REPUBLIQUE DU CAMEROUN
****
Paix-Travail-Patrie
****
UNIVERSITÉ DE YAOUNDÉ I
****
ÉCOLE NORMALE SUPÉRIEURE
DE YAOUNDÉ I
****
DEPARTEMENT DE MATHÉMATIQUES
****

N
E S

REPUBLIC OF CAMEROON
****
Peace-Work-Fatherland
****
UNIVERSITY OF YAOUNDE I
****
HIGHER TEACHER'S TRAINING
COLLEGE OF YAOUNDE I
****
DEPARTMENT OF MATHEMATICS
****

CONDITIONS D'EXISTENCE DES POLITIQUES OPTIMALES DANS UN PMDP À HORIZON FINI : CAS DES PRÉFÉRENCES COMPLÈTES

Mémoire de DIPES II de mathématiques

De

MBATCHOU Alain Cédrick

***Matricule :* 11Y651**

Licencié en Mathématiques

Sous la direction de :

Dr NJANPONG NANA Gilbert

Chargé de cours

École Nationale Supérieure Polytechnique, Université de Yaoundé I

Année académique : 2015-2016

CONDITIONS D'EXISTENCE DES POLITIQUES OPTIMALES DANS UN PMDP À HORIZON FINI : CAS DES PRÉFÉRENCES COMPLÈTES

Mémoire de DIPES II de mathématiques

De

MBATCHOU Alain Cédrick

Matricule: **11Y651**

Licencié en Mathématiques

Sous la direction de :

Dr NJANPONG NANA Gilbert

Chargé de Cours

École Nationale Supérieure Polytechnique, Université de Yaoundé I

Année Académique **2015-2016**

♣ Dédicace ♣

Je dédie ce mémoire à mon père ***TONFEU Honoré****.*

♣ Remerciements ♣

Je tiens à exprimer mes vifs remerciements et toute ma reconnaissance :

♡ À **Dieu** tout puissant pour sa charité accordée à mon éducation depuis ma naissance ;

♡ Au Professeur **ANDJIGA Nicolas Gabriel** grâce à qui ce travail a été effectué au sein de l'équipe de Mathématiques Appliquées aux Sciences Sociales (MASS) de l'Université de Yaoundé 1 ; ceci dans le cadre du projet de diversification des domaines de recherche ;

♡ Au Docteur **NJANPONG NANA Gilbert** qui n'a ménagé aucun effort pour la direction de ce travail. Sa disponibilité, ses critiques, ses suggestions, ses remarques et ses orientations pertinentes ont permis d'éviter bien des écueils ;

♡ Au Professeur **MOYOUWOU Issofa** qui m'a souvent prodigué des conseils relativement à ce mémoire ;

♡ À tous les enseignants du département de mathématiques de l'École Normale Supérieure de Yaoundé qui m'ont bien éduqués depuis mon entrée dans cette prestigieuse école ;

♡ À ma très chère mère, maman **TCHOUTA Yvette** qui m'a toujours soutenu financièrement, m'a donné le goût du savoir, m'a enseigné la sagesse, la tolérance et la patience ;

♡ À toute ma famille pour son soutien moral, matériel et financier qu'elle a eu à faire jusqu'ici pour ma réussite académique. Particulièrement à ma grand-mère, maman **NDJEUGOUA Julienne**, mes tantes **TATFO Rameline** et **YIMGA Merline** ;

♡ À tous mes amis et camarades de promotion en particuliers **FEUDJO JEMELE, NGUEFO TAKONGMO Amour, ABOUGNE Luc, SAFOKEM Adin, DJOUMESSI Joseph, TCHUISSEU Feraud** pour l'esprit d'équipe ;

♡ À tous ceux qui de près ou de loin ont contribué à la réalisation de ce travail.

♣ Table des matières ♣

♣ Déclaration sur l'honneur ♣

Le présent document est une œuvre originale du candidat et n'a été soumis nulle part ailleurs en partie ou en totalité, pour une autre évaluation académique. Les contributions externes ont été dûment mentionnées et recensées en bibliographie.

