Apprentissage par renforcement (Reinforcement Learning (RL)) Approche : Tempral Difference (TD)

T. AL-ANI
A²SI-ESIEE-PARIS

RLtoolbox

Reinforcement Learning for SCILAB

Plan de la présentation

I. Introduction

II. L'apprentissage par renforcement

III. Temporal Difference

GROUPE ESIEE PARIS

A²SI

RL toolbox

Reinforcement Learning for SCILAB

I. Introduction

- Contexte du Projet
- Objectif de notre projet
  - Étude de la théorie de l'apprentissage par renforcement
  - Implémentation des méthodes de Temporal Difference

ESIEE PARIS

A2SI

RLtoolbox

Reinforcement Learning for SCILAB

II. L'apprentissage par renforcement

- Généralités
  - Apprentissage par 'essais / erreurs'
  - Récompenses suivant les actions
  - Équilibre entre Exploration et Exploitation

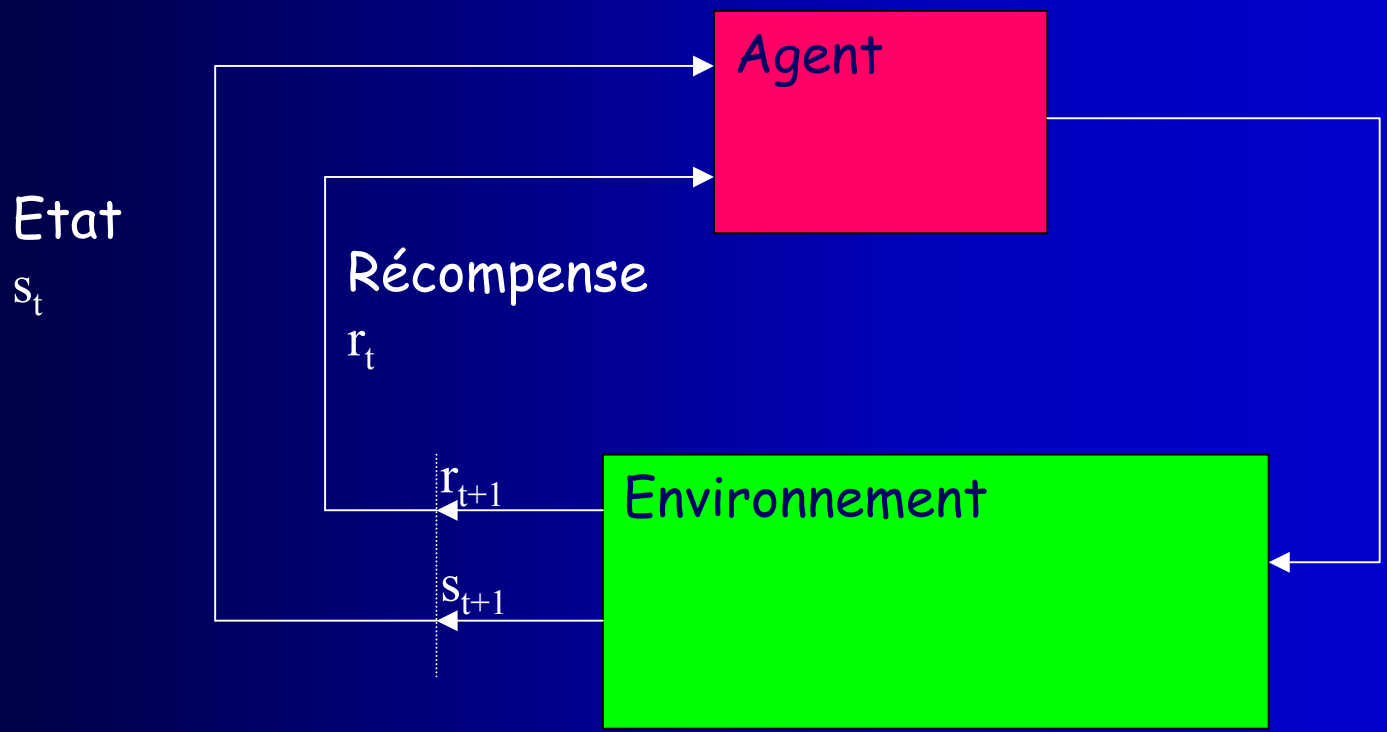
ESIEE PARIS

A2SI

RLtoolbox
Reinforcement Learning for SCILAB

II. L'apprentissage par renforcement

- Interaction Agent – Environnement



ESIEE PARIS

A²SI

RLtoolbox

Reinforcement Learning for SCILAB

II. L'apprentissage par renforcement

- Théorie
  - Évaluation de politique
  - Amélioration de politique

GROUPE ESIEE PARIS

A2SI

RLtoolbox

Reinforcement Learning for SCILAB

V. Temporal Difference

- Aucun modèle de l'environnement est nécessaire, l'expérience suffit.
- L'apprentissage se fait avant de connaître la récompense finale voire sans la connaître.
- Les méthodes convergent.

ESIEE PARIS

A²SI

RLtoolbox
Reinforcement Learning for SCILAB

V. Temporal Difference

