Chapitre 1

# Processus Décisionnels de Markov

## 1.1. Introduction

Les problèmes de décision traités dans cet ouvrage sont communément appelés *problèmes de décision séquentielle dans l'incertain*. La première caractéristique de ce type de problèmes est qu'il s'inscrit dans la durée et que ce n'est pas en fait un, mais plusieurs problèmes de décisions en séquence qu'un *agent* (ou décideur ou encore acteur) doit résoudre, chaque décision courante influençant la résolution des problèmes qui suivent. Ce caractère séquentiel des décisions se retrouve typiquement dans les problèmes de planification en intelligence artificielle et relève en particulier des méthodes de plus court chemin dans un graphe. La seconde caractéristique de ces problèmes est liée à l'incertitude des conséquences mêmes de chacune des décisions possibles. Ainsi, l'agent ne sait pas à l'avance précisément quels seront les effets des décisions qu'il prend. En tant que telle, cette problématique relève des théories de la décision dans l'incertain qui proposent de nombreuse voies de formalisation et approches de résolution, en particulier la théorie classique de maximisation de l'utilité espérée.

Les problèmes de décision séquentielle dans l'incertain couplent donc les deux problématiques de décision séquentielle et de décision dans l'incertain. Les *problèmes décisionnels de Markov* (MDP[1]) en sont une formalisation mathématique, qui généralise les approches de plus court chemin dans un environnement stochastique. A la base de ce formalisme, les *processus décisionnels de Markov* (que l'on note aussi MDP) intègrent les concepts d'*état* qui résume la situation de l'agent à chaque instant, d'*action*

Chapitre rédigé par Frédérick GARCIA.

1. Pour *Markov Decision Problem*

(ou décision) qui influence la dynamique de l'état, de *revenu* (ou *récompense*) qui est associé à chacune des transitions d'état. Les MDP sont alors des chaînes de Markov visitant les états, contrôlées par les actions et valuées par les revenus. Résoudre un MDP, c'est contrôler l'agent pour qu'il se comporte de manière optimale, c'est-à-dire de façon à maximiser son revenu. Toutefois, les solutions d'un MDP ne sont pas des décisions ou séquences de décisions, mais plutôt des *politiques*, ou *stratégies*, ou encore *règles de décision*, qui spécifient l'action à entreprendre en chacune des étapes pour toutes les situations futures possibles de l'agent. Du fait de l'incertitude, une même politique peut donner lieu à des séquences d'états / actions très variées selon les aléas.

EXEMPLE.– Illustrons ces concepts de manière plus concrète en prenant l'exemple de l'entretien d'une voiture. La question qui se pose est de décider, en fonction de l'*état* de la voiture (présence de panne, usure, âge, *etc*.), quelle est la meilleure *stratégie* (ne rien faire, remplacer préventivement, réparer, changer de voiture, *etc*.) pour minimiser le coût de l'entretien sur le long terme. Si on fait l'hypothèse que l'on connaît les *conséquences* et le *coût* des différentes actions pour chaque état (par exemple on connaît la *probabilité* qu'un moteur lâche si on ne répare pas une fuite d'huile) alors on peut modéliser ce problème comme un MDP dont la solution nous donnera, en fonction de l'état de la voiture, l'action optimale. Ainsi, la suite des actions prises au fur et à mesure de l'évolution de l'état de la voiture permettra, en moyenne, de minimiser son coût d'entretien.

Le cadre des problèmes décisionnels de Markov et ses généralisations que nous développerons dans des chapitres ultérieurs forment les modèles les plus classiques pour les problèmes de décision séquentielle dans l'incertain. Nous en exposons les bases dans ce chapitre, dans le cas d'un agent qui dispose a priori d'une connaissance parfaite du processus et de son état à tout instant, dont la tâche consiste donc à planifier a priori une politique optimale qui maximise son revenu au cours du temps.

## 1.2. Problèmes décisionnels de Markov

### 1.2.1. *Processus décisionnels de Markov*

Les processus décisionnels de Markov sont définis comme des processus stochastiques contrôlés satisfaisant la propriété de Markov, assignant des récompenses aux transitions d'états [BER 87, PUT 94]. On les définit par un quintuplet : $(S, A, T, p, r)$ où :

– $S$ est l'espace d'états dans lequel évolue le processus ;

– $A$ est l'espace des actions qui contrôlent la dynamique de l'état ;

– $T$ est l'espace des temps, ou axe temporel ;

– $p()$ sont les probabilités de transition entre états ;

– $r()$ est la fonction de récompense sur les transitions entre états.

La figure 1.1 représente un MDP sous la forme d'un diagramme d'influence. A chaque instant $t$ de $T$, l'action $a_t$ est appliquée dans l'état courant $s_t$, influençant le processus dans sa transition vers l'état $s_{t+1}$. La récompense $r_t$ est émise au cours de cette transition.

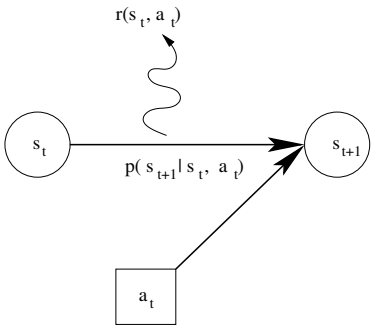


**Figure 1.1.** *Processus décisionnel de Markov.*

Le domaine $T$ des étapes de décision est un ensemble discret, assimilé à un sous-ensemble de $\mathbb{N}$, qui peut être fini ou infini (on parle d'horizon fini ou d'horizon infini).

Les domaines $S$ et $A$ sont supposés finis, même si de nombreux résultats peuvent être étendus aux cas où $S$ et $A$ sont dénombrables ou continus (voir [BER 95] pour une introduction au cas continu). Dans le cas général, l'espace $A$ peut être dépendant de l'état courant ($A_s$ pour $s \in S$). De même, $S$ et $A$ peuvent être fonction de l'instant $t$ ($S_t$ et $A_t$). Nous nous limiterons ici au cas classique où $S$ et $A$ sont constants tout au long du processus.

Les probabilités de transition caractérisent la dynamique de l'état du système. Pour une action $a$ fixée, $p(s'|s,a)$ représente la probabilité que le système passe dans l'état $s'$ après avoir exécuté l'action $a$ dans l'état $s$. On impose classiquement que $\forall s, a$, $\sum_{s'} p(s'|s,a) = 1$. Par ailleurs, on utilise classiquement une représentation matricielle de ces probabilités de transition, en notant $P_a$ la matrice de dimension $|S| \times |S|$ dont les éléments sont $\forall s, s'\ P_{a\,s,s'} = p(s' \mid s, a)$. Les probabilités décrites par $p()$ se décrivent donc par $|A|$ matrices $P_a$, chacune des lignes de ces matrices ayant pour somme 1 : les $P_a$ sont des matrices *stochastiques*.

Les distributions $p()$ vérifient la propriété fondamentale qui donne son nom aux processus décisionnels de Markov considérés ici. Si on note $h_t$ l'historique à la date $t$ du processus, $h_t = (s_0, a_0, \ldots, s_{t-1}, a_{t-1}, s_t)$, alors la probabilité d'atteindre un nouvel état $s_{t+1}$ suite à l'exécution de l'action $a_t$ n'est fonction que de $a_t$ et de l'état courant $s_t$ et ne dépend pas de l'historique $h_t$. Si on note de façon standard $P(x|y)$ la

probabilité conditionnelle de l'événement $x$ sachant que $y$ est vrai, on a :

$$\forall h_t, a_t, s_{t+1} \quad P(s_{t+1} \mid h_t, a_t) = P(s_{t+1} \mid s_t, a_t) = p(s_{t+1} \mid s_t, a_t)$$

Il faut noter que cela n'implique pas que le processus stochastique induit $(s_t)_{t \in T}$ soit lui-même markovien, tout dépend de la politique de choix des actions $a_t$.

Comme résultat d'avoir choisi l'action $a$ dans l'état $s$ à l'instant $t$, l'agent décideur reçoit une récompense, ou revenu, $r_t = r(s,a) \in \mathbb{R}$. Les valeurs de $r_t$ positives peuvent être considérées comme des gains et les valeurs négatives comme des coûts. Cette récompense peut être instantanément perçue à la date $t$, ou accumulée de la date $t$ à la date $t+1$, l'important est qu'elle ne dépende que de l'état et de l'action choisie à l'instant courant. La représentation vectorielle de la fonction de récompense $r(s,a)$ consiste en $|A|$ vecteurs $r_a$ de dimension $|S|$.

Une extension classique est de considérer des récompenses $r(s,a)$ aléatoires et l'on considère alors la valeur moyenne $r_t = \bar{r}(s,a)$. En particulier, $r_t$ peut ainsi dépendre de l'état d'arrivée $s'$ selon $r(s,a,s')$. On considère alors la valeur moyenne est $\bar{r}(s,a) = \sum_{s'} p(s'|s,a) r(s,a,s')$. Dans tous les cas, on suppose $r_t$ bornée dans $\mathbb{R}$.

Par ailleurs, comme pour $S$ et $A$, les fonctions de transition et de récompense peuvent elles-mêmes varier au cours du temps, auquel cas on les note $p_t$ et $r_t$. Lorsque ces fonctions de varient pas, on parle de processus *stationnaires* : $\forall t \in T, \quad p_t() = p(), \; r_t() = r()$. Par la suite, nous supposerons vérifiée cette hypothèse de stationnarité dans l'étude des MDP à horizon infini.

### 1.2.2. *Les politiques d'actions*

Les processus décisionnels de Markov permettent de modéliser la dynamique de l'état d'un système soumis au contrôle d'un agent, au sein d'un environnement stochastique. On nomme alors politique (notée $\pi$), ou stratégie, la procédure suivie par l'agent pour choisir à chaque instant l'action à exécuter. Deux distinctions sont essentielles ici. Tout d'abord, une politique peut déterminer précisément l'action à effectuer, ou simplement définir une distribution de probabilité selon laquelle cette action doit être sélectionnée. Ensuite, une politique peut se baser sur l'historique $h_t$ du processus, ou peut ne simplement considérer que l'état courant $s_t$. Nous obtenons ainsi quatre familles distinctes de stratégies, comme indiqué sur le tableau 1.1 :

Pour une politique déterministe, $\pi_t(s_t)$ ou $\pi_t(h_t)$ définit l'action $a$ choisie à l'instant $t$. Pour une politique aléatoire, $\pi_t(a, s_t)$ ou $\pi_t(a, h_t)$ représente la probabilité de sélectionner $a$.