Signature du candidat

MBATCHOU Alain Cédrick

♣ Résumé ♣

Le modèle de processus de décision markovien probabiliste (PDMP) est le modèle standard pour la résolution des problèmes de planification dans l'incertain afin de trouver une politique optimale. L'objet principal de ce mémoire est de déterminer les conditions d'existence d'une politique optimale dans un PDMP à horizon fini dans le cadre des préférences complètes.
Nous étudions l'équation de Bellman pour le critère fini et énonçons les propriétés suffisantes (hypothèse de complétude, transitivité, invariance par translation, indépendance) des relations de préférences sur les loteries pour permettre l'utilisation des méthodes fondées sur la programmation dynamique. Nous fournissons enfin un exemple d'application de l'approche par la fonction de valeur pour le critère fini.

Mots clés : PDMP, politique optimale, fonction de valeur, loterie, programmation dynamique.

♣ Abstract ♣

The model of probabilist Markov decision process (PMDP) is the standard model used in solving uncertain planification problems in order to find the optimal policy. This dissertation aims at determining the existence's conditions of one optimal policy in a PMDP at a finite horizon within the framework of the complete preferences. We study Bellman's equation in finite criterion and we state sufficient properties (hypothesis of completion, transitivity, invariance by translation, independence) of preferred relations on lotteries to allow the use of founded methods on dynamic programming. Finally, we provide an example of application of the approach by the value function in finite criterion.

Keys words : PMDP, optimal policy, value function, lottery, dynamic programming.

♣ Liste des abréviations ♣

- **argmax** : argument de maximun.
- **MDP** : Markov decision process.
- **PDMP** : processus de décision Markovien probabiliste.
- **PMDP** : probabilist Markov decision process.

♣ Table des figures ♣

♣ Liste des tableaux ♣

♣ Introduction générale ♣

Un processus de décision markovien probabiliste (PDMP) est un modèle stochastique issu de la théorie de la décision et la théorie des probabilités. Le modèle PDMP peut être vu comme une chaîne de Markov à laquelle on ajoute deux composantes décisionnelles (l'ensemble des actions possibles et la fonction de récompense). Comme les autres modèles de sa famille, il est entre autres utilisé en robotique, en recherche opérationnelle, en intelligence artificielle pour le contrôle des systèmes complexes comme les agents intelligents, etc. Un PDMP permet de prendre des décisions dans un environnement lorsque l'on a une incertitude sur l'état dans lequel on se trouve et aussi sur l'effet des actions entreprises.

Le modèle des processus de décision markoviens a été assez peu étudié du point de vue des structures de préférence. A notre connaissance, les travaux les plus généraux dans cette optique sont ceux de Sobel 1975. En identifiant un PDMP à un problème déterministe défini sur les distributions de probabilité sur l'ensemble d'états, il montre que sous certains axiomes, il est possible d'appliquer un algorithme de type itération de la politique pour déterminer des politiques optimales. Mais ces résultats sont difficilement applicables dans la pratique car les politiques sont de forme complexe, définies sur des espaces infinis et de ce fait, n'ont pas d'interprétation évidente. On peut néanmoins mentionner quelques études utilisant des préférences non classiques dans le cadre des PDMP, comme Cavazos-Cadena et al. 2000 qui utilisent un critère d'utilité sensible au risque. Dans ces travaux, l'hypothèse de l'existence de coûts scalaires numériques est conservée. Par ailleurs, dans certains travaux , on utilise des valuations numériques non scalaires (vecteurs de réels), donnant naissance aux PDMP multicritères. D'autres travaux ont étudié des PDMP exploitant des préférences qualitatives (Bonet et al. 2002). Leur modèle des processus de décision markoviens qualitatifs traite les problèmes pour lesquels l'information sur les données numériques n'est pas assez riche. Ils utilisent alors des ordres de grandeur pour les probabilités et la fonction de coût. Quand l'incertain est modélisé par la

théorie des possibilités, Sabbadin et al. 1998 et 1999 ont étendu les critères qualitatifs, définis axiomatiquement par Dubois et al. 1995. Les processus de décision markoviens font partie d'une classe plus large de problèmes de décision dynamique dans l'incertain. Dans le cadre des PDMP, la notion de cohérence dynamique est proche du principe de Bellman qui dit que toute sous politique d'une politique optimale est optimale. Dans tous ces travaux, il est supposé que les relations de préférence sont complètes et numériquement représentées.