- Théorie: TD prédiction

1. Initialize $V(s)$ arbitrarily, π to the policy to be evaluated :

2. Repeat (for each episode) :
   Initialize s
   Repeat (for each step of episode) :
   $a \leftarrow$ action given by π for s
   Take action a; observe reward, r, and next state, s'
   $V(s) \leftarrow V(s) + \alpha[r + \gamma V(s') - V(s)]$
   $s \leftarrow s'$
   until s is terminal

ALGO. 5.1: Algorithme *TD(0)* d'évaluation d'une politique

GROUPE ESIEE PARIS

A2SI

RLtoolbox
Reinforcement Learning for SCILAB

V. Temporal Difference

- Théorie: TD Sarsa

1. Initialize $Q(s,a)$ arbitrarily :

2. Repeat (for each episode) :
   Initialize s
   Choose a from s using policy derived from Q (e.g., $\epsilon-$greedy)
   Repeat (for each step of episode) :
      Take action a, observer r, s'
      Choose a' from s' using policy derived from Q (e.g., $\epsilon-$greedy)
      $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$
      $s \leftarrow s'$; $a \leftarrow a'$
   until s is terminal

ALGO. 5.2: Algorithme *Sarsa*

GROUPE ESIEE PARIS

A²SI

RLtoolbox
Reinforcement Learning for SCILAB

V. Temporal Difference

- Théorie: TD Q learning

1. Initialize $Q(s,a)$ arbitrarily :

2. Repeat (for each episode) :
   Initialize s
   Repeat (for each step of episode) :
   Choose a from s using policy derived from Q (e.g., $\epsilon-$greedy)
   Take action a, observer r, s'
   $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$
   $s \leftarrow s'$;
   until s is terminal