Ces quatre familles de politique définissent les quatre ensembles suivants :

| politique $\pi_t$ | déterministe | aléatoire |
|---|---|---|
| markovienne | $s_t \longrightarrow a_t$ | $a_t,\ s_t \longrightarrow [0,1]$ |
| histoire-dépendante | $h_t \longrightarrow a_t$ | $h_t,\ s_t \longrightarrow [0,1]$ |

**Tableau 1.1.** *Différentes familles de politiques pour les* MDP.

– $\Pi^{HA}$ pour les politiques histoire-dépendantes aléatoires ;
– $\Pi^{HD}$ pour les politiques histoire-dépendantes déterministes ;
– $\Pi^{MA}$ pour les politiques markoviennes aléatoires ;
– $\Pi^{MD}$ pour les politiques markoviennes déterministes.

Ces différentes familles de politiques sont imbriquées entre elles, de la plus générale (histoire-dépendante aléatoire) à la plus spécifique (markovienne déterministe), comme le montre la figure 1.2.

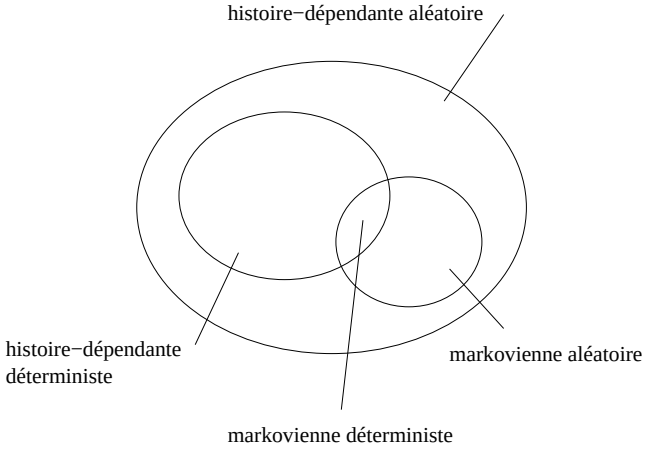


**Figure 1.2.** *Relations entre les différentes familles de politiques*

Indépendamment de cela et comme pour le processus décisionnel de Markov lui-même, la définition des politiques peut ou non dépendre explicitement du temps. Ainsi, une politique est *stationnaire* si $\forall t\ \pi_t = \pi$. Parmi ces politiques stationnaires, les politiques markoviennes déterministes sont centrales dans l'étude des MDP. Il s'agit du modèle le plus simple de stratégie décisionnelle, on nomme leur ensemble $\mathcal{D}$ :

DÉFINITION 1.1.– *Politiques markoviennes déterministes stationnaires*
*$\mathcal{D}$ est l'ensemble des fonctions $\pi$ qui à tout état de $S$ associent une action de $A$ :*

$$\pi : s \in S \longrightarrow \pi(s) \in A$$

Un autre ensemble important, noté $\mathcal{D}^A$ est constitué des politiques markoviennes aléatoires stationnaires. Les politiques de $\mathcal{D}$ et $\mathcal{D}^A$ sont très importantes car, comme nous le verrons, $\mathcal{D}$ et $\mathcal{D}^A$ contiennent les politiques optimales pour les principaux critères.

### 1.2.3. *Critères de performance*

Se poser un problème décisionnel de Markov, c'est rechercher parmi une famille de politiques celles qui optimisent un critère de performance donné pour le processus décisionnel markovien considéré. Ce critère a pour ambition de caractériser les politiques qui permettront de générer des séquences de récompenses les plus importantes possibles. En termes formels, cela revient toujours à évaluer une politique sur la base d'une mesure du *cumul espéré* des récompenses instantanées le long d'une trajectoire, comme on peut le voir sur les critères les plus étudiés au sein de la théorie des MDP, qui sont respectivement :

– le critère fini : $E[r_0 + r_1 + r_2 + \cdots + r_{N-1} \mid s_0]$

– le critère $\gamma$-pondéré : $E[r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots + \gamma^t r_t + \cdots \mid s_0]$

– le critère total : $E[r_0 + r_1 + r_2 + \cdots + r_t + \cdots \mid s_0]$

– le critère moyen : $\lim_{n\to\infty} \frac{1}{n} E[r_0 + r_1 + r_2 + \cdots + r_{n-1} \mid s_0]$

Les deux caractéristiques communes à ces quatre critères sont en effet d'une part leur formule additive en $r_t$, qui est une manière simple de résumer l'ensemble des récompenses reçues le long d'une trajectoire et, d'autre part, l'espérance $E[.]$ qui est retenue pour résumer la distribution des récompenses pouvant être reçues le long des trajectoires, pour une même politique et un même état de départ. Ce choix d'un cumul espéré est bien sûr important, car il permet d'établir le *principe d'optimalité de Bellman* [BEL 57] (« les sous-politiques de la politique optimale sont des sous-politiques optimales »), à la base des nombreux algorithmes de programmation dynamique permettant de résoudre efficacement les MDP. On verra au chapitre 5 d'autres critères qui étendent les MDP, pour lesquels le principe d'optimalité n'est plus nécessairement respecté.

Dans la suite de ce chapitre, nous allons successivement caractériser les politiques optimales et présenter les algorithmes permettant d'obtenir ces politiques optimales pour chacun des précédents critères.

### 1.3. Fonctions de valeur

Les critères fini, $\gamma$-pondéré, total et moyen que nous venons de voir permettent de définir une *fonction de valeur* qui, pour une politique $\pi$ fixée, associe à tout état initial $s \in S$ la valeur du critère considéré en suivant $\pi$ à partir de $s$ : $\forall \pi \quad V^\pi : S \longrightarrow \mathbb{R}$.

On note $\mathcal{V}$ l'espace des fonctions de $S$ dans $\mathbb{R}$, identifiable à l'espace vectoriel $\mathbb{R}^{|S|}$. L'ensemble $\mathcal{V}$ est muni d'un ordre partiel naturel :

$$\forall U, V \in \mathcal{V} \quad U \leq V \Leftrightarrow \forall s \in S \quad U(s) \leq V(s).$$

L'objectif d'un problème décisionnel de Markov est alors de caractériser et de rechercher – si elles existent – les politiques optimales $\pi^* \in \Pi^{HA}$ telles que

$$\forall \pi \in \Pi^{HA} \quad \forall s \in S \quad V^\pi(s) \leq V^{\pi^*}(s)$$

soit encore

$$\pi^* \in \underset{\pi \in \Pi^{HA}}{\operatorname{argmax}} V^\pi.$$

On note alors $V^* = \max_{\pi \in \Pi^{HA}} V^\pi = V^{\pi^*}$. Dans le cadre des MDP, on recherche donc des politiques optimales meilleures que toute autre politique, quel que soit l'état de départ. Remarquons que l'existence d'une telle politique optimale n'est pas en soi évidente.

La spécificité des problèmes décisionnels de Markov est alors de pouvoir être traduits en terme d'équations d'optimalité portant sur les fonctions de valeur, dont la résolution est de complexité moindre que le parcours exhaustif de l'espace global des politiques de $\Pi^{HA}$ (la taille du simple ensemble $\mathcal{D}$ est déjà de $|A|^{|S|}$).

#### 1.3.1. *Le critère fini*

On suppose ici que l'agent doit contrôler le système en $N$ étapes, avec $N$ fini. Le critère fini conduit naturellement à définir la fonction de valeur qui associe à tout état $s$ l'espérance de la somme des $N$ prochaines récompenses en suivant la politique $\pi$ à partir de $s$ :

DÉFINITION 1.2.– *Fonction de valeur pour le critère fini*
*Si* $T = \{0, \ldots, N-1\}$*, on pose*

$$\forall s \in S \quad V_N^\pi(s) = E^\pi[\sum_{t=0}^{N-1} r_t \mid s_0 = s].$$

Dans cette définition, $E^\pi[.]$ dénote l'espérance mathématique sur l'ensemble des réalisations du MDP en suivant la politique $\pi$. $E^\pi$ est associée à la distribution de probabilité $P^\pi$ sur l'ensemble de ces réalisations.

Notons qu'il est parfois utile d'ajouter au critère une récompense terminale $r_N$ fonction du seul état final $s_N$. Il suffit pour cela de considérer une étape artificielle supplémentaire où $\forall s, a \ \ r_N(s,a) = r_N(s)$. C'est le cas par exemple lorsqu'il s'agit de piloter un système vers un état but en $N$ étapes et à moindre coût.

### 1.3.2. *Le critère $\gamma$-pondéré*

Le critère $\gamma$-pondéré, ou critère actualisé, est le critère à horizon infini le plus classique. La fonction de valeur du critère $\gamma$-pondéré est celle qui associe à tout état $s$ la limite lorsque $N$ tend vers l'infini de l'espérance en suivant la politique $\pi$ à partir de $s$ de la somme des $N$ futures prochaines récompenses, pondérées par un *facteur d'actualisation*[2] $\gamma$ :

DÉFINITION 1.3.– *Fonctions de valeur pour le critère $\gamma$-pondéré*
*Pour $0 \leq \gamma < 1$, on pose*

$$\begin{aligned} \forall s \in S \quad V_\gamma^\pi(s) &= E^\pi[r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots + \gamma^t r_t + \cdots \mid s_0 = s] \\ &= E^\pi[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s] \end{aligned}$$

Le terme $\gamma$ représente la valeur à la date $t$ d'une unité de récompense reçue à la date $t+1$. Reçue à la date $t+\tau$, cette même unité vaudrait $\gamma^\tau$. Cela implique que les instants de décision $t = 0, 1, 2, \ldots$ de $T$ soient régulièrement répartis sur $\mathbb{N}$. Ce facteur $\gamma$ a pour principal intérêt d'assurer la convergence de la série en horizon infini. D'un point de vue pratique, il est naturellement utilisé au sein des MDP modélisant des processus de décision économique en posant $\gamma = \frac{1}{1+\tau}$, où $\tau$ est le taux d'actualisation.

### 1.3.3. *Le critère total*

Il est toutefois possible de choisir $\gamma = 1$ dans certains cas particuliers à horizon infini. Lorsque cela a un sens, on pose ainsi :

2. *discount factor* en anglais

DÉFINITION 1.4.– *Fonction de valeur pour le critère total*

$$V^{\pi}(s) = E^{\pi}[\sum_{t=0}^{\infty} r_t \mid s_0 = s]$$

Ce critère est en pratique souvent utilisé pour des problèmes à horizon temporel aléatoire fini non borné : on sait que le processus de décision va s'arrêter à une étape terminale, mais on ne peut borner cet instant. Ce type de modèle est particulièrement utilisé dans des applications de type *optimal stopping* (à tout instant, la décision de l'agent porte simplement sur l'arrêt ou non du processus aléatoire), ou plus généralement de type jeux et paris.