Ce mémoire cherche à déterminer les conditions d'existence des politiques optimales en explicitant l'approche par la fonction de valeur et la résolution par les loteries, à horizon fini. Nos travaux se démarquent de ceux évoqués précédemment sur le fait qu'avec l'approche par les loteries, nous ne supposons pas nécessairement l'existence de récompenses additives. Pour ce fait, notre travail s'articule autour de quatre chapitres. Le chapitre un, présente le cadre général de la notion de PDMP ainsi que les outils devant nous aider dans la suite du travail. Dans le chapitre deux, consacré aux méthodes de résolutions, nous recherchons les conditions d'existence d'une politique optimale. Dans le chapitre trois, nous présentons un exemple pratique de PDMP illustrant l'approche par la fonction de valeur pour le critère fini et enfin au chapitre quatre, nous présentons quelques implications pédagogiques de notre travail.

⋆ ⋆ **Chapitre Un** ⋆ ⋆

PRÉLIMINAIRES ET NOTION DE PDMP

Dans ce chapitre, nous présentons quelques outils qui seront très utiles dans la résolution du problème fondamental et dégageons quelques résultats.

1.1 Préliminaires

Soient X et Y deux ensembles non vides.

Définition 1.1. *Une **relation binaire** de X vers Y est une partie R de $X \times Y$. Si $(x, y) \in R$, on dit que x est en relation avec y et on note xRy, sinon on dit que x n'est pas en relation avec y et on note $non(xRy)$. Si $X = Y$, on dit que R est une relation binaire sur X.*

Définition 1.2. *Une **relation de préférence** (ou **préférence**) sur X est une relation binaire sur X indiquant la manière dont les éléments de X sont comparés entre-eux. Si R est une préférence sur X alors $\forall x, y \in X$, xRy signifie que " x est préféré (au sens large) à y " ou que "x est au moins aussi bon que y " et se note $x \succsim_R y$.*
Soit R une préférence sur X. La partie stricte de R sera notée $\underset{R}{\succ}$ et l'indifférence de R notée $\sim_R$ et définies respectivement par :

(i) $\forall x, y \in X$, $x \underset{R}{\succ} y$ *si* $x \succsim_R y$ *et non* $(y \succsim_R x)$;

(ii) $\forall x, y \in X$, $x \sim_R y$ *si* $x \succsim_R y$ *et* $y \succsim_R x$.

L'écriture $x \underset{R}{\succ} y$ signifie que " x est strictement meilleur que y et $x \sim_R y$ signifie qu'il y a "indifférence entre x et y".

Définition 1.3. *Une relation binaire R sur X est dite :*

(i) ***réflexive*** *si* $\forall x \in X,\ x \succsim_R x$ *;*

(ii) ***symétrique*** *si* $\forall x, y \in X,\ x \succsim_R y \Rightarrow y \succsim_R x$ *;*

(iii) ***antisymétrique*** *si* $\forall x, y \in X,\ (x \succsim_R y$ *et* $y \succsim_R x) \Rightarrow x = y$ *;*

(iv) ***transitive*** *si* $\forall x, y, z \in X,\ (x \succsim_R y$ *et* $y \succsim_R z\ ) \Rightarrow x \succsim_R z$ *;*

(v) ***complète*** *ou* ***totale*** *si* $\forall x, y \in X,\ x \succsim_R y$ *ou* $y \succsim_R x$ *;*

(vi) ***partielle*** *si* $\exists x, y \in X$, *non* $(x \succsim_R y)$ *et non* $(y \succsim_R x)$.

Définition 1.4. *On dit qu'une relation binaire* R *sur* X *est :*

(i) une ***relation d'équivalence*** *sur* X*, si elle est réflexive, symétrique et transitive ;*

(ii) un ***préordre*** *(****totale****) sur* X*, si* R *est réflexive, transitive (et totale) ;*

(iii) un ***ordre*** *(****totale****) sur* X*, si elle est réflexive, antisymétrique, transitive (et total).*

Définition 1.5. *Soit* R *une relation d'ordre sur* X *et* A *une partie non vide de* X.

(i) On appelle ***maximun*** *de* A *tout élément* $x \in A$ *vérifiant :*
$(\forall y \in X)\ (y \in A \Rightarrow x \underset{R}{\succsim} y)$.

(ii) On appelle ***minimun*** *de* A *tout élément* $x \in A$ *vérifiant :*
$(\forall y \in X)\ (y \in A \Rightarrow y \underset{R}{\succsim} x)$.

(iii) On appelle ***élément maximal*** *de* A *tout élément* $x \in A$ *vérifiant :*
$(\forall y \in X)\ (y \in A$ *et* $yRx \Rightarrow y \sim_R x)$.