ALGO. 5.3: Algorithme *Q-learning*

GROUPE ESIEE PARIS

A2SI

RLtoolbox

Reinforcement Learning for SCILAB

V. Temporal Difference

- Exemple du Random Walk

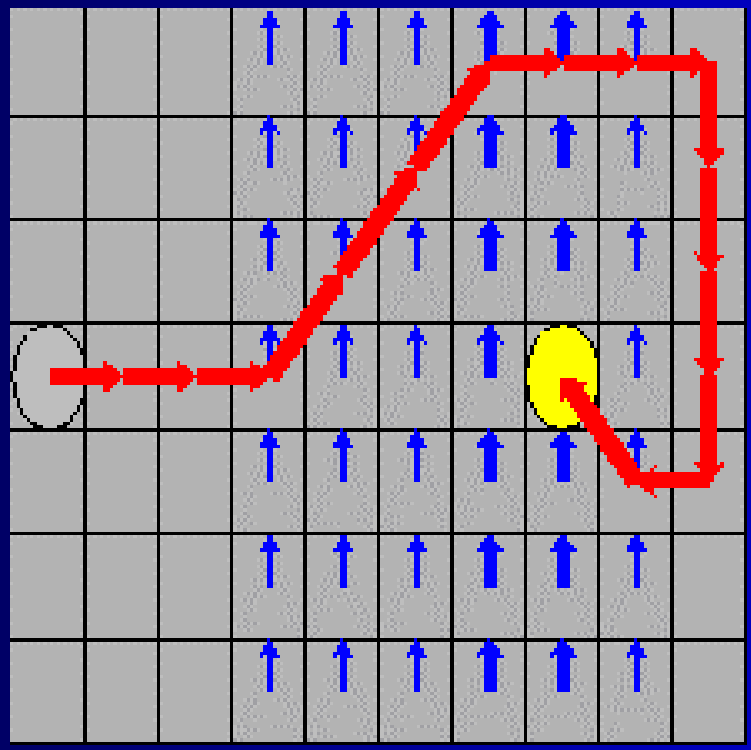
ESIEE PARIS

A2SI

RLtoolbox
Reinforcement Learning for SCILAB

V. Temporal Difference

- Exemple 1 du *Windy Grid*
  - Exemple du livre de Sutton



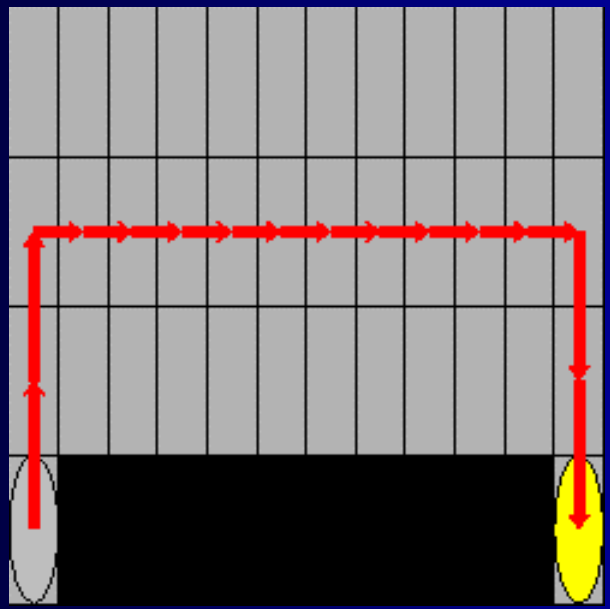
ESIEE PARIS

A²SI

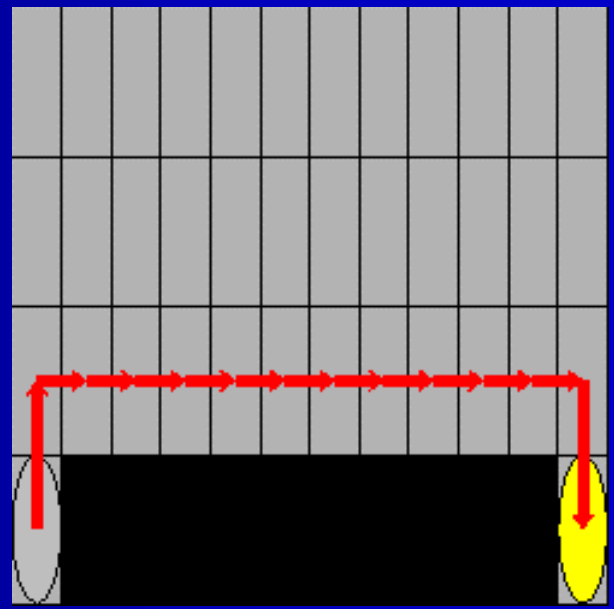
RLtoolbox
Reinforcement Learning for SCILAB

