Rapport de Stage de L3

Léo Gayral

Étude en temps long des Processus de Décision Markoviens

Encadrants :

Bruno Ziliotto, CEREMADE, Université Paris Dauphine, brunoziliotto01@gmail.com

Xavier Venel, MSE, Université Paris 1 Panthéon-Sorbonne, xavier.venel@gmail.com

Université Paris-Dauphine

30 Mai 2016 — 10 Juillet 2016

Table des matières

Déroulement du stage

Le modèle de Processus de Décision Markovien décrit une situation où un individu fait face à un problème de décision répété : il doit à chaque étape choisir une action qui va influencer la transition de l'*état du monde* actuel vers l'état suivant. Le joueur reçoit en retour un paiement qui est fonction de l'état actuel et de l'action effectuée. Le but du joueur est d'adopter une stratégie lui permettant de maximiser son gain "global", une moyenne pondérée des paiements. L'objectif initial du stage était de lire et de comprendre les trois articles suivants :

— Blackwell [1], qui traite du cas où le preneur de décision observe parfaitement l'*état du monde*,
— Rosenberg, Solan et Vieille [2], qui traite du cas où le preneur de décision n'a qu'une observation partielle de l'état,
— Renault [3], qui traite le problème dans le cas où les ensembles d'actions, d'état et de signaux sont infinis.

L'étude des MDPs fait généralement appel aux équations de Bellman, un résultat classique dans ce domaine, qu'on peut trouver dans les notes de cours de Renault [4] par exemple. L'étude de ces différents articles, et de problèmes auxiliaires qu'ils ont soulevés, m'a occupé durant la grosse première moitié du stage.

Par la suite, mon travail s'est scindé en deux axes principaux. D'une part, une approche pratique du sujet, avec l'étude et l'implémentation de plusieurs algorithmes qui calculent de bonnes façons de jouer, des stratégies optimales. D'autre part, une approche plus théorique, avec l'étude de la vitesse de convergence du gain, et la recherche d'un POMDP à la convergence la plus lente possible, un problème actuellement ouvert.

Concernant l'étude d'algorithmes, tout est parti de l'article de Blackwell qui propose entre-autres un algorithme de construction de stratégie optimale en temps fini. J'ai implémenté cet algorithme, puis je me suis penché sur d'autres algorithmes proches à l'aide de *Discrete Stochastic Dynamic Programming* [5], un ouvrage assez complet sur ce qui existe dans la littérature mais avec une approche trop portée sur l'économie à mon goût. Pour élargir un peu le sujet, j'ai entamé la lecture du cours de Barto et Sutton [6], qui introduit d'autres classes d'algorithmes, dans le contexte général des Sciences Cognitives : la grande nouveauté est l'introduction des algorithmes de *learning* fonctionnels même sur un jeu où le joueur ne connait pas le modèle, le fonctionnement interne du jeu, a priori. C'est pendant cette lecture que mon stage s'est terminé. Dans le cadre d'un stage plus long, c'est sur cet aspect que j'aurais voulu passer plus de temps : implémenter proprement un des algorithmes de learning dans le cas des MDPs, au lieu de simplement jouer avec sur quelques exemples simples.

Au sujet des vitesses de convergence, sans entrer trop dans les détails, il s'agit de l'étude du comportement du gain en faisant tendre vers 0 le taux $\lambda \in]0;1]$, qui traduit la patience du joueur. Dans le cas des MDPs, on peut montrer que cette convergence est au pire en $O(\lambda)$ mais pour les POMDPs, le problème est ouvert. L'objet de mon étude personnelle a été d'expliciter un jeu à convergence aussi lente que possible : en me basant sur un exemple en $O(\sqrt{\lambda})$ proposé par mon tuteur, j'ai pu définir un exemple où la convergence est au mieux en $O(\sqrt{\lambda}\ln(\lambda))$.

J'ai choisi dans ce rapport de me concentrer sur les modèles et les résultats introduits par Blackwell [1] puis par Rosenberg [2], sans trop creuser les autres articles et sujets étudiés ; traiter ces deux articles en premier lieu est de toute façon nécessaire pour comprendre les résultats de Renault [3], le fonctionnement des algorithmes et le problème des vitesses de convergence. Compte tenu des contraintes de longueur, j'ai pris le parti d'éviter globalement les démonstrations à l'exception d'une principale, assez longue mais qui illustre différentes méthodes utilisées dans ce domaine.

Mes remerciements à Bruno Ziliotto et Xavier Venel pour m'avoir encadré et permis de réorienter la fin de mon stage sur les documents de mon choix.

1 Introduction

1.1 Motivation

Le modèle du Processus de Décision Markovien (*MDP*) a été l'objet central de mon stage. Cette section vise à introduire le modèle étudié ainsi que les problèmes d'optimisation associés, auxquels on répondra dans les parties suivantes.

Un MDP est, en résumé, un jeu à un seul joueur, à temps discret et avec observation totale. Il peut être décrit comme un ensemble A de transitions markoviennes, les *actions*, sur un ensemble d'états Ω (dans un cadre général, Ω doit être un espace probabilisé, mais on ne considère ici que des cas dénombrables où la tribu discrète est naturellement utilisée).

Au n-ième tour, en partant de l'état ω_n, une action a_n est choisie et permet d'atteindre un état ω_{n+1} suivant une loi aléatoire $q(\omega_n, a_n) \in \Delta(\Omega)$, où $\Delta(X)$ désigne les mesures de probabilités sur un espace probabilisé X.

Effectuer une action a en partant d'un état ω donne lieu à une certaine récompense $r(\omega, a) \in [0,1]$. Pour le joueur, le *preneur de décision*, le but est alors d'adopter une stratégie σ lui permettant de maximiser son gain global, fonction des paiements $r(\omega_n, a_n)$ à chaque étape du jeu.

Ce modèle, ainsi que les algorithmes et méthodes d'optimisation qui en sont issus, ont des applications dans de multiples domaines. Son lien intime avec la théorie des jeux en fait un objet de choix des économistes, mais ont peut également trouver ce modèle au cœur de la Robotique et du *Reinforcement Learning* [6], en Sciences Cognitives.

1.2 Exemple introductif

Pour commencer, considérons un exemple déterministe simple à concevoir : un système de pension de retraite. On cherche à modéliser un système où le joueur travaille pendant un certain nombre de tours, puis prend sa retraite avant de recevoir un paiement pendant autant de tours qu'il a travaillé ; une fois sa retraite prise, le joueur n'a alors plus la possibilité de travailler à nouveau.

Soit $\Omega = \mathbb{N}^2$. Pour $\omega = (i, j)$, i traduit le temps passé à travailler et j le temps écoulé depuis le début de la retraite. L'état initial est assez naturellement $\omega_1 = (0,0)$. On définit $A = \{a_t, a_r\}$ avec les transitions déterministes $q(i,0,a_t) = (i+1, 0)$ (continuer à travailler) et $q(i,0,a_r) = (i,1)$ (prendre sa retraite). Une fois la retraite prise, quand $j \geq 1$, on a la transition $q(i,j,a) = (i, j+1)$ indépendamment de l'action a puisqu'on ne peut plus travailler.

On a alors la loi de paiement r suivante :

$$r(\omega, a) = \begin{cases} 1 & \text{si } 1 \leq j < i \\ 1 & \text{si } j = 0, i \geq 1 \text{ et } a = a_r \\ 0 & \text{sinon} \end{cases}$$

On reviendra sur cet exemple après avoir présenté les principaux résultats de Blackwell [1] dans la seconde section.

1.3 Notations générales

Par la suite, on aura toujours Ω l'ensemble des états, A l'ensemble des actions, $q : \Omega \times A \to \Delta(\Omega)$ la loi de transition et $r : \Omega \times A \to [0,1]$ le paiement immédiat. On note alors $\Gamma = (\Omega, A, q, r)$ le MDP associé.

