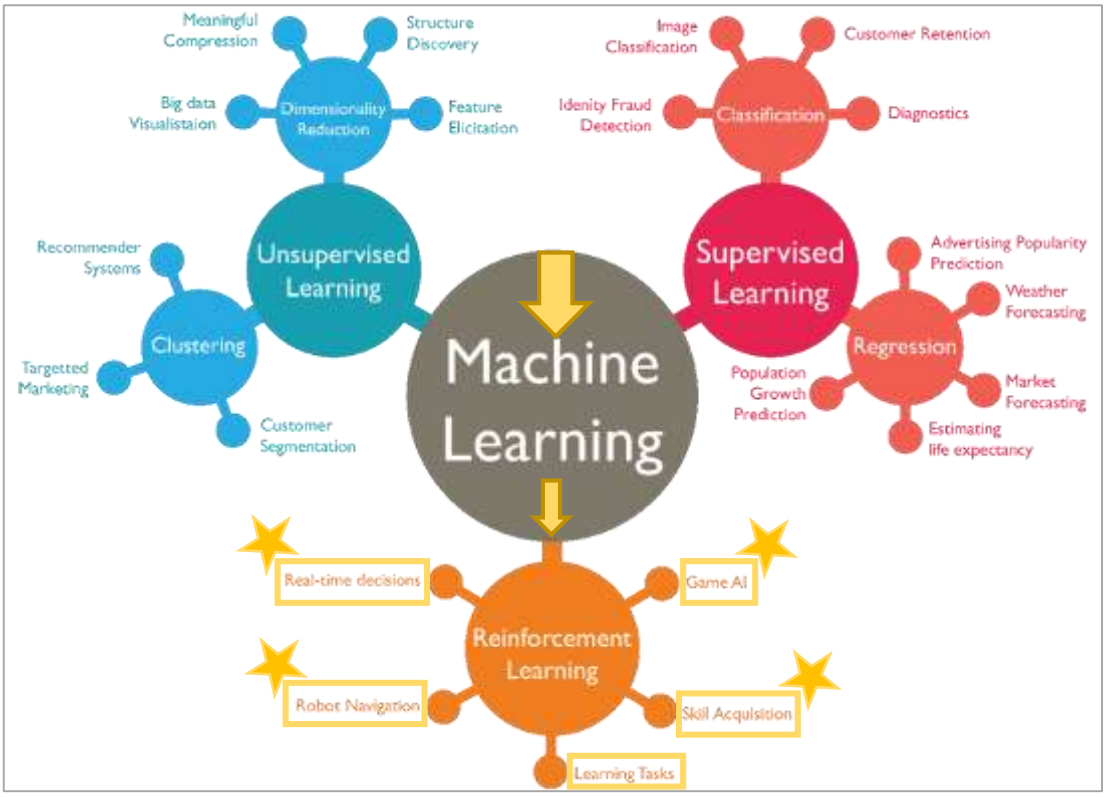
UVSQ

université PARIS-SACLAY

2ème Année Master Informatique, Semestre 2

Module : IA & JEUX

Lecture d'Article : GAMBLING THEORY



Rapport réalisé par :

SARAH OUHOCINE

DJILLALI BOUTOUILI

Professeur :

M. DAVID AUGER

Année universitaire 2022/2023

SOMMAIRE

1. INTRODUCTION

Pour la réalisation de notre travail, nous allons nous intéresser à la section 2 « *Application to Gambling Theory* » du chapitre 4 « *Maximizing Rewards* » du livre « *Introduction to Stochastic Dynamic Programming* » de Sheldon Ross.

Dans cette section, l'auteur essaye de résoudre des problèmes de décision pour maximiser les gains de joueurs dans les jeux de hasard. Dans ce contexte, on parle souvent de la « Gambling Theory ».

2. DÉFINITION DE LA « GAMBLING THEORY »

La « Gambling Theory » ou la théorie des **jeux de hasard** est une branche de la théorie des jeux, qui étudie des situations de décision stratégique de pari et de risque, où plusieurs acteurs doivent prendre des décisions qui influencent les résultats finaux. Elle est utilisée pour étudier les jeux de hasard, tels que les jeux de casino ou les loteries.

Dans cette section, nous allons étudier le jeu de hasard « Gambler's Ruin ».

3. PRÉSENTATION DU JEU « GAMBLER'S RUIN »

"Gambler's Ruin" est un modèle mathématique qui décrit une situation de jeu où un joueur dispose d'une certaine somme d'argent et doit prendre des décisions pour tenter de gagner davantage.