(iv) On appelle ***élément minimal*** *de* A *tout élément* $x \in A$ *vérifiant :*
$(\forall y \in X)\ (y \in A$ *et* $xRy \Rightarrow x \sim_R y)$.

1.2 Notion de PDMP

1.2.1 Processus stochastique

Soit $I \subseteq \mathbb{R}$ et non vide. Soient $(\Omega, \mathcal{T}(\Omega), P)$ un espace probabilisé et $(\mathcal{S}, \mathcal{T}(\mathcal{S}))$ un espace probabilisable. Pour tout $t \in I$, on définit l'application $X_t : \Omega \to \mathcal{S}$.

Définition 1.6. *Un* ***processus stochastique*** *est une suite de variables aléatoires* $(X_t)_{t \in I}$*, qui décrit l'évolution d'un phénomène aléatoire. Autrement dit, c'est une famille de variable aléatoire indexée par un sous ensemble de* $\mathbb{R}$ *ou de* $\mathbb{N}$*, souvent assimilé à l'espace temporel.*

Pour un t fixé dans I, la variable aléatoire X_t représente l'état du processus au temps t et $\mathcal{S}$ est l'ensemble de toutes les valeurs possibles pour cette variable. On travaillera en temps discret (indexé par des entiers) et on supposera $\mathcal{S}$ fini.

Remarque: 1.2.1. $\mathcal{S}$ est l'ensemble des états du processus et I est l'ensemble des instants ou étapes d'observation du processus.

Exemple 1.2.1. Voici quelques exemples de processus stochastiques :

- en **météorologie**, X_t peut désigner la température en un lieu donné au jour t, ou encore la hauteur des précipitations pour l'année t ;
- en **bourse de valeur**, X_t peut désigner la valeur sur le marché d'une action pour le jour t ou le chiffre d'affaire d'une société pour l'année t ;
- en **assurance**, X_t peut désigner le montant des indemnités versées par une compagnie d'assurance pour des sinistres survenus le mois t ;
- en **épidémiologie**, X_t peut désigner le nombre d'individus infectés par une maladie contagieuse au bout de t jours.

1.2.2 Propriété de Markov et Matrice de transition

Soient (Ω, $\mathcal{T}(\Omega)$, P) un espace probabilisé, ($\mathcal{S}$, $\mathcal{T}(\mathcal{S})$) un espace probabilisable et $(X_t)_{t\in I}$ un processus stochastique.

Définition 1.7. *Un processus stochastique $(X_t)_{t\in I}$ vérifie la propriété de Markov lorsque son évolution après une date t, ne dépend du passé X_0, X_1, X_2, ..., X_{t-1} qu'à travers sa position au temps t (et non pas du trajet qu'il a suivi pour atteindre cet état).*

Relativement à la notion de probabilité conditionnelle $P(B|A)$, nous nous intéresserons au cas où $A = \{X_0, X_1, X_2, ..., X_t\}$ et $B = \{X_{t+1} = s_{t+1}\}$.

Définition 1.8. *Formellement, un processus stochastique $(X_t)_{t\in I}$ vérifie la **propriété de Markov** si $\forall\ s_i, s_j, s_0, ..., s_{t-1} \in \mathcal{S}$:*

$P(X_{t+1} = s_i \mid X_0 = s_0, ..., X_{t-1} = s_{t-1}, X_t = s_j) = P(X_{t+1} = s_i \mid X_t = s_j) = P_{s_i s_j}$ *où*

- $P(X_{t+1} = s_i \mid X_0 = s_0, ..., X_{t-1} = s_{t-1}, X_t = s_j)$ *est la probabilité que le processus passe à l'état s_i à l'instant $t+1$ sachant qu'il a été successivement aux états s_0, ..., s_{t-1}, s_j pendant les instants 0, 1, ..., $t-1$, t ;*

- $P(\ X_{t+1} = s_i \mid X_t = s_j)$ *est la probabilité que le processus passe à l'état* s_i*, à l'instant* $t+1$ *sachant qu'il est à l'état* s_j*, à l'instant* t*.*

Interprétation 1.1. L'évolution future ne dépend que de l'état actuel du système. On peut constater que la dynamique du système n'intègre pas la mémoire de l'historique des états passés. Dans un tel processus, dit processus sans mémoire, la prédiction du futur à partir du présent ne nécessite pas la connaissance du passé.