### 1.3.4. *Le critère moyen*

Lorsque la fréquence des décisions est importante, avec un facteur d'actualisation proche de 1, ou lorsqu'il n'est pas possible de donner une valeur économique aux récompenses, on préfère considérer un critère qui représente la moyenne des récompenses le long d'une trajectoire et non plus leur somme pondérée. On associe ainsi à une politique l'espérance du gain moyen par étape. On définit alors le gain moyen $\rho^{\pi}(s)$ associé à une politique particulière $\pi$ et à un état $s$ :

DÉFINITION 1.5.– *Le gain moyen*

$$\rho^{\pi}(s) = \lim_{n \to \infty} E^{\pi}[\frac{1}{n} \sum_{t=0}^{n-1} r_t \mid s_0 = s]$$

Pour le critère moyen, une politique $\pi^*$ est dite *gain-optimale* si $\rho^{\pi^*}(s) \geq \rho^{\pi}(s)$ pour toute politique $\pi$ et tout état $s$.

Ce critère est particulièrement utilisé dans des applications de type gestion de file d'attente, de réseau de communication, de stock etc.

## 1.4. Politiques markoviennes

### 1.4.1. *Equivalence des politiques histoire-dépendantes et markoviennes*

Nous allons établir ici une propriété fondamentale des MDP pour ces différents critères, qui est d'accepter comme politiques optimales des politiques simplement markoviennes, sans qu'il soit nécessaire de considérer l'espace total $\Pi^{HA}$ des politiques histoire-dépendantes.

PROPOSITION 1.1.– *Soit $\pi \in \Pi^{HA}$ une politique aléatoire histoire-dépendante. Pour chaque état initial $x \in S$, il existe alors une politique aléatoire markovienne $\pi' \in \Pi^{MA}$ telle que*

*1)* $V_N^{\pi'}(x) = V_N^{\pi}(x)$,

*2)* $V_\gamma^{\pi'}(x) = V_\gamma^{\pi}(x)$,

*3)* $V^{\pi'}(x) = V^{\pi}(x)$,

*4)* $\rho^{\pi'}(x) = \rho^{\pi}(x)$

PREUVE.– Soit $x \in S$ et $\pi$ une politique aléatoire histoire-dépendante. Soit $\pi'$ la politique aléatoire markovienne définie à partir de $\pi$ et $x$ selon :

$$\forall t = 0, 1, \ldots, \forall s \in S, \forall a \in A \quad \pi'(a_t = a, s_t = s) = P^{\pi}(a_t = a \mid s_t = s, s_0 = x)$$

On a ainsi $P^{\pi'}(a_t = a \mid s_t = s) = P^{\pi}(a_t = a \mid s_t = s, s_0 = x)$. On montre alors par récurrence sur $t$ que $P^{\pi}(s_t = s, a_t = a \mid s_0 = x) = P^{\pi'}(s_t = s, a_t = a \mid s_0 = x)$.

L'égalité est directe pour $t = 0$. Pour $t > 0$, en supposant établie la propriété jusqu'à $t - 1$, on a

$$\begin{aligned} P^{\pi}(s_t = s \mid s_0 = x) &= \sum_{i \in S} \sum_{a \in A} P^{\pi}(s_{t-1} = i, a_{t-1} = a \mid s_0 = x) p(s \mid i, a) \\ &= \sum_{i \in S} \sum_{a \in A} P^{\pi'}(s_{t-1} = i, a_{t-1} = a \mid s_0 = x) p(s \mid i, a) \\ &= P^{\pi'}(s_t = s \mid s_0 = x) \end{aligned}$$

D'où $P^{\pi'}(s_t = s, a_t = a \mid s_0 = x) = P^{\pi'}(a_t = a \mid s_t = s) P^{\pi'}(s_t = s \mid s_0 = x) = P^{\pi}(a_t = a \mid s_t = s, s_0 = x) P^{\pi}(s_t = s \mid s_0 = x) = P^{\pi}(s_t = s, a_t = a \mid s_0 = x)$, ce qui établit la récurrence.

On conclut en remarquant alors que pour tout $x \in S$

$$\begin{aligned} V_N^{\pi}(x) &= \sum_{t=0}^{t=N-1} E^{\pi}[r(s_t, a_t) \mid s_0 = x] \\ V_\gamma^{\pi}(x) &= \sum_{t=0}^{t=\infty} \gamma^t E^{\pi}[r(s_t, a_t) \mid s_0 = x] \\ V^{\pi}(x) &= \sum_{t=0}^{t=\infty} E^{\pi}[r(s_t, a_t) \mid s_0 = x] \\ \rho^{\pi}(x) &= \lim_{n \to \infty} \frac{1}{n} \sum_{t=0}^{t=n-1} E^{\pi}[r(s_t, a_t) \mid s_0 = x], \end{aligned}$$

et

$$E^{\pi}[r(s_t, a_t) \mid s_0 = x] = \sum_{s \in S} \sum_{a \in A} r(s, a) P^{\pi}(s_t = s, a_t = a \mid s_0 = x).$$

□

Ce résultat permet d'affirmer que lorsque l'on connaît l'état initial (ou une distribution de probabilité sur l'état initial), toute politique histoire-dépendante aléatoire peut être remplacée par une politique markovienne aléatoire ayant la même fonction de valeur.

### 1.4.2. *Politique markovienne et chaîne de Markov valuée*

Pour toute politique markovienne $\pi \in \Pi^{MA}$, le processus décrit par l'état $s_t$ vérifie, pour tout $s_0, s_1, \cdots, s_{t+1}$, $P^{\pi}(s_{t+1} \mid s_0, s_1, \cdots, s_t) = \sum_{a \in A} P^{\pi}(a_t = a \mid s_0, s_1, \cdots, s_t) P^{\pi}(s_{t+1} \mid s_0, s_1, \cdots, s_t, a_t = a) = \sum_{a \in A} \pi(a, s_t) P^{\pi}(s_{t+1} \mid s_t, a_t = a) = P^{\pi}(s_{t+1} \mid s_t)$.

Il s'agit donc d'un processus markovien, qui forme une chaîne de Markov dont la matrice de transition notée $P_{\pi}$ est définie par

$$\forall s, s' \quad P_{\pi s,s'} = P^{\pi}(s_{t+1} = s' \mid s_t = s) = \sum_{a} \pi(a, s) p(s' \mid s, a).$$

Dans le cas où $\pi$ est déterministe ($\pi \in \Pi^{MD}$), $P_{\pi s,s'}$ est simplement égal à $p(s' \mid s, \pi(s))$. La matrice $P_{\pi}$ est construite simplement en retenant pour chaque état $s$ la ligne correspondante dans la matrice $P_a$ avec $a = \pi(s)$. De même, on note $r_{\pi}$ le vecteur de composante $r(s, \pi(s))$ pour $\pi \in \Pi^{MD}$ et $\sum_a \pi(a, s) r(s, a)$ pour $\pi \in \Pi^{MA}$.

Le triplet $(S, P_{\pi}, r_{\pi})$ définit ce que l'on nomme un *processus de Markov valué*, ou *chaîne de Markov valuée*. Il s'agit simplement d'une chaîne de Markov avec des revenus associés aux transitions. Nous verrons qu'évaluer une politique $\pi$ consiste alors à calculer certaines grandeurs asymptotiques (pour les critères infinis) caractéristiques de la chaîne de Markov valuée associée.

## 1.5. Caractérisation des politiques optimales

### 1.5.1. *Le critère fini*

#### 1.5.1.1. *Equations d'optimalité*

Supposons que l'agent se trouve dans l'état $s$ lors de la dernière étape de décision, confronté au choix de la meilleure action à exécuter. Il est clair que la meilleure décision à prendre est celle qui maximise la récompense instantanée à venir, qui viendra

s'ajouter à celles qu'il a déjà perçues. On a ainsi :

$$\pi^*_{N-1}(s) \in \operatorname*{argmax}_{a\in A} r_{N-1}(s,a)$$

et

$$V^*_1(s) = \max_{a\in A} r_{N-1}(s,a)$$

où $\pi^*_{N-1}$ est la politique optimale à suivre à l'étape $N-1$ et $V^*_1$ la fonction de valeur optimale pour un horizon de longueur 1, obtenue en suivant cette politique optimale.

Supposons maintenant l'agent dans l'état $s$ à l'étape $N-2$. Le choix d'une action $a$ va lui rapporter de façon sûre la récompense $r_{N-2}(s,a)$ et l'amènera de manière aléatoire vers un nouvel état $s'$ à l'étape $N-1$. Là, il sait qu'en suivant la politique optimale $\pi^*_{N-1}$, il pourra récupérer une récompense moyenne $V^*_1(s')$. Le choix d'une action $a$ à l'étape $N-2$ conduit donc au mieux en moyenne à la somme de récompenses $r_{N-2}(s,a)+\sum_{s'} p_{N-2}(s'|s,a)V^*_1(s')$. Ainsi, le problème de l'agent à l'étape $N-2$ se ramène simplement à rechercher l'action qui maximise cette somme, soit :

$$\pi^*_{N-2}(s) \in \operatorname*{argmax}_{a\in A}\{r_{N-2}(s,a) + \sum_{s'} p_{N-2}(s'|s,a)V^*_1(s')\}$$

et

$$V^*_2(s) = \max_{a\in A}\{r_{N-2}(s,a) + \sum_{s'} p_{N-2}(s'|s,a)V^*_1(s')\}.$$

Ce raisonnement peut s'étendre jusqu'à la première étape de décision, où l'on a donc :

$$\pi^*_0(s) \in \operatorname*{argmax}_{a\in A}\{r_0(s,a) + \sum_{s'} p_0(s'|s,a)V^*_{N-1}(s')\}$$

et

$$V^*_N(s) = \max_{a\in A}\{r_0(s,a) + \sum_{s'} p_0(s'|s,a)V^*_{N-1}(s')\}.$$

Cela conduit ainsi à l'énoncé du théorème suivant :

THÉORÈME 1.1.– *Equations d'optimalité pour le critère fini*
*Soit $N < \infty$. Les fonctions de valeurs optimales $V^* = (V^*_N, \ldots, V^*_1)$ sont les solutions uniques du système d'équations*

$$\forall s \in S \quad V^*_{n+1}(s) = \max_{a\in A}\{r_{N-1-n}(s,a) + \sum_{s'} p_{N-1-n}(s'|s,a)V^*_n(s')\} \qquad (1.1)$$

*avec $n = 0, \ldots, N-1$ et $V_0 = 0$. Les politiques optimales pour le critère fini $\pi^* = (\pi^*_0, \pi^*_1, \ldots, \pi^*_{N-1})$ sont alors déterminées par :*

$$\forall s \in S \quad \pi^*_t(s) \in \operatorname*{argmax}_{a\in A}\{r_t(s,a) + \sum_{s'} p_t(s'|s,a)V^*_{N-1-t}(s')\}$$

*pour* $t = 0, \ldots, N-1$.

On voit donc ici dans le cadre du critère fini que les politiques optimales sont de type markovien déterministe, mais non stationnaire (le choix de la meilleure décision à prendre dépend de l'instant $t$).