Dans ce modèle, le joueur peut soit gagner une certaine somme d'argent (avec une probabilité spécifique), soit perdre une certaine somme d'argent (avec une autre probabilité spécifique), et le jeu se poursuit jusqu'à ce que le joueur soit ruiné (c'est-à-dire qu'il ait perdu tout son argent) ou qu'il atteigne un objectif prédéfini.

Le modèle est souvent utilisé pour étudier des stratégies de jeu optimales, en particulier dans les **jeux de casino** où le joueur peut utiliser des systèmes de **mise** pour **maximiser ses gains** ou **minimiser ses pertes**.

Situation de jeu étudiée :

Un joueur **A** entre dans un casino de jeux :

Si **A** possède **i** dollars avec **i > 0**, alors
- **A** peut parier tout montant **j ≤ i**, avec **j > 0**

de plus :

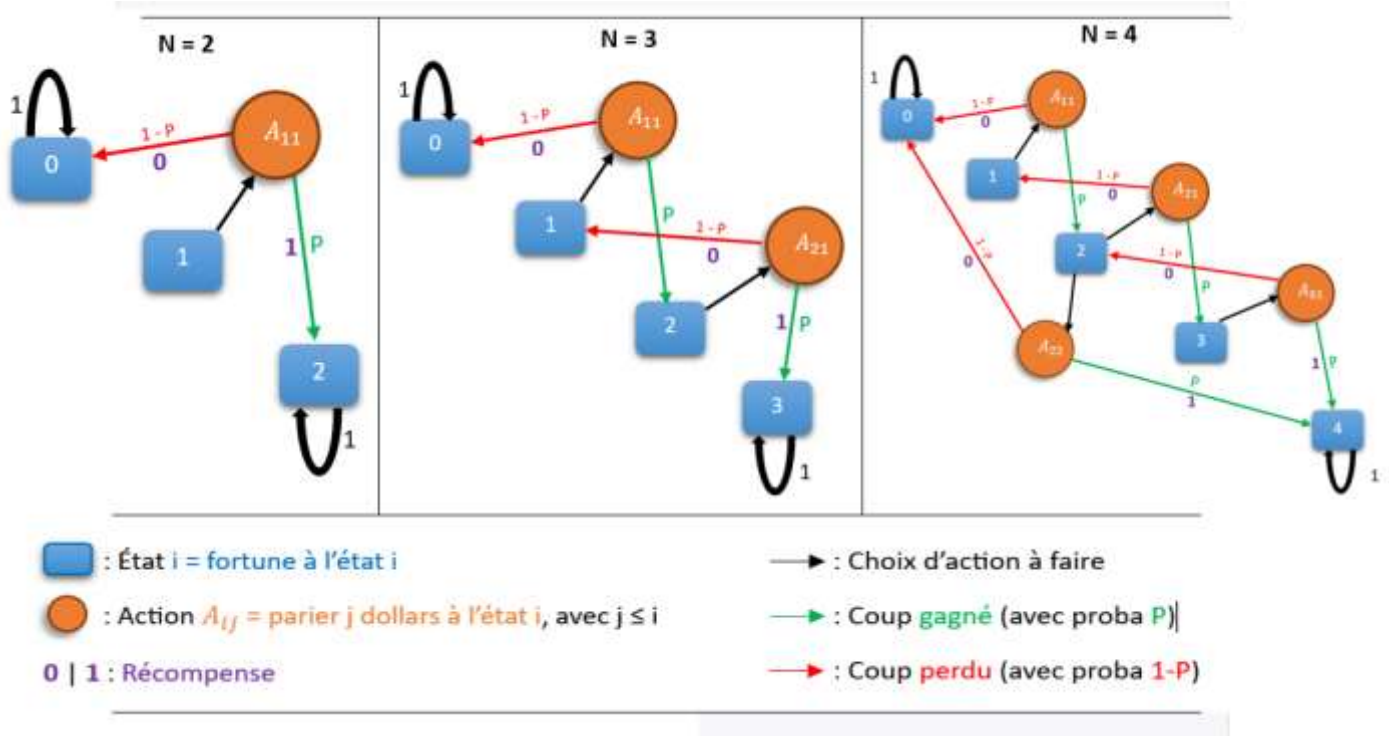
Si **A** parie **j**, alors soit il :
- **(a)** gagne **j** avec probabilité **p**
- **(b)** perd **j** avec probabilité **1-p**

On cherche quelle stratégie de jeu maximise la probabilité que le joueur A atteigne son objectif (une fortune = **N**) avant d'être ruiné (une fortune = **0**).