1.2.3 Chaîne de Markov

Définition 1.9. *Une* ***chaîne de Markov*** *est un processus stochastique* $(X_t)_{t\in\mathbb{I}}$ *qui vérifie la propriété de Markov.*

En d'autres termes, une chaîne de Markov est une suite d'états stochastiques, tels que la probabilité d'obtenir un état à un instant donné, ne dépend que de l'état précédent de la chaîne.

Notation 1.2.1. Pour tout $s_i \in \mathcal{S}$, $(P_{s_i s_j})_{s_j\in\mathcal{S}}$ désigne les probabilités de **transition** pour l'état s_i. Si l'espace des états $\mathcal{S}$ est fini, $(P_{s_i s_j})_{s_i,s_j\in\mathcal{S}}$ est alors appelée **matrice de transition** de la chaîne.

Remarque: 1.2.2.

- Pour tout s_i, $s_j \in \mathcal{S}$, $P_{s_i s_j}$ est la probabilité de transition d'état s_i à l'instant t vers l'état s_j à l'instant $t+1$.
- Chaque ligne de la matrice de transition représente une loi de probabilité, d'où la relation $\forall s_i \in \mathcal{S},\ \sum_{s_j\in S} P_{s_i s_j} = 1$. Son ordre est égal au cardinal de l'espace des états S.

Exemple 1.2.2. Chaîne à deux états : État d'une ligne de téléphone

Considérerons l'état d'une ligne de téléphone $X_t = 0$ si la ligne est libre à l'instant t et $X_{t+1} = 1$ si la ligne est occupée. Supposons que sur chaque intervalle de temps, il y a une probabilité p qu'un appel arrive (un appel au plus) et si la ligne est déjà occupée, l'appel est perdu. Supposons également que si la ligne est occupée au temps t, il y a une probabilité q qu'elle se libère au temps $t+1$.

Cherchons la matrice de transition de ce processus stochastique.

On peut alors modéliser une chaîne de Markov à valeurs dans $\mathcal{S} = \{s_0; s_1\}$ avec $s_0 = 0$ désignant l'état libre de la ligne, $s_1 = 1$ désignant l'état occupé de la ligne et ayant pour matrice de transition :

$$(P_{s_i s_j})_{s_i, s_j \in \mathcal{S}} = \begin{pmatrix} P_{s_0 s_0} & P_{s_1 s_0} \\ P_{s_0 s_1} & P_{s_1 s_1} \end{pmatrix} = \begin{pmatrix} 1-p & P \\ q & 1-q \end{pmatrix}$$

où :

- $P_{s_0 s_0}$ est la probabilité qu'à l'instant $t+1$, la ligne soit libre sachant qu'elle était également libre à l'instant t ;
- $P_{s_0 s_1}$ est la probabilité qu'à l'instant $t+1$, la ligne soit libre sachant qu'elle était occupée à l'instant t ;
- $P_{s_1 s_0}$ est la probabilité qu'à l'instant $t+1$, la ligne soit occupée sachant qu'elle était libre à l'instant t ;
- $P_{s_1 s_1}$ est la probabilité qu'à l'instant $t+1$, la ligne soit occupée sachant qu'elle était également occupée à l'instant t.

1.2.4 Des chaînes de Markov aux PDMP

Les chaînes de Markov permettent de modéliser les systèmes stochastiques mais n'autorisent pas un agent à intervenir et à agir sur l'évolution du système. Considérons un agent autonome qui évolue dans un espace d'états $\mathcal{S}$ de dimension finie, et dont les effets des actions sont stochastiques et markoviens ; la probabilité d'arriver dans un état s' ne dépend que de l'état précédent s. Ainsi pour chaque action, la dynamique de l'agent est une chaîne de Markov. À chaque observation de l'état courant, l'agent décide d'appliquer une action qui implique un changement de chaîne de Markov pour l'ensemble de ses états. Rajoutons à ce formalisme des récompenses sur les transitions entre les états de la chaîne de Markov contrôlée ; l'agent peut alors décider des actions optimales à appliquer, de sorte à optimiser un critère numérique basé sur les récompenses de la chaîne. La chaîne de Markov contrôlée à laquelle des récompenses sont associées aux transitions stochastiques est appelée **processus de décision Markovien probabiliste (PDMP)**.