Les variables aléatoires considérées sont ω_n l'*état du monde* au tour $n \in \mathbb{N}^*$ et a_n l'action effectuée du tour n au tour $n+1$. En particulier, l'état initial du jeu $\omega_1 \in \Omega$ est fixé.

1.4 Stratégies

Posons $H_n = (\Omega \times A)^{n-1} \times \Omega$ l'ensemble des histoires du jeu au temps n et $H = \cup_{n\geq 1}(H_n)$ l'ensemble de toutes les histoires du jeu. $h_n = (\omega, 1, a_1, ..., \omega_{n-1}, a_{n-1}, \omega_n)$ est alors la variable aléatoire à valeurs dans H_n, représentant l'histoire du jeu au tour n.

Dans un cadre général, une stratégie comportementale est une application $\sigma : H \to \Delta(A)$ qui, à partir de toute l'information disponible à un instant n, fait un choix (éventuellement aléatoire) sur l'action a_n à entreprendre.

σ est dite markovienne si elle ne dépend que de l'état en cours et du moment de la décision, c'est-à-dire $\sigma = (f_n) \in (\Omega \to \Delta(A))^{\mathbb{N}}$. On dit que $\sigma : H \to A$ est une stratégie pure (l'action est déterminée de façon certaine) et que $\sigma : \Omega \to \Delta(A)$ est stationnaire (l'effet de la stratégie ne dépend que de l'état actuel ω_n, pas du temps écoulé ou du reste de l'histoire du jeu).

1.5 Gains, valeurs et stratégies optimales

On va par la suite s'intéresser à deux définitions du gain. La première notion de gain ne fait appel qu'à des moyennes finies, ce qui la rend plus visuelle. La seconde considère tous les paiements reçus, tout en accordant plus d'importance aux premiers paiements.

1.5.1 Gain moyen

Pour un état de départ ω_1 et une stratégie σ, on définit le gain moyen sur les n premières étapes par l'espérance conditionnelle :

$$\gamma_n(\sigma, \omega_1) = \mathbb{E}_{\omega_1,\sigma}\left[\frac{1}{n}\sum_{k=1}^{n} r(\omega_k, a_k)\right]$$

On peut de façon analogue définir le gain moyen $\gamma_{i,j}$ entre les instants i et j (inclus).

On définit la *valeur* du jeu $\Gamma_n(\omega)$ comme le gain (moyen) maximal possible sur les n premières étapes. Formellement :

$$v_n(\omega) = \sup_{\sigma} \gamma_n(\sigma, \omega)$$

On dit alors de σ qu'elle est ϵ-*optimale* dans $\Gamma_n(\omega)$ lorsque :

$$\gamma_n(\sigma, \omega) \geq v_n(\omega) - \epsilon$$

On peut élargir cette notion au comportement asymptotique. On dit alors que σ est ϵ-optimale dans le problème asymptotique $\Gamma_\infty(\omega)$ lorsqu'elle l'est dans tous les Γ_n à partir d'un rang, c'est-à-dire :

$$\exists n_0, \forall n \geq n_0, \gamma_n(\sigma, \omega) \geq v_n(\omega) - \epsilon$$

On dit que $\Gamma_\infty(\omega)$ a la valeur limite si la suite $(v_n(\omega))$ converge, et on note alors $v_\infty(\omega)$ cette limite.

On dit enfin que σ est *asymptotiquement* ϵ-*optimale* dans $\Gamma_\infty(\omega)$ lorsque elle est δ-optimale dans ce problème pour tous les δ tels que $\delta > \epsilon$.

1.5.2 Gain escompté

La notion précédente est assez visuelle car elle se base sur un *horizon* fini, les paiements dans l'espérance sont pondérés par une loi de probabilités sur $[\![1;n]\!]$. Le gain (escompté) au taux $\lambda \in]0;1]$ est un gain à horizon infini, une somme pondérée sur $\mathbb{N}$ tout entier. Ce gain est défini par :

$$\gamma_\lambda(\sigma,\omega_1) = \mathbb{E}_{\omega_1,\sigma}\left[\lambda\sum_{k=1}^{\infty}(1-\lambda)^{k-1}r(\omega_k,a_k)\right]$$

Le taux λ traduit la patience du joueur : on peut interpréter le facteur $\lambda(1-\lambda)^{k-1}$ comme la probabilité que le jeu se termine à l'étape k. Ce faisant, le joueur perd patience avec le temps dans tous les cas mais plus λ est proche de 0, plus il sera prêt à attendre longtemps pour recevoir une récompense.

On définit alors la valeur de $\Gamma_\lambda(\omega)$ comme $v_\lambda(\omega) = \sup_\sigma \gamma_\lambda(\sigma,\omega)$. Ici, σ est ϵ-optimale dans le problème asymptotique $\Gamma_0(\omega)$ lorsque :

$$\exists\lambda_0 > 0, \forall 0 < \lambda < \lambda_0, \gamma_\lambda(\sigma,\omega) \geq v_\lambda(\omega) - \epsilon$$

Sous réserve d'existence, la valeur limite dans Γ_0 est notée $v_0(\omega) = \lim_{\lambda\to 0} v_\lambda(\omega)$.

Les autres notions d'optimalité se transposent sans plus d'efforts dans ce contexte.

Par la suite, on utilisera "optimal" pour dire "0-optimal" dans toutes les définitions précédentes.

1.6 Théorème de Kuhn

Le théorème de Kuhn est un résultat classique de la théorie des jeux et s'applique dans un cadre bien plus général que celui des MDPs. On peut le retrouver p.1181 de Rosenberg [2]. Ce résultat donne une équivalence entre plusieurs types de stratégies envisageables :

Pour toute stratégie comportementale σ, il existe π une mesure de probabilités sur A^H (l'espace des stratégies pures, ici muni de la tribu discrète puisqu'on considère A fini et H dénombrable) induisant la même probabilité que σ. On dit que π est une stratégie *mixte*.

On peut écrire ce résultat synthétiquement sous la forme :

$$\mathbb{P}_\sigma = \int_{A^H} \mathbb{P}_f \,\mathrm{d}\pi(f)$$

Dans un cadre assez général, $\mathbb{P}_\sigma$ désigne la loi de probabilité conditionnelle sachant que le joueur suit la stratégie σ (mais dont l'état initial reste encore à préciser). Ce résultat est également vrai dans le sens réciproque : $\forall\pi \in \Delta(A^H), \exists\sigma \in \Delta(A)^H, \mathbb{P}_\sigma = \int \mathbb{P}_f \,\mathrm{d}\pi(f)$.

2 Existence et construction de stratégies optimales

Le problème central des MDPs est celui de l'existence de stratégies optimales (éventuellement asymptotiquement ou à ϵ près) pour une classe de jeux donnée. Ces résultats ont un intérêt théorique, mais pour les applications pratiques il est également pertinent d'exposer des méthodes de construction de telles stratégies.

Dans toute cette section, on s'intéresse aux résultats exposés par Blackwell [1] : sauf mention explicite, on se limite au cas où Ω et A sont des ensembles finis.

2.1 Stratégies optimales dans Γ_n et équations de Bellman

Les équations de Bellman sont un des résultats fondamentaux des MDPs. Elles se présentent sous de nombreuses formes assez similaires dans de multiples contextes, comme dans le cas des *Gambling Houses* (où les actions sont toutes déterministes) dans les notes de cours de J. Renault [4].

2.1.1 Équation de Bellman dans Γ_n

Avec l'initialisation $v_1(\omega) = \sup_{a\in A} r(\omega,a)$ qui découle de la définition, les valeurs des problèmes Γ_n sont entièrement définies par la relation de récurrence :

$$(n+1)\times v_{n+1}(\omega) = \sup_{a\in A}\left(r(\omega,a) + n\times \sum_{\tilde{\omega}\in\Omega} q(\omega,a,\tilde{\omega})\times v_n(\tilde{\omega})\right)$$

De façon informelle, ce résultat traduit le fait que maximiser le gain sur les $n+1$ premiers instants revient à maximiser la somme du gain issu de la première transition, donné par r, avec le gain maximal sur les n étapes suivantes, donné par v_n.