4. MODÉLISATION DU JEU

Le modèle considéré ici est un **Processus de Décision Markovien** (**MDP**), qui est un cadre mathématique utilisé pour la modélisation des situations où les résultats futurs dépendent des actions prises dans le présent, avec une incertitude quant à la manière dont l'environnement réagira à ces actions.

Le **MDP** associé à « Gambler's Ruin » est défini par **un ensemble d'états**, **un ensemble d'actions**, **des transitions** (probabilités de transition) entre les états et **des récompenses** (rewards) pour chaque état-action.



Exemple de MDP pour N= 2, 3, 4 : (avec **N** : l'objectif / fortune cible)

Généralisation de MDP pour N quelconque :

- ✓ **Espace d'états :** S = {0, 1, 2, ..., N} qui représente la fortune actuelle du joueur, avec 0 et N les états finaux (puits).

- ✓ **Espace d'actions de chaque état :** A(s) = {1, 2, …, s} pour chaque s appartenant à S.

- ✓ **Transitions :** sont déterminées par les règles du jeu, telles que les probabilités de gagner ou de perdre.
  - P(s - a | s, a) = 1 – p
  - P(s + a | s, a) = p
  - P(s | s, 0) = **1** avec s = **N** ou s = **0** (les boucles)

- ✓ **Récompenses (rewards) :** sont définies en fonction des gains ou des pertes de chaque tour de jeu.
  - R (s, a, s') = **1** si s' = N
  - R (s, a, s') = **0** sinon

En général, le but du joueur ici est d'atteindre une fortune cible N, tout en minimisant ses pertes. Dans ce cas, N est l'état final, et une récompense est attribuée uniquement lorsque le joueur atteint cet état final.

Une fois le MDP de Gambler's Ruin défini, définissons maintenant les politiques de ce jeu.

5. POLITIQUES DU JEU

Dans l'article, deux politiques sont définies pour le jeu « Gambler's Ruin » :

- ✓ **La politique Timide** (prudente).
- ✓ **La politique Bold** (audacieuse).

Politique	Principe	Avantages	Inconvénients
Timide	consiste à jouer de manière conservatrice et à ne parier qu'une petite partie de sa fortune à chaque fois. Par exemple, un joueur timide pourrait décider de parier seulement **10 %** de sa fortune à chaque tour de jeu.	-- le joueur met des petites sommes d'argent, ce qui réduit le risque de perdre rapidement (**efficace pour minimiser les pertes à court terme et maximiser les gains à long terme**). -- la politique est considérée comme moins risquée.	-- La politique est potentiellement **moins louable à court terme**. -- La politique peut être lente pour atteindre l'objectif final.
Bold	consiste à prendre des risques plus importants et à parier une plus grande partie de sa fortune à chaque tour de jeu. Par exemple, un joueur audacieux pourrait décider de parier **50 %** de sa fortune à chaque tour de jeu.	-- Le joueur a plus de chances de gagner rapidement car il met des montants plus élevés. -- La politique est potentiellement **plus louable à court terme**	-- La politique augmente le risque de perdre rapidement toute sa fortune et donc de ne pas atteindre l'objectif final. -- La politique est considérée comme plus risquée.

Le choix de la politique dépend des objectifs du joueur, de sa tolérance au risque et de sa stratégie à long terme.

6. POLITIQUES OPTIMALES

Dans l'article, la stratégie optimale dans le jeu de Gambler's Ruin, dépend de la valeur de **p**, la probabilité de gagner à chaque tour.

Politique Optimale	Quand ?	Exemple	Comment ?	Pourquoi ?
Timide	$P \geq ½$ (chances de gagner **élevées**)	P = 0.6 1-P = 0.4	Parier toujours **1$** jusqu'à ce que la fortune atteigne **N** ou **0**	Maximise les chances de gagner petit à petit (**à long terme**) et minimise les risques de perdre (**à court terme**).

Bold	P ≤ ½ (chances de gagner **faibles**)	P = 0.4 1-P = 0.6	Parier toujours la fortune actuelle ou une partie de celle-ci >> **min (i, N-i)**, ce qui permet d'atteindre **N** si on gagne.	Maximise les chances de gagner rapidement (**à court terme**) et d'atteindre son objectif avant de perdre tout son argent.

7. ÉVALUATION DES POLITIQUES

Dans l'apprentissage par renforcement, plusieurs algorithmes peuvent être utilisés pour évaluer une stratégie dans le but d'atteindre un certain objectif.

