?

propriétés principales des méthodes d'apprentissage par renforcement

- L'exemple de morpion est très simple, mais il illustre certains des propriétés principales des méthodes d'apprentissage par renforcement:

1. Il y a l'accent sur l'apprentissage tout en agissant sur l'environnement, dans ce cas-ci avec un joueur opposé.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

propriétés principales des méthodes d'apprentissage par renforcement

2. il y a un objectif clair, et le comportement correct exige la planification ou la prévoyance qui tiennent compte des effets retardés des décisions. Par exemple, le joueur simple d'apprentissage par renforcement apprendrait à installer des pièges avec des mouvements multiples contre un adversaire myope (ne regarde pas plus loin). C'est une propriété saisissante de la solution d'apprentissage par renforcement qu'ellel peut réaliser les effets effets de la planification et de la prévoyance sans employer un modèle de l'adversaire et sans effectuer une recherche explicite sur des séquences possibles des états et actions futurs.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

propriétés principales des méthodes d'apprentissage par renforcement

- Le RL également s'applique quand il n'y a aucun adversaire externe explicite, c.-à-d., dans le cas d'un « jeu contre la nature».
- Le RL n'est pas également limitée aux problèmes dans lesquels le comportement décompose en épisodes séparés, comme les jeux séparés de morpion, avec la récompense appliquée seulement à la fin de chaque épisode. Il est aussi appliqué quand le comportement continue indéfiniment et quand des récompenses de diverses grandeurs peuvent être reçues à tout moment.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Généralisation

- Malgré que dans cet exemple le nombre d'états est relativement petit, la méthode de RL peut être appliquée à un nombre très élevé d'états, même pour un nombre infini. Beaucoup d'applications combine le RL avec les réseaux de neurones RN pour généraliser à partir de l'expérience dans le passé.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Généralisation

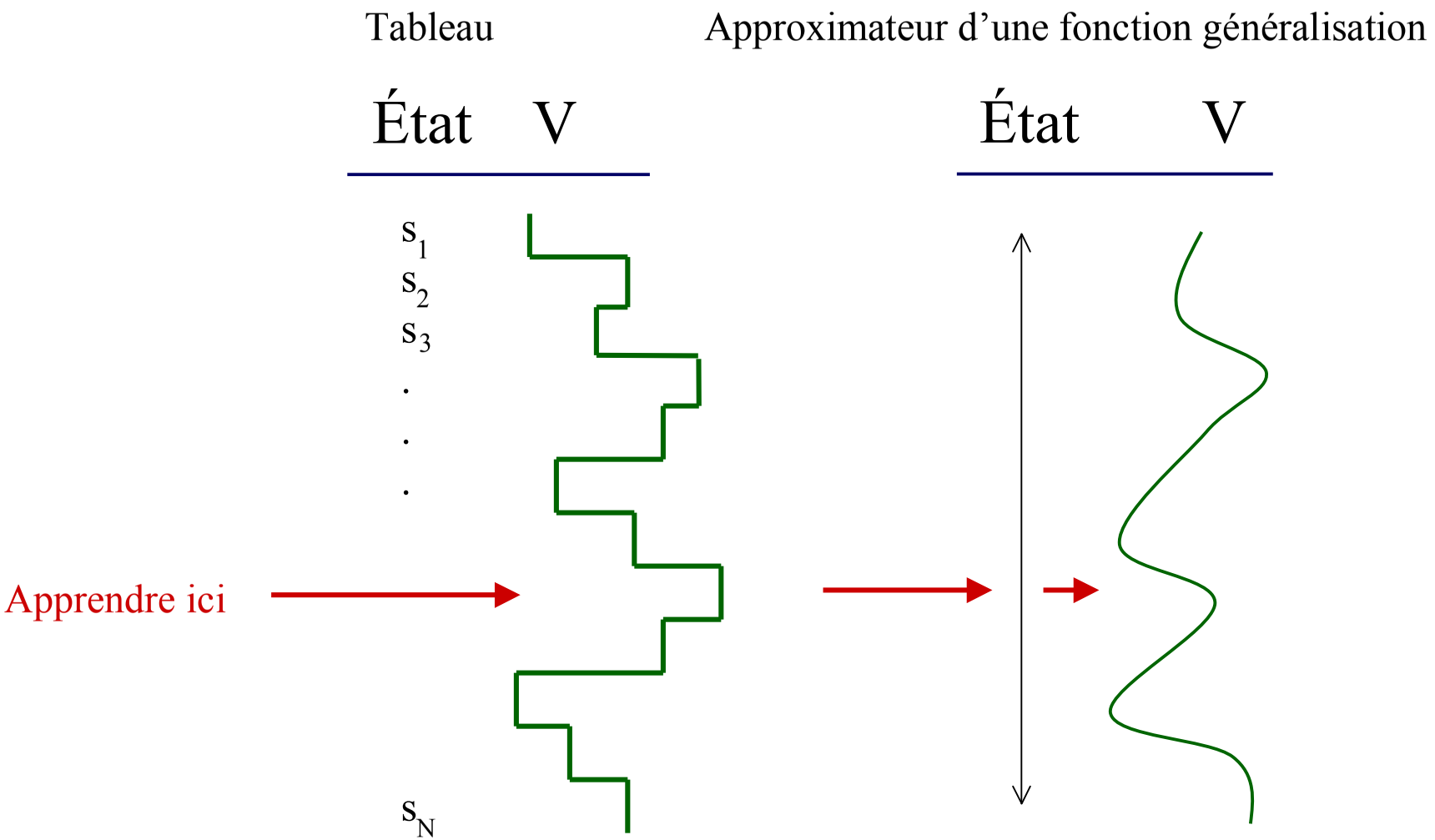
- Le réseau de neurones permettent au programme de généraliser à partir de son expérience, de sorte que dans de nouveaux états il choisisse des mouvements basés sur l'information sauvée par des états semblables rencontrés dans le passé.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Généralisation