V. Temporal Difference

- Exemple 2 du *Windy Grid*
  - Cliff Walking



TD Sarsa
500 itérations



TD Q Learning
300 itérations

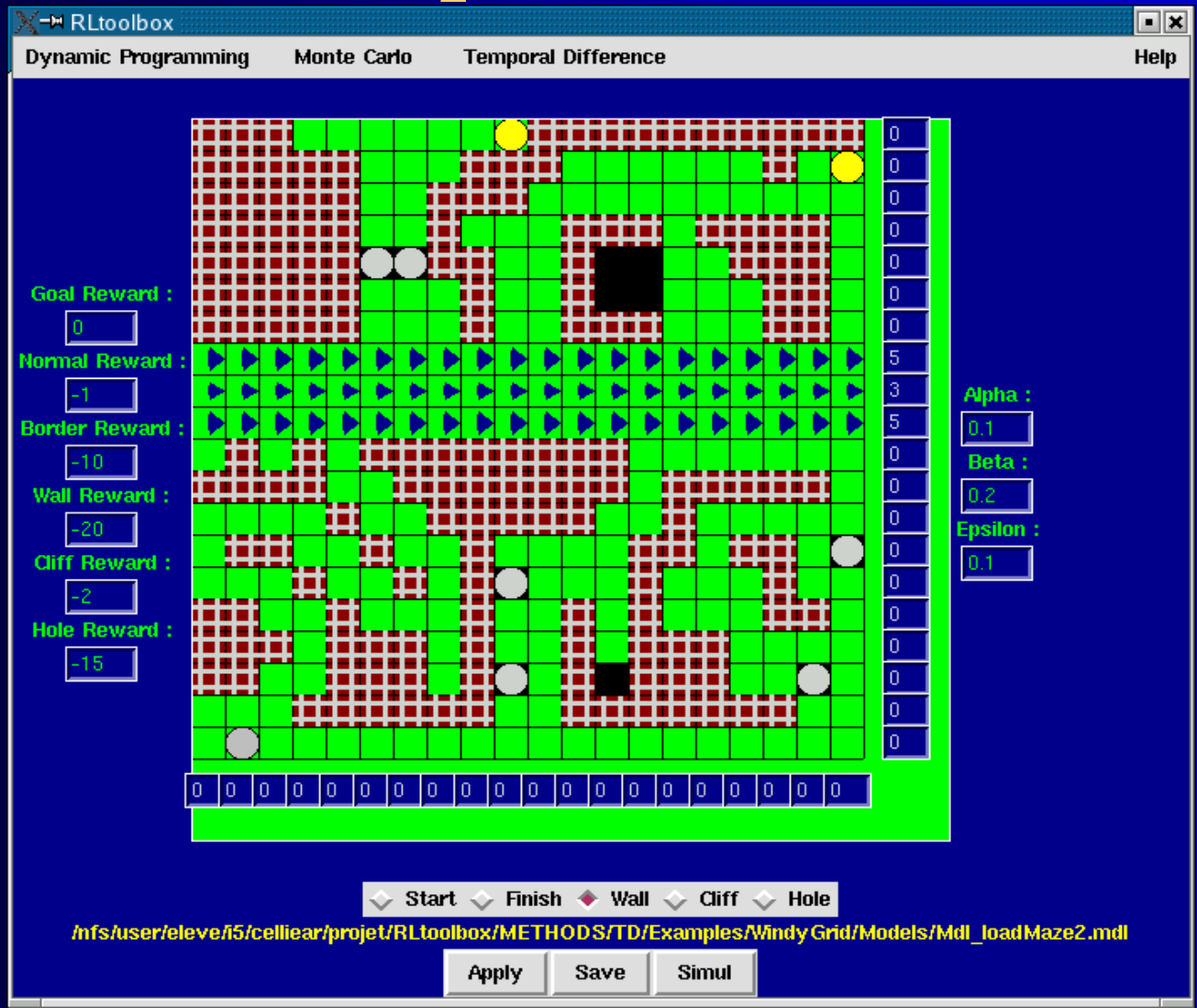
ESIEE PARIS

A²SI

RLtoolbox

Reinforcement Learning for SCILAB

V. Temporal Difference



Éditeur de grille

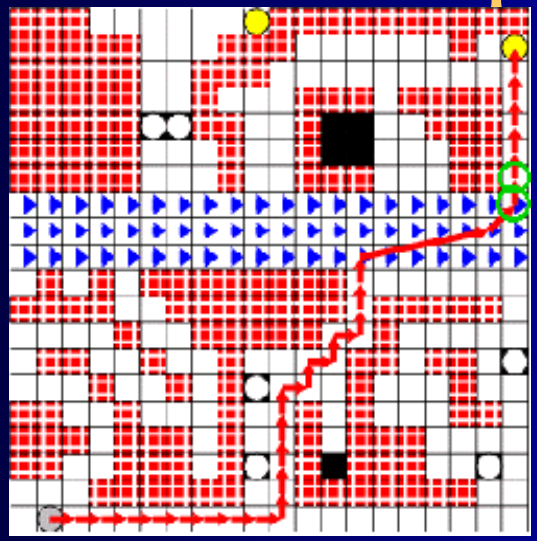
ESIEE PARIS

A²SI