Ce résultat peut se généraliser dans un cadre plus large que celui considéré ici, en remplaçant les sommes par des intégrales sur des espaces probabilisés, mais il faudrait alors définir ou donner des hypothèses sur cet espace, ce qu'on évite ici en étudiant le cas des espaces dénombrables.

Dans le cas où A est fini (mais pas nécessairement Ω), les sup sont atteints et il en découle directement l'existence d'une stratégie pure optimale dans Γ_n.

2.1.2 Démonstration

Soient σ une stratégie comportementale et ω l'état initial considéré. On remarque que $\sigma(\omega,a,\bullet) : h \mapsto \sigma(\omega,a,h)$ définit une stratégie comportementale. On a alors la relation suivante :

$$\begin{aligned}
(n+1)\gamma_{n+1}(\sigma,\omega) &= \sum_{a\in A}\sigma(\omega)(a)\times\left(r(\omega,a) + \sum_{\tilde{\omega}\in\Omega} q(\omega,a,\tilde{\omega})\times n\times\gamma_n(\sigma(\omega,a,\bullet),\tilde{\omega})\right)\\
&\leq \sup_{a\in A}\left(r(\omega,a) + n\times\sum_{\tilde{\omega}\in\Omega} q(\omega,a,\tilde{\omega})\times\gamma_n(\sigma(\omega,a,\bullet),\tilde{\omega})\right)\\
&\leq \sup_{a\in A}\left(r(\omega,a) + n\times\sum_{\tilde{\omega}\in\Omega} q(\omega,a,\tilde{\omega})\times v_n(\tilde{\omega})\right)
\end{aligned}$$

Ceci étant vrai pour toute stratégie, on en déduit une première inégalité. Réciproquement, soient $\epsilon > 0$ et $(\sigma_{\tilde{\omega}})_{\tilde{\omega}\in\Omega}$ une famille telle que $\sigma_{\tilde{\omega}}$ est $\frac{\epsilon}{n}$-optimale dans $\Gamma_n(\tilde{\omega})$. Pour toute action $a\in A$, on peut définir une stratégie σ par $\sigma(\omega) = a$ sur H_1 et $\sigma(\omega,a,h) = \sigma_{\tilde{\omega}}(h)$ où $\tilde{\omega}$ est l'état initial de l'histoire $h\in\cup_{n\geq 2}H_n$. Pour cette stratégie, on a alors :

$$\begin{aligned}
(n+1)\gamma_{n+1}(\sigma,\omega) &= r(\omega,a) + n\times\sum_{\tilde{\omega}\in\Omega} q(\omega,a,\tilde{\omega})\times\gamma_n(\sigma(\omega,a,\bullet),\tilde{\omega})\\
&= r(\omega,a) + n\times\sum_{\tilde{\omega}\in\Omega} q(\omega,a,\tilde{\omega})\times\gamma_n(\sigma_{\tilde{\omega}},\tilde{\omega})\\
&\geq \left(r(\omega,a) + n\times\sum_{\tilde{\omega}\in\Omega} q(\omega,a,\tilde{\omega})\times v_n(\tilde{\omega})\right) - \epsilon
\end{aligned}$$

En passant au sup à gauche (la stratégie σ dépend du ϵ considéré, on doit l'éliminer en premier) puis en passant à la limite $\epsilon\to 0$, on en déduit l'autre inégalité donc l'égalité voulue.

2.1.3 Équation de Bellman dans Γ_λ

La principale différence avec le cas précédent est qu'on n'a plus une simple relation de récurrence sur les sommes finies γ_n. En suivant les mêmes idées, on en déduit ici que v_λ vérifie une relation de type point-fixe :

$$v_\lambda(\omega) = \sup_{a\in A}\left(\lambda \times r(\omega,a) + (1-\lambda)\sum_{\tilde{\omega}\in\Omega} q(\omega,a,\tilde{\omega}) \times v_\lambda(\tilde{\omega})\right)$$

Lorsque A et Ω sont finis, en utilisant le théorème du point fixe de Banach sur l'opérateur défini par l'équation précédente, on peut de plus conclure que cette équation fonctionnelle caractérise v_λ, qui est donc l'unique application réelle et bornée sur Ω qui la vérifie.

2.2 Stratégies optimales dans $\Gamma_\lambda, \lambda > 0$

On va dans cette partie chercher à construire une stratégie *pure stationnaire* et optimale dans Γ_λ, c'est-à-dire optimale dans tous les $\Gamma_\lambda(\omega_1)$.

L'existence de stratégies pures stationnaires optimales dans Γ_λ découle de l'équation de Bellman ci-dessus, où il suffit de prendre un a qui réalise le sup pour chaque ω. Cependant, il n'est pas possible de calculer directement la valeur v_λ sans disposer de la stratégie qu'on cherche.

Une méthode de *brute force* sur l'espace des stratégies stationnaires pures, $F = A^\Omega$, viendrait à bout du problème mais on peut faire mieux. Par la suite, on va démontrer un algorithme de recherche de point fixe sur l'espace F, fini, qui permettra une convergence "rapide" vers une stratégie optimale.

Il est important de noter qu'on a rarement l'unicité d'une telle stratégie f. D'une part, A n'est pas un ensemble de fonctions mais de "noms" d'actions dont les effets sont donnés par q. On peut donc avoir plusieurs actions qui agissent strictement de la même façon sur un état, et si une parmi elles constitue un choix optimal dans Bellman sur l'état ω, alors on peut choisir une des autres indifféremment comme action dans cette stratégie optimale. D'autre part, deux stratégies (optimales ou non) peuvent adopter des actions différentes mais donner lieu à un même gain total en suivant deux dynamiques différentes.

Dans la suite de cette seconde partie, on travaille sur l'espace des stratégies *pures markoviennes* ($\sigma = (f_n) \in F^{\mathbb{N}^*}$) pour montrer l'algorithme, mais l'équation de Bellman nous assure que la stratégie obtenue est optimale en tant que stratégie comportementale.

2.2.1 Résultats préliminaires

Soient $R(f) = (r(\omega, f(\omega)))_{\omega\in\Omega}$ le gain associé à $f \in F$ et $Q(f) = (q(\tilde{\omega}, f(\tilde{\omega}), \omega))_{\tilde{\omega},\omega\in\Omega}$ sa matrice de transitions. On peut écrire le gain au taux λ de $\sigma = (f_n)$ sous la forme :

$$\gamma_\lambda(\sigma) = (\gamma_\lambda(\sigma,\omega))_{\omega\in\Omega} = \lambda\sum_{n=1}^{\infty}(1-\lambda)^{n-1} \times Q(f_1) \times \ldots \times Q(f_{n-1}) \times R(f_n)$$

Soit $L_\lambda(f) : w \mapsto \lambda R(f) + (1-\lambda)Q(f)\times w$ définie sur $\mathbb{R}^\Omega$. On remarque alors que $\gamma_\lambda((f,\sigma)) = L_\lambda(f)(v_\lambda(\sigma))$. On peut de la même façon définir R, Q et L pour une action a fixée.

Si pour tout $f \in F$ on a $L(f)(\gamma_\lambda(\sigma^*)) \leq \gamma_\lambda(\sigma^*)$ (inégalité large sur chaque coordonnée des vecteurs), alors σ^* est optimale dans Γ_λ. Au contraire, s'il existe $f \in F$ telle que $\gamma_\lambda((f,\sigma)) > \gamma_\lambda(\sigma)$, alors on a $\gamma_\lambda(f) > \gamma_\lambda(\sigma)$ (inégalité large et vecteurs distincts) et donc σ non optimale.