1.5.1.2. *Evaluation d'une politique markovienne déterministe*

Soit une politique $\pi$ markovienne déterministe. La même démarche permet alors de caractériser sa fonction de valeur $V_N^\pi$ :

THÉORÈME 1.2.– *Caractérisation de* $V_N$
*Soient* $N < \infty$ *et* $\pi = (\pi_0, \pi_1, \ldots, \pi_{N-1})$ *une politique markovienne. Alors* $V_N^\pi = V_N$*, avec* $(V_N, V_{N-1}, \ldots, V_1)$*, solutions du système d'équations linéaires*

$$\forall s \in S \quad V_{n+1}(s) = r_{N-1-n}(s, \pi_{N-1-n}(s)) + \sum_{s'} p_{N-1-n}(s'|s, \pi_{N-1-n}(s)) V_n(s')$$

*pour* $n = 0, \ldots, N-1$ *et* $V_0 = 0$.

### 1.5.2. *Le critère $\gamma$-pondéré*

Ce critère est le plus classique en horizon infini et celui pour lequel il est assez simple de caractériser la fonction de valeur optimale et les politiques associées. On rappelle que l'on suppose ici, dans le cas de l'horizon infini, que le MDP considéré est stationnaire.

1.5.2.1. *Evaluation d'une politique markovienne stationnaire*

Pour une politique markovienne stationnaire $\pi \in \mathcal{D}^A$, on définit l'opérateur $L_\pi$ de $\mathcal{V}$ dans $\mathcal{V}$, espace vectoriel muni de la norme max : $\forall V \in \mathcal{V}$, $\| V \| = \max_{s \in S} | V(s) |$.

DÉFINITION 1.6.– *Opérateur* $L_\pi$
*Pour* $\pi \in \mathcal{D}^A$*,*

$$\forall V \in \mathcal{V} \quad L_\pi V = r_\pi + \gamma P_\pi V$$

Un premier résultat permet alors de relier la fonction de valeur $V_\gamma^\pi$ d'une politique stationnaire $\pi \in \mathcal{D}^A$ à cet opérateur $L_\pi$ :

THÉORÈME 1.3.– *Caractérisation de* $V_\gamma^\pi$
*Soient* $\gamma < 1$ *et* $\pi \in \mathcal{D}^A$ *une politique stationnaire markovienne.*
*Alors* $V_\gamma^\pi$ *est l'unique solution de l'équation* $V = L_\pi V$ *:*

$$\forall s \in S \quad V(s) = r_\pi(s) + \gamma \sum_{s' \in S} P_{\pi,s,s'} V(s') \tag{1.2}$$

*et* $V_\gamma^\pi = (I - \gamma P_\pi)^{-1} r_\pi$.

PREUVE.– Soit $V$ solution de $V = L_\pi V$. On a donc $(I - \gamma P_\pi)V = r_\pi$. La matrice $P_\pi$ étant stochastique, toutes les valeurs propres de la matrice $\gamma P_\pi$ sont de modules inférieurs ou égaux à $\gamma < 1$ et donc la matrice $I - \gamma P_\pi$ est inversible, avec

$$(I - \gamma P_\pi)^{-1} = \sum_{k=0}^{\infty} \gamma^k P_\pi^k$$

d'où

$$V = (I - \gamma P_\pi)^{-1} r_\pi = \sum_{k=0}^{\infty} \gamma^k P_\pi^k r_\pi.$$

$$\begin{aligned}
\text{Or, } \forall s \in S \quad V_\gamma^\pi(s) &= E^\pi[r_0 + \gamma r_1 + \gamma^2 r_2 + \cdots + \gamma^t r_t + \cdots \mid s_0 = s] \\
&= \sum_{t=0}^{\infty} \gamma^t E^\pi[r(s_t, a_t) \mid s_0 = s] \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{s' \in S} \sum_{a \in A} P^\pi(s_t = s', a_t = a \mid s_0 = s) r(s', a) \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{s' \in S} \sum_{a \in A} q(a, s') P^\pi(s_t = s' \mid s_0 = s) r(s', a) \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{s' \in S} P^\pi(s_t = s' \mid s_0 = s) r_\pi(s') \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{s' \in S} P^t_{\pi,s,s'} r_\pi(s') \\
&= \sum_{t=0}^{\infty} \gamma^t P_\pi^t r_\pi(s),
\end{aligned}$$

et donc $V = V_\gamma^\pi$. □

1.5.2.2. *Equations d'optimalité*

Rappelons que l'on cherche à résoudre le problème d'optimisation $\forall s \in S, V_\gamma^*(s) = \max_{\pi \in \Pi^{HA}} V_\gamma^\pi(s)$. Une politique $\pi^*$ est dite optimale si $V_\gamma^{\pi^*} = V_\gamma^*$. De par la propriété 1.1, on a alors

$$\forall s \in S \quad V_\gamma^*(s) = \max_{\pi \in \Pi^{HA}} V_\gamma^\pi(s) = \max_{\pi \in \Pi^{MA}} V_\gamma^\pi(s).$$

Soit donc maintenant l'opérateur $L$ de l'ensemble des fonctions de valeur $\mathcal{V}$ dans lui-même, nommé *opérateur de programmation dynamique* :

DÉFINITION 1.7.– *Opérateur $L$*

$$\forall V \in \mathcal{V} \quad \forall s \in S \quad LV(s) = \max_{a \in A} \left( r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V(s') \right)$$

*soit en notation vectorielle*

$$\forall V \in V \quad LV = \max_{\pi \in \mathcal{D}} (r_\pi + \gamma P_\pi V)$$

Le théorème principal concernant l'optimalité des fonctions de valeur pour le critère $\gamma$-pondéré est alors le suivant :

THÉORÈME 1.4.– *Equation de Bellman*
*Soit $\gamma < 1$. Alors $V_\gamma^*$ est l'unique solution de l'équation $V = LV$ :*

$$\forall s \in S \quad V(s) = \max_{a \in A} \left( r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V(s') \right). \tag{1.3}$$

PREUVE.–

Montrons que $\forall V$ et pour $0 \leq \gamma \leq 1$

$$LV = \max_{\pi \in \mathcal{D}} (r_\pi + \gamma P_\pi V) = \max_{\pi \in \mathcal{D}^A} (r_\pi + \gamma P_\pi V)$$

Pour cela, considérons une fonction de valeur $V$ et $\delta \in \mathcal{D}^A$. Pour tout $s$, du fait du caractère positif des $\delta(a,s)$, on a

$$\sum_a \delta(a,s) \left( r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V(s') \right)$$

$$\begin{aligned}
&\leq \sum_a \delta(a,s) \max_{a'} \left( r(s,a') + \gamma \sum_{s' \in S} p(s' \mid s,a') V(s') \right) \\
&\leq \sum_a \delta(a,s) LV(s) \\
&\leq LV(s)
\end{aligned}$$

Ainsi, pour tout $\delta \in \mathcal{D}^A$

$$r_\delta + \gamma P_\delta V \leq \max_{\pi \in \mathcal{D}} (r_\pi + \gamma P_\pi V)$$

soit

$$\max_{\delta \in \mathcal{D}^A} (r_\delta + \gamma P_\delta V) \leq \max_{\pi \in \mathcal{D}} (r_\pi + \gamma P_\pi V)$$

L'inégalité inverse est immédiate car $\mathcal{D} \subset \mathcal{D}^A$.

Montrons alors que $\forall V, V \geq LV \Rightarrow V \geq V_\gamma^*$, et $V \leq LV \Rightarrow V \leq V_\gamma^*$.

Soit $V$ telle que $V \geq LV$. On a donc

$$V \geq \max_{\pi \in \mathcal{D}} \{r_\pi + \gamma P_\pi V\} = \max_{\pi \in \mathcal{D}^A} \{r_\pi + \gamma P_\pi V\}$$

Soit $\pi = (\pi_0, \pi_1, \ldots) \in \Pi^{MA}$. Pour tout $t$, $\pi_t \in \mathcal{D}^A$, d'où

$$\begin{aligned} V &\geq r_{\pi_0} + \gamma P_{\pi_0} V \\ &\geq r_{\pi_0} + \gamma P_{\pi_0}(r_{\pi_1} + \gamma P_{\pi_1} V) \\ &\geq r_{\pi_0} + \gamma P_{\pi_0} r_{\pi_1} + \gamma^2 P_{\pi_0} P_{\pi_1} r_{\pi_2} + \cdots + \gamma^{n-1} P_{\pi_0} \cdots P_{\pi_{n-2}} r_{\pi_{n-1}} + \gamma^n P_\pi^n V. \end{aligned}$$

On a donc

$$V - V_\gamma^\pi \geq \gamma^n P_\pi^n V - \sum_{k=n}^{\infty} \gamma^k P_\pi^k r_{\pi_k}, \text{ car } V_\gamma^\pi = \sum_{k=0}^{\infty} \gamma^k P_\pi^k r_{\pi_k},$$

avec $P_\pi^k = P_{\pi_0} P_{\pi_1} \cdots P_{\pi_{k-1}}$. Les deux termes de droite peuvent être rendus aussi petits que désiré pour $n$ suffisamment grand, car $\| \gamma^n P_\pi^n V \| \leq \gamma^n \| V \|$ et $\| \sum_{k=n}^{\infty} \gamma^k P_\pi^k r_{\pi_k} \| \leq \sum_{k=n}^{\infty} \gamma^k R \leq \frac{\gamma^n}{1-\gamma} R$ avec $R = max_{s,a} r(s,a)$. On en déduit

$$V - V_\gamma^\pi \geq 0$$

Cela tant vrai pour toute politique $\pi \in \Pi^{MA}$, on a donc

$$V \geq \max_{\pi \in \Pi^{MA}} V_\gamma^\pi = \max_{\pi \in \Pi^{HA}} V_\gamma^\pi = V_\gamma^*$$

Inversement, soit $V$ telle que $V \leq LV$. On a donc $V \leq \max_{\pi \in \mathcal{D}} \{r_\pi + \gamma P_\pi V\}$ Supposons ce max atteint en $\pi^*$. On a donc

$$\begin{aligned} V &\leq r_{\pi^*} + \gamma P_{\pi^*} V \\ &\leq r_{\pi^*} + \gamma P_{\pi^*} (r_{\pi^*} + \gamma P_{\pi^*} V) \\ &\leq r_{\pi^*} + \gamma P_{\pi^*} r_{\pi^*} + \cdots + \gamma^{n-1} P_{\pi^*}^{n-1} r_{\pi^*} + \gamma^n P_{\pi^*}^n V \\ V - V_\gamma^{\pi^*} &\leq -\sum_{k=n}^{\infty} \gamma^k P_{\pi^*}^k r_{\pi^*} + \gamma^n P_{\pi^*}^n V \end{aligned}$$

Les termes de droite pouvant être rendus aussi proches de 0 que désiré, on a donc $V - V_\gamma^{\pi^*} \leq 0$, soit $V \leq V_\gamma^{\pi^*} \leq V_\gamma^*$.