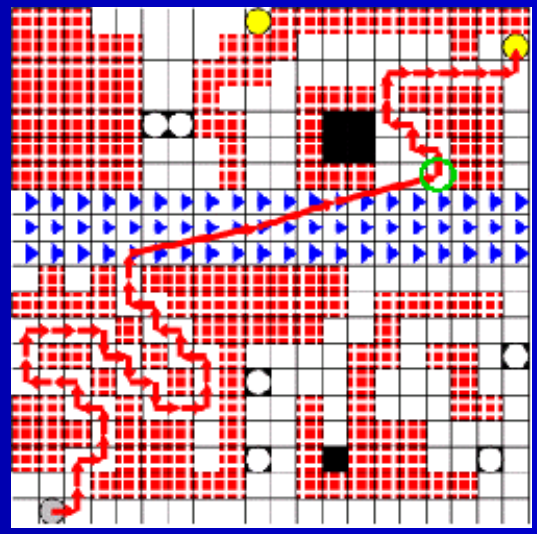
RLtoolbox

Reinforcement Learning for SCILAB

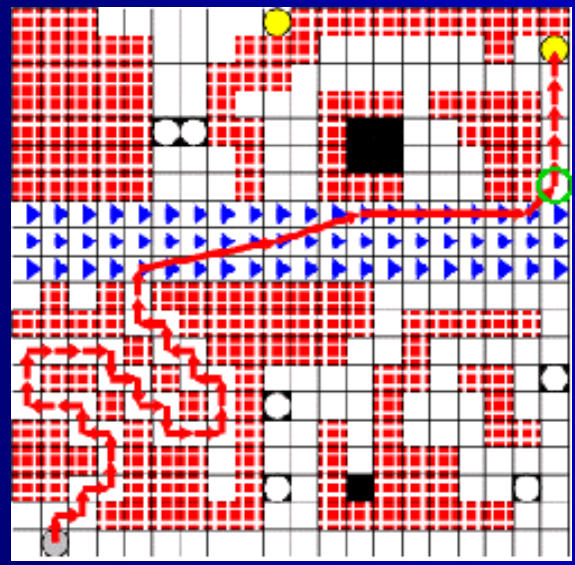
V. Temporal Difference



Mur : 25
Bord : 25



Mur : -10
Bord : -25



Mur : -25
Bord : -10

ESIEE PARIS

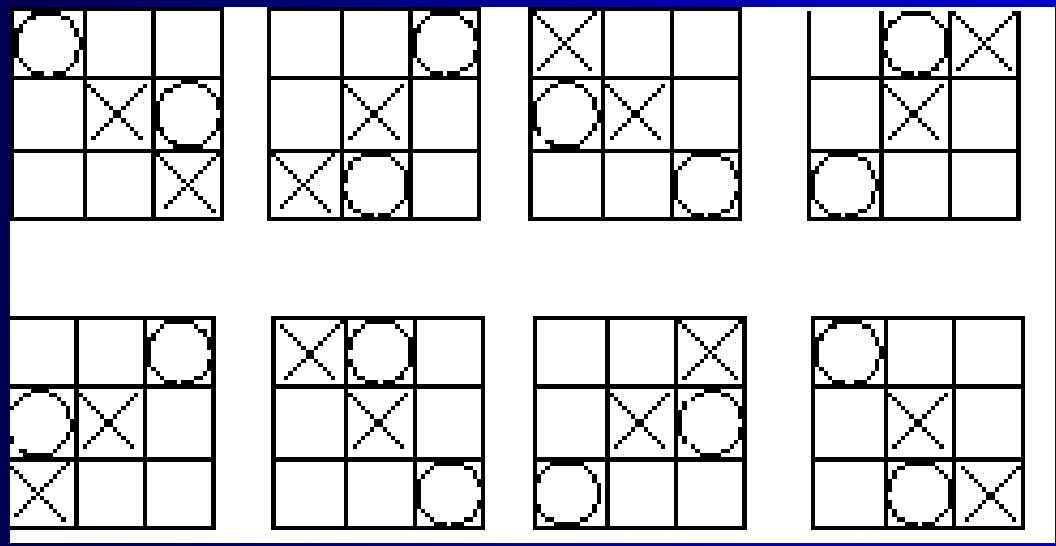
A2SI

RLtoolbox
Reinforcement Learning for SCILAB

V. Temporal Difference