D'autre part, $\gamma_\lambda(f)$ vérifie $L_\lambda(f)(\gamma_\lambda(f)) = \gamma_\lambda(f)$, qu'on peut reformuler en $(I-(1-\lambda)Q(f))\gamma_\lambda(f) = R(f)$. Sachant la matrice Q stochastique, on a alors $I-(1-\lambda)Q$ inversible : ce système définit entièrement $\gamma_\lambda(f)$ de façon calculable en temps polynômial $O(|\Omega|^3)$.

On pourra trouver une esquisse des démonstrations de ces résultats et du théorème suivant aux pages 720-721 de Blackwell [1].

2.2.2 Théorème

Soit $G(\omega, f) := \{a \in A, L(a)(\gamma_\lambda(f))_\omega > \gamma_\lambda(f)_\omega\}$. On peut voir G comme l'ensemble des actions qui offrent un meilleur gain que f sur ω.

Si pour tout $\omega \in \Omega$ on a $G(\omega, f) = \emptyset$, alors f est optimale dans Γ_λ.

A contrario, pour $g \in F$ fixée, s'il existe $\omega \in \Omega$ tel que $g(\omega) \in G(\omega, f) \neq \emptyset$ et que pour tout $\omega \in \Omega$ on a $g(\omega) \in G(\omega, f) \cup \{f(\omega)\}$, alors g offre un meilleur gain que f, $v_\lambda(g) > v_\lambda(f)$.

En itérant cet algorithme sur l'ensemble F fini, on en déduit qu'il termine en donnant une stratégie pure stationnaire optimale dans Γ_λ. Cette méthode, appelée Policy Iteration (*PI*) dans la littérature, est un des algorithmes de résolution de MDP fondamentaux.

2.2.3 Algorithme PI

Afin de confronter les résultats du théorème précédent avec mes calculs dans des exemples simples, j'ai implémenté l'algorithme en python dans le cas où on parcourt les états de façon cyclique, et en améliorant la stratégie f dès que possible :

```
from itertools import izip
from numpy import matrix, identity
from random import choice
from numpy.linalg import solve
# max_arg renvoie le maximum d'une liste et son plus petit index
max_arg = lambda array: max(izip(array, xrange(len(array))))
def Q(d,A,S): # Matrice Q(d)
    return matrix( [ [A[d[s]][s][j] for j in xrange(S)] for s in xrange(S)] )
def Q_n(a,s,A,S): # Ligne s d'une matrice Q(d)
    return matrix([ A[a][s][t] for t in xrange(S)])
def R(d,r,S): # Vecteur de gain R
    return matrix([ r(s,d[s]) for s in xrange(S) ]).T
def PI(A,r,lam):
    # r[omega,a] le gain
    # lam le taux lambda
    # A la liste des actions: A[a][s][t] = P( omega_{n+1}=t | omega_n=s , a_n=a )
    card_A = len(A)
    S = len(A[0]) # S le nombre d'etats du jeu
    # d une strategie initiale choisie au hasard
    d = [ choice(xrange(card_A)) for s in xrange(S) ]
    stop = 0
    while True :
        if stop == 0 : # Calcul de v = v_\lambda( d^\infty )
            v = solve( ( identity(S) - (1-lam) * Q(d,A,S) ) , R(d,r,S) )
            L = [ ( r(s,a) + lam * Q_n(a,s,A,S) * v ) for a in xrange(card_A) ]
            max_L,ind_L = max_arg(L)
        if max_L != L[ d[s] ] : # On ameliore d si possible
            d[s] = ind_L
            stop = 0
        else : # Sinon, on augmente le compteur d'arret
            stop += 1
        if stop == S : # Si stop vaut S, on ne peut plus ameliorer d
            break
        s = (s+1) % S
    return d
```

L'autre grande classe d'algorithmes de résolution, les *Value Iteration*, adopte également un procédé itératif mais en itérant cette fois-ci sur des gains qui croissent vers la valeur du jeu. J'ai choisi de traiter la méthode PI, introduite par Blackwell, car la terminaison de l'algorithme sur une stratégie optimale est assez instinctive, mais j'ai implémenté et testé plusieurs variations de ces deux méthodes, présentées dans l'ouvrage de Puterman [5].

Aucune des deux méthodes n'est fondamentalement meilleure que l'autre, et on peut les combiner entre elles et à d'autres méthodes pour approcher des valeurs et stratégies optimales dans d'autres types de jeux, dans

lesquels on n'a pas accès à la loi de transition q, où le gain r est lui-même aléatoire pour ω et a donnés, etc. Ces idées sont exposées dans *Reinforcement Learning : An Introduction* [6] qui introduit au passage certains enjeux des sciences cognitives, et dont j'ai commencé la lecture dans le cadre de mon stage. Parmi les algorithmes proposés, on pourra citer l'*Action-Value Method* qui correspond à une stratégie asymptotiquement ϵ-optimale dans ce contexte plus large, pour ϵ initialement fixé.

2.3 Stratégies optimales dans Γ_0

Avec des arguments similaires à ceux ci-dessus (p. 723-725 [1]), on peut également obtenir les résultats suivants :

— Γ_0 a une valeur limite (pour tout $\omega \in \Omega$, $v_0(\omega) = \lim_{\lambda \to 0} v_\lambda(\omega)$ existe).
— Il existe $f^* \in F$ une stratégie pure stationnaire optimale dans Γ_0.
— Il existe un procédé de type PI permettant de calculer la valeur limite de Γ_0 et de construire une stratégie asymptotiquement optimale.

En se ramenant à une Gambling House équivalente au MDP considéré, avec des outils topologiques plus puissants, Renault [3] montre que si $|\Omega| < \infty$ et $A \neq \emptyset$ quelconque, alors $\forall \omega \in \Omega$, $\Gamma_0(\omega)$ a une valeur limite et admet, pour tout $\epsilon > 0$, une stratégie ϵ-optimale. On perd cependant la méthode de construction des solutions du problème plus restreint traité par Blackwell.

Un des intérêts de l'approche de Blackwell est que pour une stratégie pure stationnaire σ, elle permet d'exprimer les différents coefficients de $\lambda \mapsto \gamma_\lambda(\sigma)$ sous la forme de fractions rationnelles. En considérant une telle stratégie optimale, on en déduit $\|v_\lambda - v\| = O_{\lambda \to 0}(\lambda)$: on a convergence linéaire puisqu'une fraction rationnelle admet un développement de Taylor en tout point qui n'est pas un pôle. Ce résultat de convergence linéaire est assez fort, et on verra plus tard qu'il ne se généralise pas aux MDPs à observation partielle (*POMDPs*).

2.4 Retour sur l'exemple introductif

On considère à nouveau notre système de pension de retraite où le joueur travaille pendant un certain nombre de tours puis prend sa retraite avant de recevoir un paiement pendant autant de tours. Ce modèle m'est venu assez naturellement lorsque mon tuteur m'a demandé un contre-exemple à une généralisation des résultats de Blackwell au cas où Ω est infini.

Dans ce jeu, pour calculer v_λ, il suffit de maximiser γ_λ sur les stratégies pures stationnaires. Une telle stratégie est ici équivalente à un parcours déterministe σ_N de $\mathbb{N}^2$ suivant l'axe "travail" pendant N étapes puis l'axe "retraite" pour l'éternité qui s'ensuit.

$$\gamma_\lambda(\sigma_N) = \lambda \sum_{i=N+1}^{2N} (1-\lambda)^{i-1} = (1-\lambda)^N \times (1-(1-\lambda)^N)$$

Une étude plus poussée des fonctions $f_\lambda : x \mapsto (1-\lambda)^x(1-(1-\lambda)^x)$ permet de conclure que $\lim_{\lambda \to 0} v_\lambda(\omega_1) = \frac{1}{4}$. Cependant, pour σ_N une stratégie pure fixée, $\lim_{\lambda \to 0} \gamma_\lambda(\sigma_N) = 0$. En utilisant le théorème de Kuhn, on peut de la même façon montrer que pour toute stratégie comportementale, $\gamma_\lambda(\sigma) \to 0$. Dans ce jeu où Ω est infini, malgré la convergence des valeurs du jeu en temps longs, les résultats de Blackwell ou Renault ne s'appliquent pas et le jeu n'admet pas de stratégies asymptotiquement optimales dans Γ_0, ni même ϵ-optimales pour $\epsilon > 0$ assez faible.