Agent

Définition 1.10. *Un **agent** est tout ce qui peut être considéré comme percevant son environnement grâce à des capteurs et agissant sur ce même environnement par des actionneurs.*

Un agent est donc défini, comme une entité qui reçoit des informations de son environnement et qui effectue des actions pouvant modifier l'état de celui-ci. Cependant, analyser les observations de l'environnement et procéder en conséquence à des actions, nécessite une troisième faculté, celle de raisonnement ; ce qui nous amène à élargir la définition d'un agent à un triplet : **perception, raisonnement, action**.

Exemple 1.2.3. Un robot, un opérateur économique et un homme politique sont quelques exemples d'agents dans un PDMP.

États

Définition 1.11. *Un **état** est un ensemble de descripteurs (de l'environnement) permettant de décrire la configuration de la nature à un instant donné.*

Suivant la propriété de Markov (la transition dépend uniquement de l'état courant), chaque état doit contenir l'information nécessaire à la prédiction de l'évolution du système et par conséquent à la prise de décision. L'agent peut avoir une connaissance totale ou partielle de l'état dans lequel il se trouve : on parle alors d'observabilité totale ou d'observabilité partielle de l'état.
Dans le cadre de ce travail, nous allons nous intéresser à des problèmes décisionnels dans lesquels un agent détermine comment agir afin de maximiser ses récompenses. Par conséquent, nous étudierons plus particulièrement les problèmes avec une observabilité totale.

Exemple 1.2.4. La faillite d'une banque, la crise économique et l'état d'une voiture sont quelques exemples d'états dans un PDMP.

Actions

Définition 1.12. *Une **action** est une opération qui permet la transition d'un état s à un état s' et peuvent conduire à différents états selon une distribution de probabilité sur $\mathcal{S}$.*

Notons que le modèle des PDMP considère que la durée des transitions n'est pas nécessairement identique.

Exemple 1.2.5. La création d'emplois, l'augmentation du salaire et l'achat d'une voiture neuve sont quelques exemples d'actions dans un PMDP.

Fonction de transition

On appelle **fonction de transition** l'application

$$T : \begin{array}{ccc} \mathcal{S} \times \mathcal{A} & \longrightarrow & \Pi(\mathcal{S}) \\ (s,a) & \mapsto & T(s,a) = P(X_t = s'|s,a) \end{array}$$

où $s' \in S$, désigne un nouvel état du processus.

Remarque: 1.2.3. P (X_t = s' ǀ s, a) est la probabilité que le processus passe à l'état s' à l'instant $t+1$ sachant qu'à l'instant t, il est à l'état s et l'action à appliquer est a.

La fonction T décrit l'évolution du système et vérifie la propriété de Markov ; elle se résume alors par l'équation suivante :

$\forall s' \in \mathcal{S}, \forall (s_0, a_0), (s_1, a_1)$, ..., $(s_t, a_t) \in \mathcal{S} \times \mathcal{A}$,

$$P(\ X_{t+1} = s' \text{ ǀ } s_0, a_0, s_1, a_1, ..., s_t, a_t\ ) = P(X_{t+1} = s' \text{ ǀ} s_t, a_t\ ) \quad \text{où}$$

$P(\ X_{t+1} = s' \text{ ǀ } s_0, a_0, s_1, a_1, ..., s_t, a_t\ )$ est la probabilité que le processus passe à l'état s' à l'instant $t+1$ sachant que le processus a traversé successivement les états s_i pendant l'exécution des actions a_i ($i \in \{0, 1, 2, ...,$t$\}$), respectivement pendant les instants 0, 1, 2, ..., t ;

Notation 1.2.2. P (X_{t+1} = s' ǀ s_t, a_t) sera encore noté P (s' ǀ s, a) .

Remarque: 1.2.4. La fonction de transition indique la probabilité de transition entre les états. Étant donné qu'elle produit une distribution de probabilité sur $\mathcal{S}$, on a :

$$\forall a \in \mathcal{A}, \forall s \in \mathcal{S}, \sum_{s' \in \mathcal{S}} P(s'|s,a) = 1\ .$$

Exemple 1.2.6. La **figure 1.1**, montre un exemple simple de représentation de la fonction de transition d'un modèle $(\mathcal{S}, \mathcal{A}, T, R)$ de PDMP où $\mathcal{S} = \{s_0, s_1, s_2, s_3\}$ et $\mathcal{A} = \{a_0, a_1\}$.