On a ainsi montré que $V \geq LV \Rightarrow V \geq V_\gamma^*$, $V \leq LV \Rightarrow V \leq V_\gamma^*$, ce qui implique que $V = LV \Rightarrow V = V_\gamma^*$ : toute solution de l'équation $LV = V$ est nécessairement égale à la fonction de valeur optimale $V_\gamma^*$. Montrons maintenant qu'une telle solution existe.

Rappelons pour cela le théorème du point fixe de Banach :

THÉORÈME 1.5.– *Théorème du point fixe de Banach*
*Soient $\mathcal{U}$ un espace de Banach (i.e. espace vectoriel normé complet) et $T$ une contraction sur $\mathcal{U}$ (i.e. $\forall u, v \parallel Tu - Tv \parallel \leq \lambda \parallel u - v \parallel$ pour $0 \leq \lambda < 1$). Alors*

*1) Il existe un unique $u^* \in \mathcal{U}$ tel que $Tu^* = u^*$ ;*

*2) Pour tout $u_0 \in \mathcal{U}$, la suite $(u_n)_{n \geq 0}$ définie par*

$$u_{n+1} = Tu_n = T^{n+1}u_0$$

*converge vers $u^*$.*

L'espace $\mathcal{V}$ muni de la norme max est un espace vectoriel normé de dimension fini donc complet. Il suffit donc de montrer que l'opérateur $L$ est une contraction pour cette norme.

PROPOSITION 1.2.– *Soit $\gamma < 1$. L'opérateur de programmation dynamique $L$ défini par $LV = \max_{\pi \in \mathcal{D}} (r_\pi + \gamma P_\pi V)$ est une contraction sur $\mathcal{V}$.*

PREUVE.– Soient $U$ et $V$ dans $\mathcal{V}$ et $s \in S$.

Supposons $LV(s) \geq LU(s)$. Soit $a_s^* \in \operatorname{argmax}_{a \in A} \left( r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s, a) V(s') \right)$. On a alors

$$\begin{aligned}
0 \quad \leq \quad \mid LV(s) - LU(s) \mid \quad &= \quad LV(s) - LU(s) \\
&\leq \quad r(s, a_s^*) + \gamma \sum_{s' \in S} p(s' \mid s, a_s^*) V(s') \\
&\qquad -r(s, a_s^*) - \gamma \sum_{s' \in S} p(s' \mid s, a_s^*) U(s') \\
&\leq \quad \gamma \sum_{s' \in S} p(s' \mid s, a_s^*)(V(s') - U(s')) \\
&\leq \quad \gamma \sum_{s' \in S} p(s' \mid s, a_s^*) \parallel V - U \parallel \\
&\leq \quad \gamma \parallel V - U \parallel
\end{aligned}$$

D'où

$$\parallel LV - LU \parallel = \max_s \mid LV(s) - LU(s) \mid \leq \gamma \parallel V - U \parallel .$$

Cette propriété de contraction assure donc l'existence pour l'opérateur $L$ d'un point fixe unique qui est donc égal à $V_\gamma^*$. □

On termine alors l'analyse du critère $\gamma$-pondéré avec le théorème suivant :

THÉORÈME 1.6.– *Caractérisation des politiques optimales*
*Soit $\gamma < 1$. Alors*

*1) $\pi^* \in \Pi^{HA}$ est optimale $\Leftrightarrow V_\gamma^{\pi^*}$ est solution de $LV = V$ et $V_\gamma^{\pi^*} = V_\gamma^*$ ;*

*2) toute politique stationnaire $\pi^*$ définie par*

$$\pi^* \in \underset{\pi \in \mathcal{D}}{\operatorname{argmax}}\{r_\pi + \gamma P_\pi V_\gamma^*\}$$

*est une politique optimale.*

PREUVE.– La première équivalence est évidente du fait du théorème précédent. Soit alors $\pi^* \in \operatorname{argmax}_{\pi \in \mathcal{D}}\{r_\pi + \gamma P_\pi V_\gamma^*\}$. On a alors

$$\begin{aligned} L_{\pi^*} V_\gamma^* &= r_{\pi^*} + \gamma P_{\pi^*} V_\gamma^* \\ &= \max_{\pi \in \mathcal{D}}\{r_\pi + \gamma P_\pi V_\gamma^*\} \\ &= LV_\gamma^* \\ &= V_\gamma^*. \end{aligned}$$

L'unicité de la solution de $V = L_{\pi^*} V$ démontrée par le théorème 1.3 permet de déduire que $V_\gamma^* = V_\gamma^{\pi^*}$ et donc que $\pi^*$ est optimale. □

### 1.5.3. *Le critère total*

L'existence de la limite définissant le critère total ne peut être assurée que sous certaines hypothèses. Nous considérons ici deux classes de problèmes pour lesquelles cette limite existe nécessairement, et est finie pour au moins une politique : les modèles positifs et les modèles négatifs.

DÉFINITION 1.8.– *Soit $\pi \in \Pi^{HA}$. On définit les fonctions $V_+^\pi$ et $V_-^\pi$ par*

$$\begin{aligned} V_+^\pi(s) &= E^\pi[\sum_{t=0}^{\infty} \max(r_t, 0) \mid s_0 = s] \\ V_-^\pi(s) &= E^\pi[\sum_{t=0}^{\infty} \max(-r_t, 0) \mid s_0 = s] \end{aligned}$$

On appelle alors respectivement $\mathcal{V}^+$ et $\mathcal{V}^-$ l'ensemble des fonctions positives et négatives de $\mathcal{V}$

On suppose que pour toute politique $\pi$ et état $s$, $V_+^\pi(s)$ ou $V_-^\pi(s)$ est fini, ce qui implique l'existence (finie ou infinie) de la limite $V^\pi$ avec

$$\forall s \in S \quad V^\pi(s) = V_+^\pi(s) - V_-^\pi(s).$$

Les MDP bornés positivement, ou encore positifs, sont tels que
– pour chaque $s$ il existe au moins une action $a \in A$ avec $r(s,a) \geq 0$ et
– $V_+^\pi(s) < \infty$ pour tout $s \in S$ et pour tout $\pi \in \Pi^{HA}$.

Les MDP négatifs sont tels que
– $r(s,a) \leq 0$ pour tout $s \in S$ et $a \in A$ et
– il existe $\pi \in \Pi^{HA}$ telle que $V^\pi(s) > -\infty$ pour tout $s \in S$.

L'existence d'une politique $\pi$ pour laquelle $V_\pi(s)$ soit fini pour tout $s \in S$ est typiquement assurée par la présence d'un état absorbant $s_\infty$ à récompense nulle :

$$\forall s_0 \quad P^\pi(\exists t^* \, s_{t^*} = s_\infty) = 1$$

avec

$$p(s_\infty \mid s_\infty, \pi(s_\infty)) = 1, \text{ et } r(s_\infty, \pi(s_\infty)) = 0$$

Pour un modèle positif, une politique optimale a un revenu total positif le plus éloigné de 0 possible. L'agent cherche à prolonger le plus possible les trajectoires pour accumuler des revenus positifs. Pour un modèle négatif, une politique optimale a un revenu négatif aussi proche de 0 possible. L'agent cherche à terminer aussi vite que possible en $s_\infty$ pour minimiser les pertes. Les deux modèles ne sont donc pas exactement symétriques.

Nous énonçons ci-dessous quelques résultats importants concernant ces deux modèles. Pour cela, nous introduisons une nouvelle définition des opérateurs $L_\pi$ et $L$.

DÉFINITION 1.9.– *Opérateurs $L_\pi$ et L pour le critère total*
*Soit $\pi$ stationnaire $\in \mathcal{D}^A$,*

$$\forall V \in \mathcal{V} \quad L_\pi V = r_\pi + P_\pi V$$

*et*

$$\forall V \in V \quad LV = \max_{\pi \in \mathcal{D}} (r_\pi + P_\pi V)$$

On montre alors pour les MDP positifs et négatifs les résultats suivants :

THÉORÈME 1.7 [PUT 94].– *Soit un* MDP *positif. Alors*

*1) pour tout $\pi$ stationnaire $\in \mathcal{D}$, $V^\pi$ est la solution minimale de $V = L_\pi V$ dans $\mathcal{V}^+$.*

*2) $V^*$ est la solution minimale de l'équation $V = LV$ dans $\mathcal{V}^+$.*

*3) une politique $\pi^* \in \Pi^{HA}$ est optimale $\Leftrightarrow V^{\pi^*} = LV^{\pi^*}$*

*4) si $\pi^* \in \operatorname{argmax}_{\pi \in \mathcal{D}} (r_\pi + P_\pi V^*)$ et si $\lim_{N \to \infty} P_{\pi^*}^N V^*(s) = 0$ pour tout $s \in S$, alors $\pi^*$ est optimale.*

THÉORÈME 1.8 [PUT 94].– *Soit un* MDP *négatif. Alors*

*1) pour tout $\pi$ stationnaire $\in \mathcal{D}$, $V^\pi$ est la solution maximale de $V = L_\pi V$ dans $\mathcal{V}^-$.*

*2) $V^*$ est la solution maximale de l'équation $V = LV$ dans $\mathcal{V}^-$.*

*3) toute politique $\pi^* \in \operatorname{argmax}_{\pi \in \mathcal{D}} (r_\pi + P_\pi V^*)$ est optimale.*

On note que pour un MDP négatif, il peut exister une politique $\pi$ vérifiant $V^\pi = LV^\pi$ qui ne soit pas optimale (voir l'exemple 7.3.1 dans [PUT 94]).

### 1.5.4. *Le critère moyen*

L'analyse théorique du critère moyen est plus complexe que pour les précédents critères. Elle fait intervenir le comportement limite du processus markovien valué sous-jacent. Nous nous limitons ici à présenter les résultats principaux, dans le cadre des MDP récurrents (pour toute politique markovienne déterministe, la chaîne de Markov correspondante est constituée d'une unique classe récurrente), unichaînes (chaque chaîne de Markov est constituée d'une unique classe récurrente plus éventuellement quelques états transitoires) ou multichaînes (il existe au moins une politique dont la chaîne de Markov correspondante soit constituée de deux classes récurrentes irréductibles ou plus). On suppose de plus ici que pour toute politique, la chaîne de Markov correspondante est apériodique.