3 Généralisation du modèle aux MDPs partiellement observables

Les MDPs sont certes intéressants à étudier, mais dans bon nombre d'applications le système n'est que partiellement observable et ne peut être modélisé par un MDP. Une voiture, par exemple, ne donne pas accès à

son état exact au conducteur : ce dernier n'a accès qu'à certains signaux comme le compteur kilométrique ou la jauge d'essence. Nous sommes de nos jours entourés de systèmes qui se modélisent mieux par des POMDPs que des MDPs.

Un autre exemple, omniprésent dans notre paysage ludique, est l'écran d'un jeu vidéo : le joueur reçoit une image, une grille d'informations réparties sur un certain nombre de pixels. Cette image est le signal sur lequel il se base pour réagir, mais dans les entrailles de l'ordinateur l'état du jeu n'est pas limité à ces données. Par exemple, si on considère une simple *frame* du jeu *Pong*, un seul instant, on peut déterminer la position exacte des raquettes et de la balle, ainsi que le score actuel, mais on ne peut en aucun cas déterminer dans quel sens ou à quelle vitesse se déplace la balle.

On cherche alors à mettre en œuvre des stratégies de jeu qui font face efficacement à cette incertitude, qui maximisent le gain en se basant sur la croyance qu'on a sur l'état occupé à chaque instant. Les résultats de cette section sont principalement issus de l'article de Rosenberg, Solan et Vieille [2].

3.1 Modèle du POMDP

On va par la suite s'intéresser à l'étude des γ_n plutôt qu'à celle des γ_λ comme le faisait Blackwell. Une raison est sans doute le fait que ces sommes finies sont souvent plus simple à manipuler. Cependant, un théorème taubérien (p.1181 [2]) permet, dans le contexte des résultats d'existence d'une valeur limite et de stratégies optimales ci-dessous, d'étendre ces résultats dans Γ_∞ au problème Γ_0. En particulier, si Γ_0 ou Γ_∞ a la valeur limite, l'autre problème a la même valeur limite qu'on peut alors noter v.

3.1.1 Notations

Commençons par adapter les notations des parties précédentes au modèle des *Partially Observable Markov Decision Processes* présenté dans l'article de Rosenberg, Solan et Vieille [2]. La principale nouveauté est S l'ensemble des signaux. A chaque étape du jeu, le joueur va observer un signal au lieu d'un état, et c'est ce signal qui va lui fournir une information sur l'évolution de l'état.

Comme précédemment, sauf contre-indications, les ensembles Ω, A et S sont supposés finis.

La loi de transition devient ici $q : \Omega \times A \to \Delta(\Omega \times S)$. L'état initial devient une distribution de probabilités $x_1 \in \Delta(\Omega)$, ce qui ramène les problèmes d'optimisation sous la forme $\Gamma(x_1)$. L'histoire sur n étapes devient $H_n = (S \times A)^{n-1}$.

On note alors s_n la variable du signal reçu après l'action a_n, qui nous fournit une information sur ω_{n+1}, et $h_n = (a_1, s_1, ..., a_{n-1}, s_{n-1})$ la variable de l'histoire du jeu dans H_n.

Puisqu'on n'a plus une connaissance parfaite sur l'état mais seulement une croyance, on définit maintenant les stratégies markoviennes (resp. stationnaires) sous la forme $\sigma = (f_n) \in (\Delta(\Omega) \to \Delta(A))^{\mathbb{N}}$ (resp. $\sigma : \Delta(\Omega) \to \Delta(A)$) où l'antécédent dans $\Delta(\Omega)$ fourni à f_n est la probabilité postérieure y_n, explicitée ci-dessous.

3.1.2 Calcul de la probabilité postérieure et de l'espérance conditionnelle

Ce genre de formules est assez fastidieux à obtenir et de nombreux articles, dont celui sur lequel je me suis basé pour cette partie, évitent de les expliciter. Cependant, je pense que voir ces formules écrites en clair au moins une fois aide à comprendre comment un POMDP évolue, surtout lorsque l'on n'est pas familier avec la théorie des jeux en général.

On définit $y_n(\overline{h}_n, \omega) = \mathbb{P}_{h_n = \overline{h}_n}(\omega_n = \omega)$ la probabilité postérieure sur ω_n. On a alors la relation de récurrence suivante :

$$y_n(\overline{h}_n, \omega) = \frac{\sum_{\tilde{\omega} \in \Omega} y_{n-1}(\overline{h}_{n-1}, \tilde{\omega}) \times q(a, \tilde{\omega}; s, \omega)}{\mathbb{P}_{\overline{h}_{n-1}}(s_{n-1} = s | a_{n-1} = a)}$$

avec $\mathbb{P}_{\bar{h}_{n-1}}(s_{n-1}=s|a_{n-1}=a)=\sum_{\tilde{\omega}\in\Omega} y_{n-1}(\bar{h}_{n-1},\tilde{\omega})\times\left(\sum_{\omega\in\Omega} q(a,\tilde{\omega};s,\omega)\right)$ le facteur de normalisation.

La stratégie σ influence les probabilités du système via $\mathbb{P}_{\bar{h}_{n-1}}(a_{n-1}=a)=\sigma(\bar{h}_{n-1},a_{n-1})$. Posons $\mathbb{P}(h_1=\emptyset)=1$. On a alors la relation suivante sur la probabilité des histoires du jeu en suivant la stratégie σ :

$$\mathbb{P}(h_{n+1}=\overline{h_{n+1}})=\mathbb{P}(h_n=\bar{h}_n)\times\sigma(h_n=\bar{h}_n,a_n=a)\times\sum_{\omega,\tilde{\omega}\in\Omega} y_n(\bar{h}_n,\tilde{\omega})\times q(a,\tilde{\omega};s,a)$$

On déduit enfin de ceci l'expression de l'espérance conditionnelle du gain au temps n :

$$\mathbb{E}_{x_1,\sigma}[r(\omega_n,a_n)]=\sum_{\bar{h}_n\in H_n,\omega\in\Omega}\mathbb{P}(h_n=\bar{h}_n)\times y_n(\bar{h}_n,\omega)\times\left(\sum_{a\in A}\sigma(\bar{h}_n,a)\times r(\omega,a)\right)$$

3.1.3 Résultats préliminaires

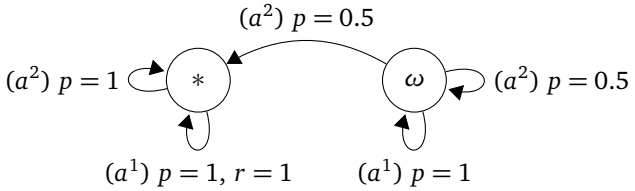
A l'instar des MDPs, en adaptant les équations de Bellman à ce modèle, on en déduit, pour $x_1\in\Delta(\Omega)$ donné, l'existence d'une stratégie σ_N pure markovienne optimale dans $\Gamma_N(x_1)$.

De façon générale, pour toute stratégie σ, l'espérance conditionnelle sur le gain est 1-lipschitzienne par rapport à la distribution de départ sur $(\mathbb{R}^\Omega,\|.\|_1)$. Autrement dit :

$$\forall x,x'\in\Delta(\Omega),\forall 0<m\leq n,|\gamma_{m,n}(\sigma,x)-\gamma_{m,n}(\sigma,x')|\leq\|x-x'\|_1$$

3.2 Un exemple de POMDP à convergence "lente"

Avant de s'intéresser au résultat central du rapport, considérons un exemple afin de nous familiariser un peu avec la manipulation de ce modèle. Ce POMDP va jouer en premier lieu le rôle de contrexemple à la généralisation du résultat sur la vitesse de convergence des v_λ pour les MDPs. En second lieu, on va s'intéresser à l'hypothétique lien entre les stratégies optimales dans les Γ_n et une stratégie ϵ-optimale dans Γ_∞.