- À quel point un agent de RL peut fonctionner dans des problèmes possédant de tels grands ensembles d'état est intimement attaché à la façon dont il peut généraliser convenablement de l'expérience antérieure. C'est dans ce rôle que nous avons le plus grand besoin de méthodes d'apprentissage supervisé pour le RL. Les réseaux de neurones ne sont pas les seuls, ou nécessairement les meilleurs, pour faire ceci.

Généralisation



Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Informations a priori

- Dans cet exemple de morpion, l'apprentissage a commencé sans la connaissance a priori au delà des règles du jeu, mais le RL nécessite nullement une vue de tabula rasa de l'apprentissage et de l'intelligence.

  Par contre, une information a priori peut être incorporée au RL par une multitude de moyens qui peuvent être importantes pour un apprentissage efficace.

Informations a posteriori

- Enfin, le joueur de morpion pouvait penser a posteriori et connaître les états qui résulteraient de chacun de ses mouvements possibles. Pour faire ceci, il a dû avoir un modèle du jeu qui lui permet de « penser » comment son environnement changerait en réponse aux démarches qu'il ne peut jamais entreprendre. Beaucoup de problèmes sont semblables à ce cas, mais dans d'autres il manque même un modèle limité des effets des actions. Le RL peut être appliquée dans les deux cas. Aucun modèle n'est exigé, mais des modèles peuvent facilement être employés s'ils sont disponibles ou peuvent être appris.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

États cachés

Nous avons également eu accès à l'état dans l'exemple de morpion, tandis que le RL peut également être appliqué quand une partie de l'espace d'état est cachée [3], ou quand les différents états semblent à l' « apprenteur » être identiques. Ce cas, cependant, est essentiellement plus difficile, et nous ne le couvrons pas de manière significative dans cette introduction.

Quelques applications remarquables de RL

- **Jeu de backgammon (TD-Gammon**: Tesauro)

  G. Tesauro (1992, 1995) a combiné le RL avec les RN pour apprendre à jouer (à un niveau de meilleur joueur du monde!) le jeu de backgammon (~10^{20} états).