#### 1.5.4.1. *Evaluation d'une politique markovienne stationnaire*

Soit $\pi \in \mathcal{D}^A$ une politique stationnaire et $(S, P_\pi, r_\pi)$ le processus de Markov valué qui lui est associé. Rappelons que le gain moyen ou critère moyen est défini par :

$$\forall s \in S \quad \rho^\pi(s) = \lim_{N \to \infty} E^\pi \left[ \frac{1}{N} \sum_{t=0}^{N-1} r_\pi(s_t) \mid s_0 = s \right].$$

Sous forme matricielle, on a ainsi

$$\rho^\pi = \lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} P_\pi^t r_\pi.$$

Soit $P_\pi^* = \lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} P_\pi^t$ la *matrice limite* de $P_\pi$. On montre que $P_\pi^*$ existe et est une matrice stochastique pour tout $S$ fini. De plus, $P_\pi^*$ vérifie

$$PP_\pi^* = P_\pi^* P = P_\pi^* P_\pi^* = P_\pi^*.$$

Le coefficient $P^*_{\pi,s,s'}$ peut être interprété comme la fraction de temps que le système passera dans l'état $s'$ en étant parti de l'état $s$. Pour des MDP apériodiques, on a de plus $P_\pi^* = \lim_{N\to\infty} P_\pi^N$ et $P^*_{\pi,s,s'}$ peut être interprété comme la probabilité à l'équilibre d'être dans l'état $s'$ en étant parti de $s$. Enfin, pour un MDP unichaîne, $P_\pi^*$ est alors la matrice dont toutes les lignes sont identiques et égales à la *mesure invariante* $\mu_\pi$ de la chaîne contrôlée par la politique $\pi$. Ainsi $\forall s, s'\ P^*_{\pi,s,s'} = \mu_\pi(s')$.

De la définition précédente de $\rho^\pi$, on déduit que $\rho^\pi = P_\pi^* r_\pi$. Pour un MDP unichaîne on établit ainsi que $\rho(s) = \rho$ est constant pour tout $s$, avec

$$\rho = \sum_{s\in S} \mu_\pi(s) r_\pi(s)$$

Dans le cas général d'un MDP multichaîne, $\rho(s)$ est constant sur chaque classe de récurrence.

Cette première caractérisation de $\rho_\pi$ fait intervenir $P_\pi^*$ qu'il n'est pas facile de calculer. Il est toutefois possible d'obtenir autrement $\rho_\pi$, en introduisant une nouvelle fonction de valeur pour le critère moyen, dite *fonction de valeur relative* :

DÉFINITION 1.10.– *Fonction de valeur relative pour le critère moyen*

$$\forall s \in S \quad U^\pi(s) = E^\pi[\sum_{t=0}^{\infty} (r_t - \rho^\pi) \mid s_0 = s]$$

En termes vectoriels, on a ainsi

$$\begin{aligned} U^\pi &= \sum_{t=0}^{\infty} P_\pi^t (r_\pi - \rho^\pi) \\ &= \sum_{t=0}^{\infty} P_\pi^t (r_\pi - P_\pi^* r_\pi) \\ &= \sum_{t=0}^{\infty} (P_\pi^t - P_\pi^*) r_\pi \\ &= (I - P_\pi^* + \sum_{t=1}^{\infty} (P_\pi - P_\pi^*)^t) r_\pi \\ &= (-P_\pi^* + (I - P_\pi + P_\pi^*)^{-1}) r_\pi, \end{aligned}$$

car on montre que la matrice $(I - P_\pi + P_\pi^*)$ est inversible et pour $t > 0$ : $(P_\pi - P_\pi^*)^t = P_\pi^t - P_\pi^*$. En multipliant à gauche par $(I - P_\pi + P_\pi^*)$, on en déduit

$$\begin{aligned} U^\pi &= (I - P_\pi + P_\pi^*)^{-1} (I - P_\pi^* (I - P_\pi + P_\pi^*)) r_\pi \\ &= (I - P_\pi + P_\pi^*)^{-1} (I - P_\pi^*) r_\pi. \end{aligned}$$

On note $H_{P_\pi} = (I - P_\pi + P_\pi^*)^{-1}(I - P_\pi^*)$ la *matrice de déviation* de $P_\pi$, qui est donc telle que

$$U^\pi = H_{P_\pi} r_\pi.$$

On peut vérifier que $H_{P_\pi}$ est matrice pseudo-inverse de $(I - P_\pi)$, ce qui établit un lien clair entre cette définition de $U^\pi$ et l'expression de $V_\gamma^\pi$ établie au théorème 1.3.

On montre alors le résultat général suivant, valable pour tout processus de Markov valué qu'il soit récurrent, unichaîne ou multichaîne :

THÉORÈME 1.9 [PUT 94].– *Soit $(S, P_\pi, r_\pi)$ un processus de Markov valué associé à une politique stationnaire $\pi \in \mathcal{D}^A$. Alors*

*1) si $\rho^\pi$ et $U^\pi$ sont le gain moyen et la fonction de valeur relative de $\pi$*
*a) $(I - P_\pi)\rho^\pi = 0$,*
*b) $\rho^\pi + (I - P_\pi)U^\pi = r_\pi$.*

*2) réciproquement, si $\rho$ et $U$ vérifient les deux égalités précédentes, alors*
*a) $\rho = P_\pi^* r_\pi = \rho^\pi$,*

*b)* $U = H_{P_\pi} r_\pi + u$, *où* $(I - P_\pi)u = 0$.
*c) si de plus* $P_\pi^* U = 0$, *alors* $U = H_{P_\pi} r_\pi = U^\pi$.

On retiendra que la fonction de valeur relative $U^\pi$ est l'unique solution de $(I - P_\pi)U = (I - P_\pi^*)r_\pi$ telle que $P_\pi^* U = 0$, obtenue en utilisant la pseudo-inverse $H_{P_\pi}$ de $(I - P_\pi)$.

Dans le cas simplifié d'un processus unichaîne, la première équation se simplifie en $\rho_\pi(s) = \rho_\pi$ et la seconde peut s'écrire selon :

$$\forall s \in S \quad U(s) + \rho = r_\pi(s) + \sum_{s' \in S} P_{\pi,s,s'} U(s'). \tag{1.4}$$

Toute solution $(\rho, U)$ de cette équation vérifie alors $\rho = \rho_\pi$ et $U = U_\pi + ke$, où $k$ est un scalaire quelconque et $e$ le vecteur dont toutes les composantes sont égales à 1. Si de plus $\sum_{s \in S} \mu_\pi(s) U(s) = 0$ alors $U = U_\pi$. Cette équation est bien sûr à rapprocher de celle établie pour le critère $\gamma$-pondéré.

1.5.4.2. *Equations d'optimalité*

Enonçons maintenant les conditions d'optimalité qu'il est possible d'établir pour le critère moyen. Rappelons que l'on recherche les politiques $\pi^* \in \Pi^{HA}$ telles que

$$\rho_{\pi^*} = \max_{\pi \in \Pi^{HA}} \rho_\pi = \rho^*$$

Le résultat principal énoncé dans le cadre général des MDP multichaînes est le suivant :

THÉORÈME 1.10 [PUT 94].– *Equations d'optimalité multichaîne*
*Il existe une solution* $(\rho, U)$ *au système d'équations définies pour tout* $s \in S$ *:*

$$\rho(s) = \max_{a \in A} \sum_{s' \in S} p(s' \mid s, a) \rho(s')$$

$$U(s) + \rho(s) = \max_{a \in B_s} \left( r(s,a) + \sum_{s' \in S} p(s' \mid s, a) U(s') \right)$$

*avec*

$$B_s = \left\{ a \in A \mid \sum_{s' \in S} p(s' \mid s, a) \rho(s') = \rho(s) \right\}$$

*On a alors* $\rho = \rho^*$.

Dans le cas unichaîne, le gain $\rho$ est constant et $B_s = A$. Les équations d'optimalité se réduisent alors à

$$\forall s \in S \quad U(s) + \rho = \max_{a \in A} \left( r(s,a) + \sum_{s' \in S} p(s' \mid s,a) U(s') \right) \tag{1.5}$$

Le lien entre solutions des équations d'optimalité et politiques optimales est alors établi avec le théorème suivant :

THÉORÈME 1.11 [PUT 94].– *Soit $(\rho, U)$ une solution aux équations d'optimalité. Il existe alors une politique stationnaire markovienne déterministe gain-optimale $\pi^* \in \mathcal{D}$, déterminée par :*

$$\forall s \in S \quad \pi^*(s) \in \operatorname*{argmax}_{a \in B_s} \left( r(s,a) + \sum_{s' \in S} p(s' \mid s,a) U(s') \right)$$

Notons qu'il peut exister des politiques gain-optimales $\pi^*$ telles que $(\rho^{\pi^*}, U^{\pi^*})$ ne vérifient pas les équations d'optimalité.

Remarquons enfin que les solutions des équations d'optimalité ne sont pas uniques, car si $(\rho^*, U)$ est solution, il en est au moins de même pour $(\rho^*, U + ke)$ pour tout scalaire $k$. Surtout, il peut y avoir plusieurs fonctions de valeur relative solutions, définissant des politiques différentes, associées au même $\rho^*$ optimal. Il est alors utile de rechercher parmi ces différentes solutions celles qui maximisent la fonction de valeur relative, on parle alors de *bias-optimality*.

## 1.6. Algorithmes de résolution des MDP

### 1.6.1. *Le critère fini*

Le cas de l'horizon fini est assez simple. Les équations d'optimalité permettent en effet de calculer récursivement à partir de la dernière étape les fonctions de valeur optimales $V_1^*, \ldots, V_N^*$ selon l'algorithme 1.1.

La complexité temporelle et spatiale de cet algorithme est en $O(N|S|^2|A|)$.

### 1.6.2. *Le critère $\gamma$-pondéré*

Trois grandes familles de méthodes existent pour résoudre de tels MDP : la programmation linéaire, l'itération sur les valeurs et l'itération sur les politiques. Toutes recherchent des politiques optimales dans $\mathcal{D}$.

**Algorithme 1.1** : Programmation dynamique à horizon fini

$V_0 \leftarrow 0$
**pour** $n \leftarrow 0$ **jusqu'à** $N-1$ **faire**
  **pour** $s \in S$ **faire**
    $V^*_{n+1}(s) = \max_{a \in A}\{r_{N-1-n}(s,a) + \sum_{s'} p_{N-1-n}(s'|s,a)V^*_n(s')\}$
    $\pi_{N-1-n}(s) \in$
    $\operatorname{argmax}_{a \in A}\{r_{N-1-n}(s,a) + \sum_{s'} p_{N-1-n}(s'|s,a)V^*_n(s')\}$
**retourner** $V^*$, $\pi^*$

1.6.2.1. *Programmation linéaire*

Il est immédiat de vérifier que si $V \in \mathcal{V}$ minimise la fonction $\sum_{s \in S} V(s)$ sous la contrainte $V \geq LV$, alors $V = V^*_\gamma$. En effet, nous avons montré au cours de la preuve du théorème 1.4 que $V \geq LV$ impliquait $V \geq V^*_\gamma$ et donc que $\sum_{s \in S} V(s) \geq \sum_{s \in S} V^*_\gamma(s)$. Une manière de rechercher la fonction de valeur optimale $V^*_\gamma$ est donc de résoudre le système linéaire associé, comme décrit dans l'algorithme 1.2 ci-dessous.