Dans ce jeu, $A=\{a^1,a^2\}$, $|S|=1$ et $\Omega=\{\omega,*\}$. On note $p_n=(y_n)_\omega$ la probabilité postérieure de $\omega_n=\omega$, avec la condition initiale $p_1=1$. Le seul gain non nul est $r(*,a^1)=1$. On constate que $(a_n=a^1)\Rightarrow(p_{n+1}=p_n)$ et $(a_n=a^2)\Rightarrow(p_{n+1}=\frac{p_n}{2})$. Pour une stratégie pure σ, le gain à l'étape n vérifie :

$$\mathbb{E}_\sigma\left[r(\omega_n,a_n)\right]=(1-p_n(a_1,...,a_{n-1}))\times\delta\left(a_n=a^1\right)$$

Dans ce contexte, comme $|S|=1$, on ne reçoit aucune information sur l'état atteint après chaque action. On peut alors voir ce POMDP comme un MDP sur l'espace infini $\Delta(\Omega)$ ou $[0,1]$ de façon équivalente car $p\in[0,1]$, la probabilité d'être en ω, décrit donc entièrement notre croyance sur ce jeu.

3.2.1 Étude des v_λ et de Γ_0

En adaptant les actions et les récompenses en conséquence, on peut alors exprimer l'équation d'optimalité de Bellman sous la forme :

$$v_\lambda(p) = \max\left\{\lambda \times (1-p) + (1-\lambda)v_\lambda(p)\ ;\ (1-\lambda)v_\lambda\left(\frac{p}{2}\right)\right\}$$

Le terme de gauche (resp. droite) est ici la conséquence de l'action a^1 (resp. a^2).

Si $v_\lambda(p) = \lambda \times (1-p) + (1-\lambda)v_\lambda(p)$, alors $v_\lambda(p) = 1-p$:

$$v_\lambda(p) = \max\left\{(1-p)\ ;\ (1-\lambda)v_\lambda\left(\frac{p}{2}\right)\right\}$$

Comme $r \leq 1$ (et donc $v \leq 1$), si $p \leq \lambda$ alors $(1-\lambda)v_\lambda\left(\frac{p}{2}\right) \leq 1-p$ d'où $v_\lambda(p) = 1-p$ à coup sûr. A partir d'un rang $n_0 \in \mathbb{N}$, $\frac{1}{2^n} < \lambda$ donc $\exists n \in \mathbb{N}, v_\lambda(1) = (1-\lambda)^n \times \left(1-\frac{1}{2^n}\right)$. Soit $f_\lambda : x \mapsto \left(\frac{(1-\lambda)}{2}\right)^x \times (2^x - 1)$. On a la majoration $v_\lambda(1) \leq \max_{x\in\mathbb{R}^+} f_\lambda(x)$.

f_λ est maximale en $x_\lambda \geq 0$ tel que $f'_\lambda(x_\lambda) = 0$, ce qui aboutit au résultat $x_\lambda = \frac{\ln\left(1-\frac{\ln(2)}{\ln(1-\lambda)}\right)}{\ln(2)}$. En réinjectant x_λ dans f_λ, on obtient le résultat suivant, où $\theta = -\frac{\ln(1-\lambda)}{\ln(2)} > 0$:

$$f_\lambda(x_\lambda) = \frac{(\theta)^\theta}{(1+\theta)^{(1+\theta)}}$$

Après avoir effectué un développement en 0, on obtient $f_\lambda(x_\lambda) = 1 + \lambda \times \ln(\lambda) + O(\lambda)$. L'étude des $v_n(1)$ permet de confirmer que $\Gamma_\infty(1)$ a la valeur limite 1 donc d'après le théorème taubérien ce résultat est également valable dans $\Gamma_0(1)$. On en déduit que dans le meilleur des cas, la convergence de v_λ se fait à la vitesse $\lambda\ln(\lambda)$, ce qui est effectivement plus lent que le $O(\lambda)$ obtenu dans le cas des MDPs.

L'existence de modèles à convergence aussi lente qu'on le veut est une question ouverte. J'ai travaillé sur ce problème durant la dernière semaine de mon stage. Mon tuteur m'a montré un exemple de modèle à "prise de risque" où la convergence se fait en $\sqrt{\lambda}$ au mieux. En combinant ces deux modèles j'ai réussi à définir formellement un POMDP pour lequel la vitesse de convergence est minorée par $\sqrt{\lambda}\ln(\lambda)$ (à un facteur multiplicatif près). Une conjecture qui semble raisonnable à l'heure actuelle est l'existence d'un POMDP à vitesse de convergence en λ^ϵ pour tout $\epsilon > 0$.

3.2.2 Étude des v_n et de Γ_∞

Soit $\sigma_N = (a^2)^{(N)}, (a^1)^{(\infty)}$ une stratégie pure. Pour $n \geq N$, on a $\gamma_n(\sigma_N) = (1-p_N) \times \frac{n-N}{n}$ donc $\lim_{n\to\infty} \gamma_n(\sigma_N) = (1-p_N) = 1-\frac{1}{2^N} \leq v(1) := v(\omega)$. N est pris arbitrairement grand donc $v(1) \geq 1$, or $r \leq 1$ d'où $v(1) = 1$.

Soit σ une stratégie pure telle qu'il existe $n_0 \in \mathbb{N}^*$ pour lequel $a_{n_0} = a^1$ et $a_{n_0+1} = a^2$. Posons $\tilde{\sigma} = (a_1, ..., a_{n_0-1}, a_{n_0+1}, a_{n_0}, a_{n_0+2}, ...)$. On a, pour $n \notin \{n_0, n_0+1\}$, $p_n(\sigma) = p_n(\tilde{\sigma})$ et la même action a_n donc la même espérance de gain. De plus, $E_\sigma[r(\omega_{n_0+1}, a_{n_0+1})] = E_{\tilde{\sigma}}[r(\omega_{n_0}, a_{n_0})] = 0$. Les deux termes restants font apparaître une inégalité stricte :

$$E_{\tilde{\sigma}}[r(\omega_{n_0+1}, a_{n_0+1})] = 1 - \frac{p_{n_0}(\tilde{\sigma})}{2} > 1 - p_{n_0}(\tilde{\sigma}) = 1 - p_{n_0}(\sigma) = E_\sigma[r(\omega_{n_0}, a_{n_0})]$$

Il en découle :

$$\forall n > n_0, \gamma_n(\sigma) < \gamma_n(\tilde{\sigma})$$

Les équations de Bellman nous permettent de conclure qu'il existe une stratégie pure optimale dans $\Gamma_n(1)$. Sachant qu'on peut optimiser la stratégie σ ci-dessus en "décalant" les a^2 sur la gauche, on en déduit que la stratégie optimale dans $\Gamma_n(1)$ est un des σ_N ci-dessus. Pour $1 \leq N \leq n$, on a ainsi $\gamma_n(\sigma_N) = (1-\frac{1}{2^N}) \times (n-N)$.

Chercher $v_n(1)$ revient à maximiser $f_n : x \mapsto (1-\frac{1}{2^x})(n-x)$ sur $\mathbb{N}$. Soit x_n l'entier positif où f_n maximale : par une étude similaire à celle de f_λ, on en déduit $x_n \geq \lfloor \frac{\ln(n\ln(2)+1)}{\ln(2)} \rfloor - 1$.

Il en découle $\lim\limits_{n\to\infty} x_n = \infty$. Dans ce cas, si on considère la limite terme à terme de nos stratégies optimales, on obtient la suite $(a^2)^{(\infty)}$ de gain nul, donc en aucun cas ϵ-optimale dans $\Gamma_\infty(1)$. Cela met en évidence la nécessité d'un procédé de construction plus astucieux, brièvement introduit ci-dessous.