Ici, comme on a un MDP ou la seule chose que l'on puisse faire c'est de simuler des épisodes aléatoires. On va pouvoir **estimer les valeurs** en utilisant les méthodes d'échantillonnage.

Dans ce contexte nous avons choisi d'évaluer nos deux politiques (Timide et Bold) avec les deux méthodes d'échantillonnage (algorithmes de simulation) **vues en cours** :

- ✓ Monte-Carlo (méthode générale).
- ✓ TD-Learning (propre aux MDP).

7.1 ALGORITHMES UTILISÉS

1) Algorithme de **Monte-Carlo** (tous passages) qui converge plus rapidement pour un MDP :

Monte Carlo est une méthode basée sur la simulation aléatoire et elle permet de calculer les valeurs d'état et d'action en se basant sur l'échantillonnage de plusieurs épisodes du jeu. La méthode Monte Carlo permet de mettre à jour les valeurs de manière itérative en fonction des récompenses obtenues à chaque épisode. Elle est simple à implémenter et converge vers la solution optimale dans la plupart des cas.

Dans l'algorithme de Monte-Carlo, après chaque nouvelle réalisation de G_s pour un état, on met à jour V(s) :

$$G_s = \gamma * G_s + R$$
$$V(s) = V(s) + \alpha * (G_s - V(s))$$

R : Reward
α : pas d'apprentissage
γ : discount du MDP

2) Algorithme de **TD-Learning** :

TD Learning, quant à lui, est une méthode basée sur l'apprentissage par renforcement qui utilise des évaluations de valeurs d'état et d'action obtenues au cours des épisodes pour estimer les valeurs optimales. Cette méthode utilise des mises à jour basées sur une estimation de l'erreur de prédiction des valeurs actuelles par rapport aux valeurs cibles. TD Learning est simple à mettre en œuvre et peut être utilisé pour des problèmes à grande échelle.

Dans l'algorithme deTD-Learning, on utilise l'estimation de V pour S_{t+1} afin de mettre à jour l'estimation de V pour S_t par :

$$V(S_t) = V(S_t) + \alpha (R_{t+1} + \gamma * V(S_{t+1}) - V(S_t))$$

$R_{t+1} + \gamma * V(S_{t+1})$: Une réalisation de $V(S_t)$

$V(S_t)$: Ancienne estimation de $V(S_t)$

7.2 PROGRAMME (Cliquez ici)

Le code est commenté et est structuré comme suit :

1) Création de l’environnement du jeu : exemple de joueur.
2) Simulation d’un épisode avec la stratégie Bold.
3) Simulation d’un épisode avec la stratégie Timide.
4) Évaluation et comparaison des deux politiques (Bold et Timide) avec la méthode Monte-Carlo.
5) Évaluation et comparaison des deux politiques (Bold et Timide) avec la méthode TD-Learning.
6) Comparaison entre les deux méthodes Monte-Carlo et TD-Learning.
7) Influence de l’état initial sur les deux politiques (Bold et Timide).

7.3 RÉSULTATS

Les résultats, les graphiques ainsi que toutes les interprétations correspondantes sont présentés dans le Notebook avec le code, dans les sections de discussion situées à la fin de chaque partie.

8. CONCLUSION

En conclusion, le modèle MDP a été utilisé pour modéliser le jeu de Gambler's Ruin, permettant ainsi de déterminer les stratégies optimales pour le joueur. Nous avons exploré deux politiques différentes, la politique timide et la politique bold, et avons déterminé dans quelles circonstances chacune d'entre elles est optimale. Nous avons également évoqué l'utilisation de Monte Carlo et TD-learning pour apprendre la valeur des états et des actions dans le contexte de ce jeu.

L'analyse a montré que la stratégie optimale dépend fortement de la probabilité de gagner, avec la politique timide étant préférée dans les cas où cette probabilité est élévée et la politique bold étant préférée lorsque cette probabilité est faible. Nous avons également constaté que l'apprentissage par renforcement peut être utilisé pour déterminer la valeur des états et des actions dans le contexte de Gambler's Ruin.

En effet, la modélisation MDP et l'analyse des politiques dans le cadre du jeu de Gambler's Ruin ont fourni une illustration intéressante de l'application de la théorie de l'apprentissage par renforcement à un problème concret.

9. BIBLIOGRAPHIE

[1] : https://en.wikipedia.org/wiki/Gambling_and_information_theory

[2] : https://en.wikipedia.org/wiki/Gambler%27s_ruin

[3] : Temporal difference learning — Wikipédia (wikipedia.org)

[4] : Algorithme de Monte-Carlo — Wikipédia (wikipedia.org)