**Algorithme 1.2** : Programmation linéaire pour le critère $\gamma$-pondéré

résoudre

$$\min_{V \in \mathcal{V}} \sum_{s \in S} V(s)$$

avec

$$V(s) \geq r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a)V(s'), \quad \forall s \in S,\, a \in A$$

**pour** $s \in S$ **faire**
  $\pi(s) \in \operatorname{argmax}_{a \in A}\{r(s,a) + \gamma \sum_{s'} p(s'|s,a)V(s')\}$
**retourner** $V$, $\pi$

Cette approche a été proposée initialement par [D'E 63]. Si $n$ et $m$ sont les tailles respectives de $S$ et $A$, avec $p()$ et $r()$ codées sur $b$ bits, la complexité d'un tel algorithme de programmation linéaire sur les rationnels est polynomiale en $|S|$, $|A|$, $b$, avec des temps de résolution assez lents [LIT 95c]. Nous verrons toutefois au chapitre 9 que des méthodes de programmation linéaire peuvent s'avérer très efficaces dans le cadre des MDP admettant une représentation factorisée.

1.6.2.2. *Algorithme d'itération sur les valeurs*

L'approche la plus classique se base aussi sur la résolution directe de l'équation d'optimalité de Bellman $V = LV$, en utilisant pour cela une méthode itérative de type point fixe, d'où son nom anglais de *value iteration* [BEL 57, BER 87, PUT 94].

Comme le prouve le théorème 1.4, la solution de l'équation de Bellman est obtenue comme limite de la suite $V_{n+1} = LV_n$, quelle que soit l'initialisation de $V_0$. Il est alors établi qu'un nombre maximum d'itérations polynomial en $|S|$, $|A|$, $b$, $1/(1-\gamma)\log(1/(1-\gamma))$ est nécessaire pour atteindre $\pi^*$, chaque itération étant de complexité $O(|A||S|^2)$ [PAP 87]. Au delà de ce nombre d'itérations, la suite $V_n$ est de plus en plus proche de $V^*$ mais la politique correspondante $\pi_n = \pi^*$ ne change plus.

En pratique, plusieurs conditions d'arrêt de l'itération peuvent être envisagées. La plus classique consiste à stopper l'itération lorsque $\| V_{n+1} - V_n \| < \epsilon$, où $\epsilon$ est un seuil d'erreur fixé a priori. On aboutit à l'algorithme 1.3 suivant :

**Algorithme 1.3** : Algorithme d'itération sur les valeurs - Critère $\gamma$-pondéré

initialiser $V_0 \in \mathcal{V}$
$n \leftarrow 0$
**répéter**
  **pour** $s \in S$ **faire**
    $V_{n+1}(s) = \max_{a \in A} \left\{ r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V_n(s') \right\}$
  $n \leftarrow n+1$
**jusqu'à** $\| V_{n+1} - V_n \| < \epsilon$
**pour** $s \in S$ **faire**
  $\pi(s) \in \operatorname{argmax}_{a \in A} \{ r(s,a) + \gamma \sum_{s'} p(s'|s,a) V_n(s') \}$
**retourner** $V_n$, $\pi$

On montre alors que $\| V_n - V_\gamma^* \| < \epsilon'$ avec $\epsilon' = \frac{2\gamma}{1-\gamma}\epsilon$ (voir chapitre 11).

Il est possible d'améliorer la vitesse de convergence de l'algorithme d'itération sur les valeurs en modifiant légèrement le calcul de $V_{n+1}$. L'idée consiste à utiliser $V_{n+1}(s)$ à la place de $V_n(s)$ lorsque cette valeur a déjà été calculée. On définit ainsi l'algorithme de Gauss-Seidel, en numérotant les états de $S$ de 1 à $|S|$ (algorithme 1.4).

Cette idée peut encore être généralisée au cas où les états mis à jour à chaque itération sont sélectionnés aléatoirement parmi $S$. On définit ainsi la programmation dynamique asynchrone [BER 89].

Il est enfin possible d'optimiser encore l'algorithme en éliminant dès que possible des actions qui s'avèrent être définitivement non optimales. Cela permet ainsi de réduire la complexité de l'opération de maximisation sur $A$.

#### 1.6.2.3. *Algorithme d'itération sur les politiques*

La dernière classe importante d'algorithmes de résolution est constituée des méthodes itérant sur les politiques elles-mêmes.

**Algorithme 1.4** : Algorithme d'itération sur les valeurs - Gauss-Seidel

initialiser $V_0 \in \mathcal{V}$
$n \leftarrow 0$
**répéter**
  **pour** $i \leftarrow 1$ **jusqu'à** $|S|$ **faire**
    $V_{n+1}(s_i) = \max_{a \in A} \Big\{ r(s,a) + \gamma \sum_{1 \leq j < i} p(s_j \mid s,a) V_{n+1}(s_j) + \gamma \sum_{i \leq j \leq |S|} p(s_j \mid s,a) V_n(s_j) \Big\}$
  $n \leftarrow n+1$
**jusqu'à** $\| V_{n+1} - V_n \| < \epsilon$
**pour** $s \in S$ **faire**
  $\pi(s) \in \operatorname{argmax}_{a \in A} \{ r(s,a) + \gamma \sum_{s'} p(s'|s,a) V_n(s') \}$
**retourner** $V_n$, $\pi$

Considérons une politique stationnaire $\pi \in \mathcal{D}$ et $V_\gamma^\pi$ sa fonction de valeur. L'algorithme d'itération sur les politiques exploite la propriété suivante :

PROPRIÉTÉ 1.1.– *Amélioration sur 1 coup de la politique*
*Soit $\pi \in \mathcal{D}$. Toute politique $\pi^+$ définie par*

$$\pi^+ \in \operatorname*{argmax}_{\delta \in \mathcal{D}} \{ r_\delta + \gamma P_\delta V_\gamma^\pi \}$$

*vérifie*

$$V_\gamma^{\pi^+} \geq V_\gamma^\pi$$

*avec $V_\gamma^{\pi^+} = V_\gamma^\pi \Leftrightarrow \pi = \pi^*$.*

PREUVE.– On a

$$\begin{aligned} r_{\pi^+} + \gamma P_{\pi^+} V_\gamma^\pi &= \max_{\delta \in \mathcal{D}} \{ r_\delta + \gamma P_\delta V_\gamma^\pi \} \\ &\geq r_\pi + \gamma P_\pi V_\gamma^\pi \\ &\geq V_\gamma^\pi \end{aligned}$$

car $V_\gamma^\pi = r_\pi + \gamma P_\pi V_\gamma^\pi$. D'où

$$\begin{aligned} r_{\pi^+} + \gamma P_{\pi^+} V_\gamma^{\pi^+} + \gamma P_{\pi^+} (V_\gamma^\pi - V_\gamma^{\pi^+}) &\geq V_\gamma^\pi \\ V_\gamma^{\pi^+} - \gamma P_{\pi^+} V_\gamma^{\pi^+} &\geq V_\gamma^\pi - \gamma P_{\pi^+} V_\gamma^\pi \\ (I - \gamma P_{\pi^+}) V_\gamma^{\pi^+} &\geq (I - \gamma P_{\pi^+}) V_\gamma^\pi \\ V_\gamma^{\pi^+} &\geq V_\gamma^\pi \end{aligned}$$

car si $u \geq v$, $(I - \gamma P_{\pi^+})^{-1}u = u + \gamma P_{\pi^+}u + \gamma^2 P_{\pi^+}^2 u^2 \cdots \geq v + \gamma P_{\pi^+}v + \gamma^2 P_{\pi^+}^2 v^2 \cdots \geq (I - \gamma P_{\pi^+})^{-1}v$.

L'égalité n'est possible que si $\max_{\delta \in \mathcal{D}}\{r_\delta + \gamma P_\delta V_\gamma^\pi\} = V_\gamma^\pi$, soit $V_\gamma^\pi = V_\gamma^*$. $\square$

L'algorithme d'itération sur les politiques se décline donc ainsi (algorithme 1.5) : soit la politique $\pi_n$ à l'itération $n$. Dans une première étape, on résout le système d'équations linéaires $V_n = L_{\pi_n}V_n$ puis, dans un second temps, on améliore la politique courante en posant $\pi_{n+1} \in \text{argmax}_{\delta \in \mathcal{D}}\{r_\delta + \gamma P_\delta V_n\}$. On stoppe l'algorithme lorsque $\pi_n = \pi_{n+1}$.

La suite $V_n$, croissante et bornée par $V_\gamma^*$, converge. Comme il y a un nombre fini de politiques, la suite $\pi_n$ converge alors en un nombre fini d'itération. A la limite, $V_n = V_\gamma^*$ et $\pi_n$ est optimale.

**Algorithme 1.5** : Algorithme d'itération sur les politiques - Critère $\gamma$-pondéré

initialiser $\pi_0 \in \mathcal{D}$
$n \leftarrow 0$
**répéter**
  résoudre

$$V_n(s) = r(s, \pi_n(s)) + \gamma \sum_{s' \in S} p(s' \mid s, \pi_n(s))V_n(s'), \quad \forall s \in S$$

  **pour** $s \in S$ **faire**
    $\pi_{n+1}(s) \in \text{argmax}_{a \in A}\left\{r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s, a)V_n(s')\right\}$
  $n \leftarrow n + 1$
**jusqu'à** $\pi_n = \pi_{n+1}$
**retourner** $V_n$, $\pi_{n+1}$

La complexité de l'algorithme d'itération sur les politiques est en $O(|A||S|^2) + O(|S|^3)$ par itération, avec un nombre maximum d'itérations polynomial en $|S|$, $|A|$, $b$ à $\gamma$ constant [PAP 87].

Là aussi, il est possible d'améliorer l'efficacité de cet algorithme en simplifiant la phase d'évaluation de la politique courante $\pi_n$. Une approche classique consiste à résoudre l'équation $V_n = L_{\pi_n}V_n$ de manière itérative, comme pour l'itération sur les valeurs, mais à s'arrêter au bout d'un faible nombre d'itérations. L'utilisation de ce principe conduit à l'algorithme modifié d'itération sur les politiques (algorithme 1.6).

Cet algorithme combine les caractéristiques de l'itération sur les valeurs et de l'itération sur les politiques. Il converge pour tout $\delta$ vers une politique optimale pour

**Algorithme 1.6** : Algorithme modifié d'itération sur les politiques - Critère $\gamma$-pondéré

initialiser $V_0 \in \mathcal{V}$ tel que $LV_0 \geq V_0$
$flag \leftarrow 0$
$n \leftarrow 0$
**répéter**
  **pour** $s \in S$ **faire**
    $\pi_{n+1}(s) \in \text{argmax}_{a \in A} \left\{ r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V_n(s') \right\}$
    ($\pi_{n+1}(s) = \pi_n(s)$ si possible)
    $V_n^0(s) = \max_{a \in A} \left\{ r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V_n(s') \right\}$
  $m \leftarrow 0$
  **si** $\| V_n^0 - V_n \| < \epsilon$ **alors** $flag \leftarrow 1$
  **sinon**
    **répéter**
      **pour** $s \in S$ **faire**
        $V_n^{m+1}(s) = r(s, \pi_{n+1}(s)) + \gamma \sum_{s' \in S} p(s' \mid s, \pi_{n+1}(s)) V_n^m(s')$
      $m \leftarrow m + 1$
    **jusqu'à** $\| V_n^{m+1} - V_n^m \| < \delta$
    $V_{n+1} \leftarrow V_n^m$
    $n \leftarrow n + 1$
**jusqu'à** $flag = 1$
**retourner** $V_n$, $\pi_{n+1}$

$\epsilon \rightarrow 0$, sous l'hypothèse $LV_0 \geq V_0$. Cette condition est par exemple vérifiée pour le choix suivant de $V_0$ :

$$V_0(s) = \frac{1}{1-\gamma} min_{s' \in S} min_{a \in A} r(s', a)$$

pour tout $s \in S$.