3.3 Stratégies optimales en chambre noire

Dans cette sous-partie, on s'intéresse au cas où le jeu est non dégénéré, non observable, c'est-à-dire $|S| = 1$. Dans ce contexte, on remarque qu'une stratégie pure définit $a_{n+1} = \sigma(a_1, ...a_n)$ de façon déterministe. On peut donc considérer les stratégies pures σ comme des suites $(a_n) \in A^{\mathbb{N}}$.

En exploitant la compacité de $\Delta(\Omega)$ et la régularité de l'espérance conditionnelle en fonction des conditions initiales, on peut faire un "recollage" judicieux entre certaines stratégies σ_N optimales dans les $\Gamma_N(x_1)$ afin d'obtenir une suite de stratégies $(\pi_k)_{k\in\mathbb{N}}$ pour lesquelles on a un contrôle sur le comportement à partir d'un même rang N_1 fixé et jusqu'à un rang N_k de plus en plus grand. On en déduit, après extraction diagonale, l'existence d'une valeur limite (notée $w := \limsup\limits_{n\to\infty} v_n(x_1)$) et d'une stratégie $\pi = (a_n)$ pure et ϵ-optimale dans $\Gamma_\infty(x_1)$ (p.1184-1186 [2]).

On va ci-dessous énoncer puis démontrer le *théorème 2*, en se basant sur les pages 1186 à 1188 de l'article de Rosenberg [2]. L'intérêt est ici de voir une démonstration complète (dont certains points éludés dans l'article d'origine) qui est, à mon sens, assez représentative du genre de raisonnements qu'on retrouve en (PO)MDPs.

3.3.1 Théorème

Supposons A et Ω finis et $|S| = 1$. Alors pour tout $x_1 \in \Delta(\Omega)$, pour tout $\epsilon > 0$, il existe $\sigma \in A^{\Delta(\Omega)}$ une stratégie pure stationnaire ϵ-optimale dans $\Gamma_n(x_1)$.

3.3.2 Lemme préliminaire

On a une "équivalence" entre les notions d'optimalité par rapport aux valeurs v_n (celle définie dans la première partie), et par rapport à la valeur limite w ($\gamma_n \geq w - \epsilon$).

En effet, comme w est limite des v_n, pour $\epsilon > 0$ fixé, $v_n(x_1) \geq w(x_1) - \epsilon$ à partir d'un rang. Dans ce cas, s'il existe σ ϵ-optimale dans $\Gamma_0(x_1)$, alors à partir d'un rang, on a $\gamma_n(\sigma) \geq w - 2\epsilon$.

On peut de même montrer qu'une stratégie est asymptotiquement ϵ-optimale selon les v_n ou selon w de façon équivalente.

3.3.3 Démonstration

Soit $\pi = (a_n)$ pure et ϵ-optimale dans $\Gamma_\infty(x_1)$. On pose $y_n = y_n(\pi)$.

On va par la suite construire une stratégie stationnaire à partir de π, en distinguant principalement le *Cas 1* (auquel on se ramène dans le *Cas 2* et le *Cas 3*), où on peut trouver un cycle dans π, et le *Cas 4*, où on ramène π à une suite injective d'actions.

On peut distinguer un des quatre cas suivants (non mutuellement exclusifs) :

• *Cas 1 : π est ultimement d-périodique pour un certain $d > 0$.*

Cela signifie qu'à partir d'un rang, les actions effectuées par le joueur ainsi que ses croyances sur l'état occupé évoluent de façon périodique, se répètent toutes les d étapes. Sachant cela, on va modifier la partie périodique de π puis le chemin initial afin d'obtenir une stratégie stationnaire.

Formellement, l'hypothèse du *Cas 1* se traduit par $\exists n_1 \geq n_0, \exists d \geq 1, \forall n \geq n_1, a_{n+d} = a_n, y_{n+d} = y_n$. Pour $n \geq n_1$, on a la division euclidienne $n - n_1 = q \times d + r$. Alors on peut décomposer $\gamma_n(\pi, x_1)$ sous la forme $\frac{n_1-1}{n}\gamma_{n_1-1} + \frac{dq}{n}\gamma_{n_1,n_1+d-1} + \frac{r+1}{n}\gamma_{n_1+dq,n}$. Le terme périodique devient nettement dominant en temps long, ce qui donne l'inégalité suivante :

$$\gamma_{n_1,n_1+d-1} = \lim_{n\to\infty} \gamma_n(\pi) \geq w - \epsilon$$

Pour obtenir la nature stationnaire de la partie périodique, l'article de Rosenberg propose une récurrence sans l'expliciter. L'hypothèse de récurrence sous-entendue m'a semblé peu claire, alors j'ai reformulé la preuve avec une hypothèse satisfaisante. On peut atteindre le résultat voulu en montrant la propriété suivante par récurrence forte sur la période d :

S'il existe $\pi \in A^H$ éventuellement (d-)périodique à partir d'un rang n_1, telle que $\gamma_{n_1,n_1+d-1} \geq w - \epsilon$, alors il existe $\pi' \in A^{\mathbb{N}}$ d'-périodique à partir d'un rang n_1', et éventuellement stationnaire, c'est-à-dire $\exists n_2 \geq n_1', \forall m, n \geq n_2, (y_m = y_n) \Rightarrow (a_m = a_n)$.

— $d = 1$:

π et (y_n) sont constantes à partir de n_1 donc π éventuellement stationnaire.

— $d > 1$:

– $\forall n_1 \leq i < j \leq n_1 + d - 1, y_i \neq y_j$:

π est éventuellement stationnaire, à partir de n_1.

– $\exists i \neq j, y_i = y_j$:

$$\gamma_{i,j-1} \times \frac{j-i}{d} + \gamma_{j,i+d-1} \times \frac{d-(j-i)}{d} = \gamma_{i,i+d-1} = \gamma_{n_1,n_1+d-1} \geq w - \epsilon$$

On a un "croisement" dans le cycle du point de vue de la suite (y_n), on peut donc découper la phase périodique en deux termes dont au moins un vérifie l'inégalité voulue :

- $\gamma_{i,j-1} \geq w - \epsilon$:

$\pi' = (a_1, ..., a_{i-1}), (a_i, ..., a_{j-1})^\infty$ vérifie l'hypothèse de récurrence pour la période $j - i < d$.

- $\gamma_{i,j-1} \geq w - \epsilon$:

$\pi' = (a_1, ..., a_{j-1}), (a_j, ..., a_{i+d-1})^\infty$ vérifie l'hypothèse pour la période $d - (j - i) < d$.

La récurrence est montrée donc d'après les hypothèses du *Cas 1*, il existe π' éventuellement d'-périodique et stationnaire à partir de n_2, 2ϵ-optimale. On les note respectivement π, d et ϵ par la suite.

Soient maintenant $Y := \{y_n, 1 \leq n \leq n_2 + d - 1\}$ et $S := \{(y_n, y_{n+1})\}$. (Y, S) définit le graphe de transitions orienté de π sur $\Delta(\Omega)$. Par construction, on peut atteindre tout Y depuis y_1.

Il existe alors un chemin $(y_{i_1}, ..., y_{i_k})$ de y_1 vers $\{y_n, n \geq n_2\}$ de longueur minimale, $i_k \leq n_2$. Ce chemin est fini, injectif, $y_{i_1} = y_1$, et quitte à changer l'indice de début du cycle n_2, $y_{i_k} = y_{n_2}$.

Soit π" $= (a_{i_1}, ..., a_{ik-1}), (a_{n_2}, ..., a_{n_2+d-1})^\infty$:

$$\forall n \leq k, y_n(\pi") = y_{i_n}(\pi) \text{ et } (y_n(\pi"))_{n\geq k} = (y_n(\pi))_{n\geq n_2}$$

Par construction, $\pi" \in A^{\Delta(\Omega)}$ est purement stationnaire. Or $\gamma_{k,k+d-1}(x_1, \pi") = \gamma_{n_2,n_2+d-1}(x_1, \pi) \geq w - \epsilon$.