- **Contrôle d'un ascenseur (Elevator Control**: Crites & Barto)

  Contrôleur de haute performance

- **Planification de l'inventaire (Inventory Management**: Van Roy, Bertsekas, Lee&Tsitsiklis)

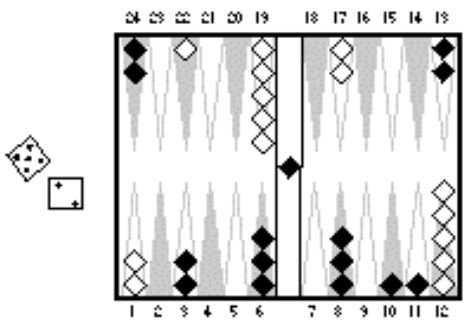
  Amélioration de 10 à15% sur les méthodes industrielles standards

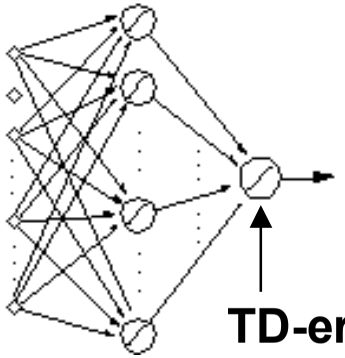
- **Allocation des canaux dynamiques (Dynamic Channel Assignment**: Singh & Bertsekas, Nie & Haykin)

  Allocation de haute performance des bandes radios pour les appels téléphones mobiles.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

TD-Gammon, Tesauro [6]





~10^{20} états

Sélection de l'action

Commencer par un réseau aléatoire

Jouer plusieurs fois contre vous même

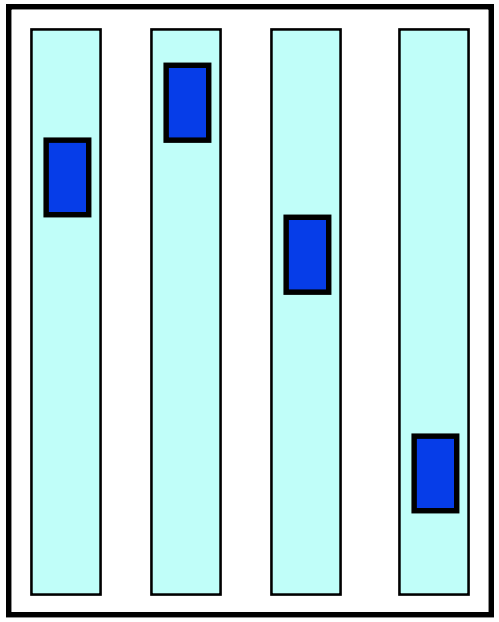
Apprendre une fonction de valeur à partir de cette expérience simulée

Ceci produit le meilleur joueur du monde

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Elevator Dispatching, Crites and Barto [6]

10 étages, 4 ascenseurs **~10^{22} états**



États : états des boutons; positions, directions, et états des mouvements des ascenseurs; passagers dans les ascenseurs & dans les halls

Actions : arrêter, ou aller à l'étage suivant

Récompenses : –1 par unité de temps pour chaque personne en attente.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Résumé de l'historique de RL

Trial-and-Error learning	Temporal-difference learning	Optimal control, value functions
Thorndike (Ψ) 1911	Secondary reinforcement (Ψ)	Hamilton (Physics) 1800s
		Shannon
Minsky	Samuel	Bellman/Howard (OR)
Klopf	Holland	
	Witten	Werbos
Barto et al.		
	Sutton	
		Watkins

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS

Références

1. Richard S. Sutton, Andrew G. Barto. « Reinforcement Learning (Adaptive Computation and Machine Learning). Ed. MIT Press, Cambidge, MA, 1998.
2. http://www-anw.cs.umass.edu/~rich/book/the-book.html
3. Andrew Kachites McCallum. « Reinforcement Learning with Selective Perception and Hidden State » (PhD Thesis), Department of Computer Science, University of Rochester, Rochester, NY, 14627, USA. http://www.cs.rochester.edu/u/mccallum/phd-thesis/
4. Sebastian Siegel. « Training an artificial neural network to play tic tac toe », ECE 539 Term Project http://www.cae.wisc.edu/~ece539/project/f01/
5. Crites, R. H. and Barto, A. G. (1996). Improving elevator performance using reinforcement learning. In D. S. Touretzky, M. C. Mozer, M. E. H., editor, *Advances in Neural Information Processing Systems: Proceedings of the 1995 Conference*, pages 1017--1023, Cambridge, MA. MIT Press.
6. Tesauro, G. J. (1994). TD--gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 6(2):215--219.

Adapté de (R. S. Sutton and A. G. Barto: Reinforcement Learning: An Introduction http://www-anw.cs.umass.edu/~rich/book/the-book.html) par Tarik AL ANI, A²SI-ESIEE-PARIS