En pratique, les algorithmes d'itération sur les politiques et, en particulier, l'algorithme modifié d'itération sur les politiques, apparaissent plus efficaces que les algorithmes d'itération sur les valeurs et doivent leur être préférés.

### 1.6.3. *Le critère total*

#### 1.6.3.1. MDP *positifs*

On montre pour les modèles bornés positivement que l'algorithme d'itération sur les valeurs converge de manière monotone vers $V^*$ sous l'hypothèse que $V_0$ vérifie $0 \leq V_0 \leq V^*$.

En ce qui concerne l'algorithme d'itération sur les politiques adapté au cas des MDP positifs (algorithme 1.7), on impose une condition sur $V_0$ qui implique que $V_n$ reste dans $\mathcal{V}^+$ pour tout $n$. Le calcul de $V_n$ peut être mené en forçant à 0 la valeur $V_n(s)$ pour tous les états récurrents de la chaîne définie par $P_{\pi_n}$. On montre alors que cet algorithme converge en un nombre fini d'itérations vers $V^*$ et $\pi^*$. De même, sous les hypothèses $LV_0 \geq V_0$ et $V_0 \leq V^*$, on montre que l'algorithme modifié d'itération sur les politiques converge vers une politique optimale. En pratique $V_0 = 0$ est une condition suffisante.

**Algorithme 1.7** : Algorithme d'itération sur les politiques - Critère total - MDP positifs

initialiser $\pi_0 \in \mathcal{D}$ avec $r_{\pi_0} \geq 0$
$n \leftarrow 0$
**répéter**
  calculer la solution minimale de

$$V_n(s) = r(s, \pi_n(s)) + \sum_{s' \in S} p(s' \mid s, \pi_n(s)) V_n(s'), \quad \forall s \in S$$

  **pour** $s \in S$ **faire**
    $\pi_{n+1}(s) \in \operatorname{argmax}_{a \in A} \left\{ r(s,a) + \sum_{s' \in S} p(s' \mid s, a) V_n(s') \right\}$
    ($\pi_{n+1}(s) = \pi_n(s)$ si possible)
  $n \leftarrow n + 1$
**jusqu'à** $\pi_n = \pi_{n+1}$
**retourner** $V_n$, $\pi_{n+1}$

#### 1.6.3.2. MDP *négatifs*

On montre que l'algorithme d'itération sur les valeurs converge de manière monotone vers $V^*$ pour toute condition initiale $V^* \leq V_0 \leq 0$.

Par contre, ce n'est pas le cas pour les algorithmes d'itération sur les politiques, qui peuvent s'arrêter sur des politiques sous-optimales. Il en est de même pour l'algorithme modifié d'itération sur les politiques.

### 1.6.4. *Le critère moyen*

En ce qui concerne le critère moyen, de même que pour le critère $\gamma$-pondéré, on dispose de nombreux algorithmes de programmation dynamique pour calculer des politiques gain-optimales. On présente ici les deux principaux, dans le cas simplifié de MDP unichaînes (toutes les politiques sont unichaînes) pour lesquels le gain moyen $\rho$ est constant. Le test d'arrêt est ici basé sur l'emploi de la semi-norme span sur $\mathcal{V}$ : $\forall V \in \mathcal{V}$, $\operatorname{span}(V) = \max_{s \in S} V(s) - \min_{s \in S} V(s)$. Contrairement à $\| V \|$ qui

mesure l'écart de $V$ à 0, la semi-norme $\mathrm{span}(V)$ mesure l'écart de $V$ à un vecteur constant.

1.6.4.1. *Algorithme d'itération sur les valeurs relatives*

L'algorithme 1.8 est un algorithme d'itération sur les valeurs relatives $U(s) = V(s) - \rho$.

**Algorithme 1.8** : Algorithme d'itération sur les valeurs relatives - Critère moyen

initialiser $U_0 \in \mathcal{V}$
choisir $s^* \in S$
$n \leftarrow 0$
**répéter**
  $\rho_{n+1} = \max_{a \in A} \left\{ r(s^*, a) + \sum_{s' \in S} p(s' \mid s^*, a) U_n(s') \right\}$
  **pour** $s \in S$ **faire**
    $U_{n+1}(s) = \max_{a \in A} \left\{ r(s, a) + \sum_{s' \in S} p(s' \mid s, a) U_n(s') \right\} - \rho_{n+1}$
  $n \leftarrow n + 1$
**jusqu'à** $\mathrm{span}(U_{n+1} - U_n) < \epsilon$
**pour** $s \in S$ **faire**
  $\pi(s) \in \mathrm{argmax}_{a \in A} \{ r(s, a) + \sum_{s'} p(s' | s, a) U_n(s') \}$
**retourner** $\rho_n$, $U_n$, $\pi$

Sous différentes hypothèses techniques, on peut montrer sa convergence pour $\epsilon \rightarrow 0$ vers une solution $(\rho^*, V^*)$ des équations d'optimalité (1.5) et donc vers une politique optimale $\pi^*$ (voir [PUT 94], théorème 8.5.3). Pour un MDP unichaîne, c'est le cas par exemple si $p(s \mid s, a) > 0$ pour tout $s$ et $a$.

1.6.4.2. *Algorithme modifié d'itération sur les politiques*

L'algorithme 1.9 est un algorithme modifié d'itération sur les politiques, qui ne nécessite pas la résolution de l'équation (1.4) pour évaluer la fonction de valeur relative.

Pour $\delta$ élevé, l'algorithme est équivalent à l'itération sur les valeurs (non relative car on ne gère pas ici explicitement le revenu moyen $\rho_n$). Pour $\delta$ proche de 0, on retrouve une itération sur les politiques classique. Sous les mêmes conditions techniques précédentes, on montre que cet algorithme converge pour tout $\delta$ vers une politique optimale pour $\epsilon \rightarrow 0$. Plus précisément, lorsque l'algorithme s'arrête, on a

$$\min_{s \in S}(V_n^0(s) - V_n(s)) \leq \rho^{\pi_{n+1}} \leq \rho^* \leq \max_{s \in S}(V^0(s) - V_n(s)),$$

ce qui assure $\mid \rho^{\pi_{n+1}} - \rho^* \mid \leq \epsilon$.

**Algorithme 1.9** : Algorithme modifié d'itération sur les politiques - Critère moyen

initialiser $V_0 \in \mathcal{V}$
$flag \leftarrow 0$
$n \leftarrow 0$
**répéter**
  **pour** $s \in S$ **faire**
    $\pi_{n+1}(s) \in \text{argmax}_{a \in A} \left\{ r(s,a) + \sum_{s' \in S} p(s' \mid s,a) V_n(s') \right\}$
    ($\pi_{n+1}(s) = \pi_n(s)$ si possible)
    $V_n^0(s) = \max_{a \in A} \left\{ r(s,a) + \sum_{s' \in S} p(s' \mid s,a) V_n(s') \right\}$
  $m \leftarrow 0$
  **si** $\text{span}(V_n^0 - V_n) < \epsilon$ **alors** $flag \leftarrow 1$
  **sinon**
    **répéter**
      **pour** $s \in S$ **faire**
        $V_n^{m+1}(s) = r(s, \pi_{n+1}(s)) + \sum_{s' \in S} p(s' \mid s, \pi_{n+1}(s)) V_n^m(s')$
      $m \leftarrow m + 1$
    **jusqu'à** $\text{span}(V_n^{m+1} - V_n^m) < \delta$
    $V_{n+1} \leftarrow V_n^m$
    $n \leftarrow n + 1$
**jusqu'à** $flag = 1$
**retourner** $V_n$, $\pi_{n+1}$

## 1.7. Conclusion et perspectives

Le cadre des processus décisionnels de Markov, avec les modèles de décision, critères d'optimalité et algorithmes d'optimisation que nous venons de présenter constitue un outil méthodologique de base en intelligence artificielle. Il est en particulier devenu incontournable pour concevoir et analyser les méthodes formelles développées aujourd'hui sur le thème de la décision séquentielle dans l'incertain.

Malgré sa généralité, le cadre théorique que nous avons exposé au cours de ce chapitre n'est toutefois pas exempt de limites en termes théoriques. Tout d'abord, ce cadre suppose de la part de l'agent une parfaite connaissance des fonctions de transition et de récompense qui définissent le problème auquel il est confronté. Nous verrons au chapitre 2 comment l'apprentissage par renforcement permet de relâcher cette hypothèse. Par ailleurs, on suppose que l'agent a directement accès à son état. Or, dans la plupart des situations où l'on représente un agent en interaction avec son environnement, l'agent ne dispose pas d'un tel accès à son état, mais plutôt à des perceptions différenciées qui le renseignent exhaustivement ou non sur sa situation vis-à-vis de son environnement. Nous verrons au chapitre 3 comment formaliser une observation partielle de l'état du monde. De même, une autre limite formelle concerne le caractère

mono-agent du cadre des MDPs. De nombreux problèmes requièrent la modélisation de plusieurs agents évoluant et agissant ensemble au sein du même environnement. Nous présenterons dans les chapitres 4 et 8 les travaux qui étendent les MDPs au cadre multi-agents. Enfin, une autre limitation théorique du cadre des MDP provient de l'expression du critère à optimiser, sous la forme de l'espérance d'une somme de récompenses à maximiser. Nous verrons alors dans le chapitre 5 comment il est possible d'étendre les représentations de l'incertitude et des préférences de l'agent à d'autres formalismes.

Les utilisations de plus en plus nombreuses du cadre des MDP et de ses extensions évoquées ci-dessus pour aborder des problèmes de décision dans des domaines finalisés variés, qui vont de la gestion indutrielle aux agro-écosystèmes en passant par la robotique et les applications militaires, ont amené à considérer de manière de plus en plus sérieuse la question de l'efficacité des algorithmes de résolution proposés. Plusieurs chapitres de cet ouvrage seront ainsi consacrés à des méthodes récentes permettant de dépasser les limitations traditionnelles des algorithmes de programmation dynamique, dont les principales sont l'approximation de la fonction de valeur (chapitre 11), les représentations factorisées (chapitre 9), l'optimisation de politiques paramétrées (chapitre 12) ou encore l'optimisation de décision en ligne (chapitre 10).