Tic Tac Toe

Compression des états : *5890 -> 825*



$$S(X) = \sum_{i=1}^{9} 0.5^i . I_X(i)$$

$$S(O) = \sum_{i=1}^{9} 0.5^i . I_O(i)$$

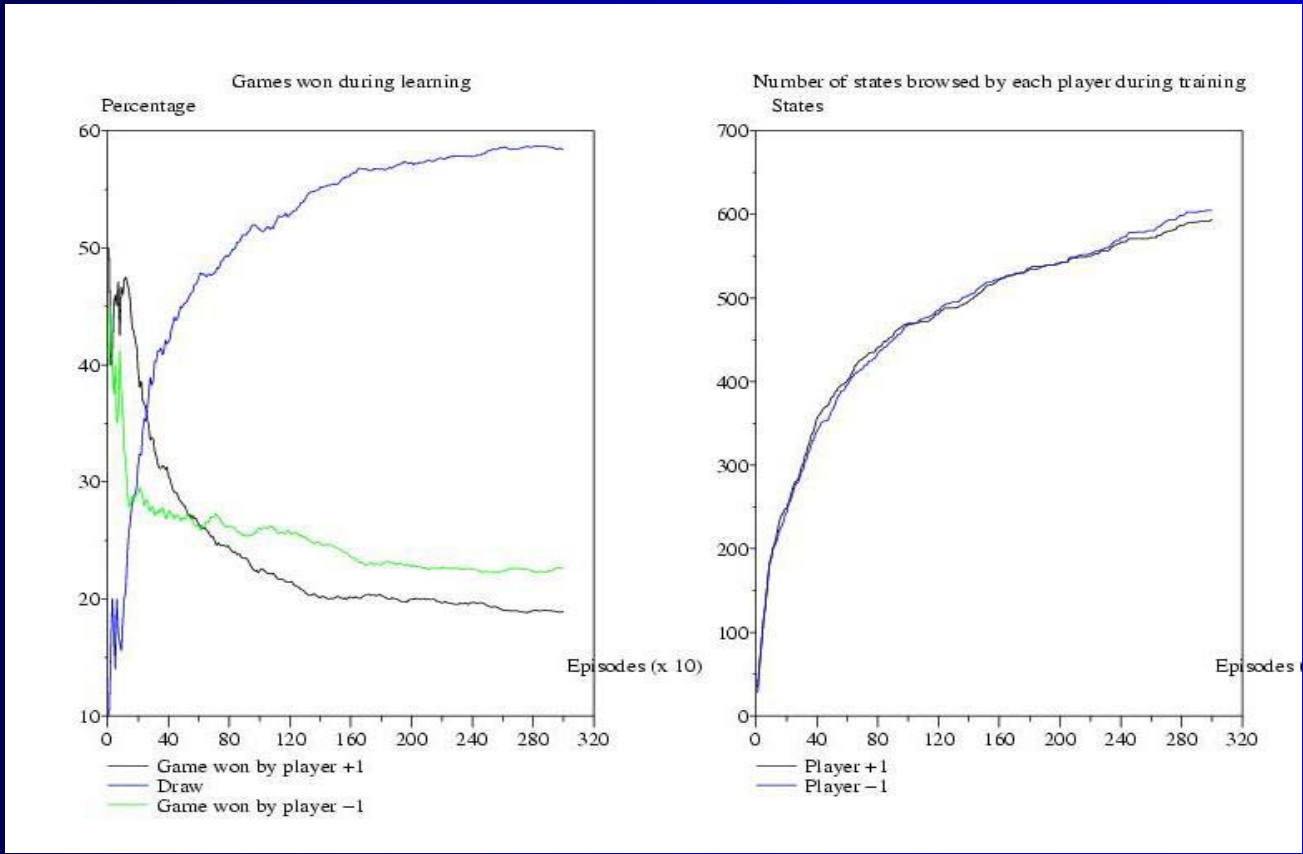
GROUPE ESIEE PARIS

A²SI

RLtoolbox
Reinforcement Learning for SCILAB

V. Temporal Difference

Tic Tac Toe



Entraînement sur 3000 parties

ESIEE PARIS