Avec la division euclidienne $n-k=qd+r$, on peut alors décomposer le gain moyen :

$$\gamma_n(x_1,\pi") = \frac{k-1}{n}\times\gamma_{k-1}(\pi") + \frac{qd}{n}\times\gamma_{k,k+d-1} + \frac{r+1}{n}\times\gamma_{k+qd,n}(\pi")$$

Pour $n \geq n_3 = \frac{k(d+n_2}{\epsilon}$, on a $\gamma_n(x_1,\pi") \geq (1-\epsilon)(w-\epsilon) \geq w-2\epsilon$. En conclusion, en reprenant les notations initiales, si π vérifie le *Cas 1*, on peut obtenir $\pi" \in A^{\Delta(\Omega)}$ stationnaire, pure et 4ϵ-stationnaire.

• *Cas 2 : π permet directement une périodicité.*

Dans ce cas, une croyance se répète dans (y_n) et le gain entre ces répétitions est assez élevé. Formellement, $\exists n_1 < n_2, y_{n_1} = y_{n_2}$ et $y_{n_1,n_2-1}(x_1,\pi) \geq w-\epsilon$.

Cela implique que la suite $\pi' = (a_1,...,a_{n_1-1}),(a_{n_1},...,a_{n_2-1})^\infty$ est 2ϵ-périodique et vérifie le *Cas 1*, d'où l'existence d'une stratégie $\pi"$ stationnaire et 8ϵ-optimale.

• *Cas 3 : π repasse infiniment souvent en un $y \in \Delta(\Omega)$.*

Dans ce cas, il existe $y \in \Delta(\Omega)$ tel que $\{n, y_n(\pi) = y\}$ est infini. On pose $(n_i)_{i\geq 1}$ le parcours croissant de cet ensemble. $\forall i \geq 1, y_{n_{i+1}} = y_{n_i}$.

Supposons $\forall i \geq 1, \gamma_{n_i,n_{i+1}}(\pi) < w-2\epsilon$. Alors à partir d'un rang, $\gamma_n(\pi) < w-\epsilon$, ce qui contredit la nature ϵ-optimale de π. Ainsi, $\exists n_i, \gamma_{n_i,n_{i+1}}(\pi) \geq w-2\epsilon$.

On s'est ramené au *Cas 2* avec 2ϵ, d'où l'existence d'un $\pi"$ stationnaire et 16ϵ-optimal.

• *Cas 4 : π ne vérifie aucun des cas précédents.*

On ne peut extraire de π un comportement périodique. Le but va alors être au contraire de filtrer toutes les répétitions jusqu'à obtention d'une stratégie "injective" et donc stationnaire.

Ici, π contredit les cas précédents. Notamment, pour tout $y \in \Delta(\Omega)$, l'ensemble $\{n \in \mathbb{N}, y_n = y\}$ est fini et pour tout $n \geq m \geq 1$, on a $y_n = y_m \Rightarrow \gamma_{n,m-1} < w-\epsilon$.

Soit alors la suite $(i_n)_{n\geq 0}$ définie par $i_0 = 0$ et $i_{k+1} = \sup\{n \geq 1, y_n = y_{i_k+1}\} < \infty$. On remarque que (i_n) est bien une suite infinie, $y_{i_k+1} = y_{i_{k+1}}$ et $(y_{i_k})_{k\geq 1}$ injective.

Posons $\pi' := (a_{i_k})_{k\geq 1}$. Par construction, $(y_k(\pi'))_{k\geq 1} = (y_{i_k})_{k\geq 1}$ d'où π' stationnaire. Soit $k_0 \geq n_0$. Pour $n = i_{k_0}$, on a l'égalité :

$$n\gamma_n(\pi) = k_0\gamma_{k_0}(\pi') + \sum_{0\leq k<k_0, i_{k+1}>i_k+1} (i_{k+1}-i_k-1)\gamma_{i_k+1,i_{k+1}-1}(\pi) \geq w-\epsilon$$

Or $y_{i_k+1} = y_{i_{k+1}}$ donc par hypothèse initiale $\gamma_{i_k+1,i_{k+1}-1}(\pi) < w-\epsilon$. Les termes restants doivent donc compenser la moyenne, ainsi, pour $k \geq n_0$, on a nécessairement $\gamma_{k_0}(\pi') \geq w-\epsilon$. La stratégie π' est bien stationnaire, pure et ϵ-optimale.

Dans le pire des cas précédents, on peut construire $\pi"$ 16ϵ-optimale (selon w), donc 32ϵ-optimale dans notre problème $\Gamma_\infty(x_1)$ d'après le lemme préliminaire. En partant d'une stratégie π $\frac{\epsilon}{32}$-optimale on en déduit le résultat du théorème.

3.4 Existence de stratégies optimales dans le cas général

Avec des raisonnements assez similaires, Rosenberg, Solan et Vieille [2] montrent l'existence de stratégies ϵ-optimales et d'une valeur limite $w(x_1) = \lim v_n(x_1)$ dans le problème $\Gamma_\infty(x_1)$ dans le cas d'un jeu dégénéré, où $1 < |S| < \infty$. Contrairement au cas non dégénéré, les stratégies obtenues dans l'article ne sont a priori ni pures ni stationnaires.

Dans l'article de Renault [3], à l'aide d'outils topologiques plus puissants, il étend ce résultat au cas où $|\Omega| < \infty$ mais où A et S peuvent être pris non vides quelconques, éventuellement infinis voire indénombrables.

4 Bilan

Commençons par synthétiser les idées de ce rapport. Nous avons d'abord introduit les MDPs, qui modélisent des systèmes où un joueur seul, qui a connaissance du fonctionnement du jeu et de l'état qu'il occupe à chaque instant, doit maximiser son gain. Dans ce contexte, nous avons vu que le joueur peut "bêtement" réagir à chaque état par une action fixée pour jouer de façon optimale sur le long terme. Dans un second temps, nous nous sommes intéressés aux POMDPs, où le joueur n'a plus connaissance de l'état occupé, où il doit inférer une croyance sur les états possiblement occupés à partir des signaux qu'il reçoit. Dans ce cas, nous avons vu que le joueur peut suivre une stratégie optimale à ϵ près sur le long terme, mais celle-ci ne sera plus aussi simple que dans le cas des MDPs, en particulier s'il doit réagir à plusieurs signaux.

En continuant sur la lancée des résultats de ce rapport, j'aurais pu dans mon stage étudier d'autres résultats concernant l'existence de solutions particulières dans les POMDPs. Cependant, j'ai fait le choix d'explorer plusieurs domaines plutôt que de creuser aussi loin que possible le sujet initial. Dans cette optique, dans un rapport plus long, j'aurais aimé traiter de mon étude des algorithmes de recherche de stratégies optimales en présentant par exemple la méthode *Value Iteration*, l'alter-ego de la méthode PI présentée plus haut. J'aurais également voulu détailler les raisonnements mis en œuvre pour obtenir le POMDP à convergence en $O\left(\sqrt{\lambda}\ln(\lambda)\right)$ ainsi que le jeu en question.

Références

[1] David Blackwell. Discrete dynamic programming. *The Annals of Mathematical Statistics*, 33(2) :719–726, 1962.

[2] Dinah Rosenberg, Eilon Solan, and Nicolas Vieille. Blackwell optimality in markov decision processes with partial observation. *Annals of statistics*, 30 :1178–1193, 2002.

[3] Jerome Renault. Uniform value in dynamic programming. *Journal of the European Mathematical Society*, 13(2) :309–330, 2011.

[4] Jerome Renault. Programmation dynamique et jeux stochastiques. `https://sites.google.com/site/jrenaultsite/lecturenotes`, 2012.

[5] Martin L. Puterman. *Markov Decision Processes. Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2005.

[6] Andrew G. Barto Richard S. Sutton. *Reinforcement Learning : an Introduction*. Adaptive Computation and Machine Learning. The MIT Press, 1998.