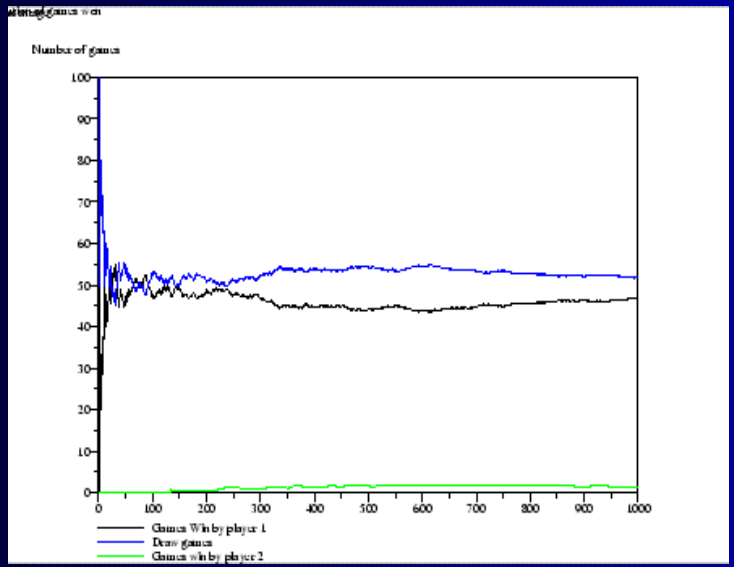
A²SI

RLtoolbox
Reinforcement Learning for SCILAB

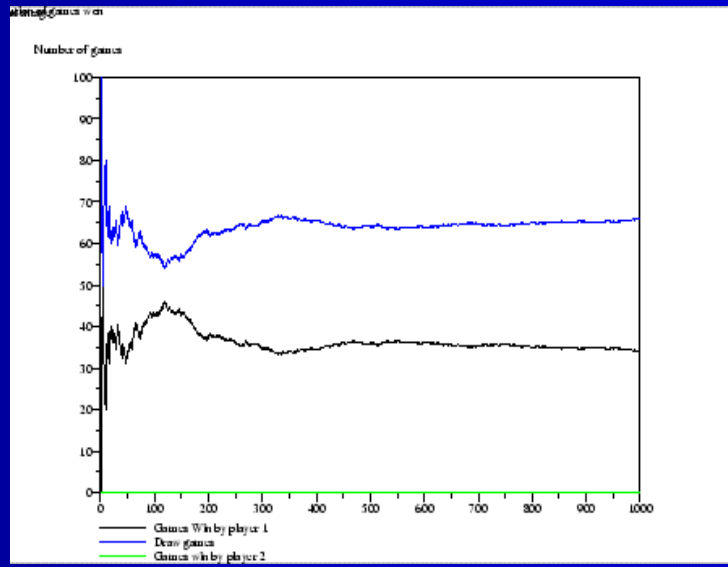
V. Temporal Difference

Tic Tac Toe :

Comparaisons d'agents



Joueur 1 : 10^4 parties, offensif
Joueur 2 : 10^2 parties, défensif



Joueur 1 : 10^4 parties, défensif
Joueur 2 : 10^2 parties, défensif

GROUPE ESIEE PARIS

A²SI