- Lorsque l'agent se trouve dans l'état s_0, il peut accomplir deux actions a_0 ou a_1.
- Lorsqu'il effectue l'action a_0 à l'instant $t = n$ $(n \in \mathbb{N})$, il a une probabilité $P(s_1|s_0, a_0)$ d'arriver dans l'état s_1 et une probabilité $P(s_2|s_0, a_0)$ d'arriver dans l'état s_2 à l'instant $t = n+1$ avec $P(s_1|s_0, a_0) + P(s_2|s_0, a_0) = 1$.
- Lorsqu'il effectue l'action a_1 à l'instant $t = n$, il ne peut se retrouver que dans l'état s_3 à l'instant $t = n+1$, donc $P(s_3|s_0, a_1) = 1$.

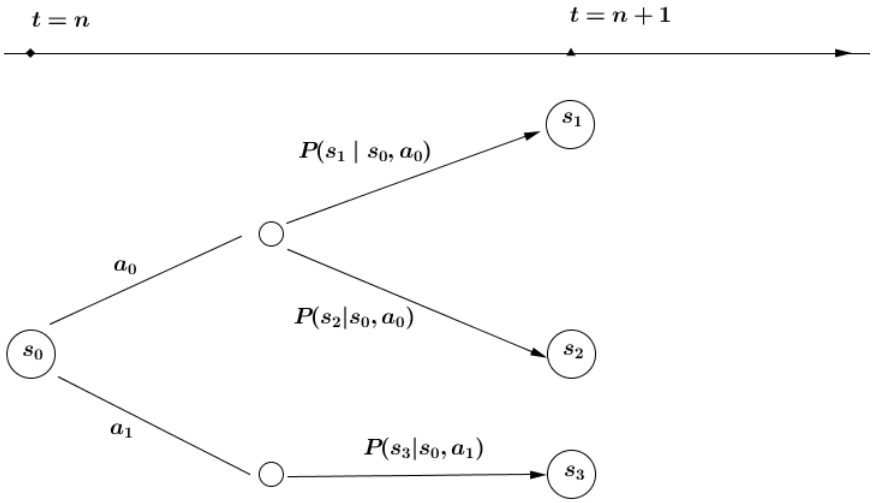

FIGURE 1.1 – Représentation d'une fonction de transition

Fonction de récompense

On appelle **fonction de récompense** l'application

$$R : \begin{array}{ccc} \mathcal{S} \times \mathcal{A} & \longrightarrow & (\mathbf{X}, o, \succsim) \\ (s, a) & \mapsto & R(s, a) \end{array} \quad \text{où}$$

$\forall s \in \mathcal{S}, \forall a \in \mathcal{A}$, $R(s, a)$ est le revenu que reçoit l'agent lorsqu'à un instant ou une étape t, il exécute l'action a à l'état s.

Remarque: 1.2.5. Cette récompense peut être instantanément perçue à l'instant t, ou accumulée de l'instant t à l'instant $t + 1$, l'important est qu'elle ne dépend que de l'état et de l'action choisie à l'instant courant. Dans le cas classique, les valeurs positives de $R(s, a)$ peuvent être considérées comme des gains et les valeurs négatives comme des pertes.

1.2.5 Cadre formel d'un PDMP

Afin de comprendre ce qu'est un PDMP, supposons que l'on ait un système évoluant dans le temps comme un automate. À chaque instant, le système est dans un état donné et il existe une certaine probabilité pour que le système évolue vers un autre état à l'instant suivant en effectuant une transition. Supposons que l'on doit contrôler ce système boite noire de la meilleure façon possible. L'objectif est de l'amener dans un état considéré comme bénéfique. Pour cela, on dispose d'un ensemble d'actions possibles sur le système dont les effets sont aléatoires. L'action entreprise peut avoir l'effet escompté ou tout autre effet et l'efficacité du contrôle est mesurée

relativement au gain ou à la pénalité reçue au long de l'expérience. Ainsi, un raisonnement à base de PDMP peut se ramener au discours suivant :

étant dans tel état et choisissant telle action, il y a tant de chance que je me retrouve dans tel nouvel état avec tel gain.

Formellement, un PDMP est la donnée du quintuplet $(I, \mathcal{S}, \mathcal{A}, T, R)$ (Puterman 1994) où :

- I est l'ensemble des instants (ou étapes) de décision ;
- $\mathcal{S}$ est l'ensemble des états de la nature ou de l'environnement de décision ;
- $\mathcal{A}$ est l'ensemble des actions qui contrôlent la dynamique de l'état du système ;
- $T : \mathcal{S} \times \mathcal{A} \to \Pi(\mathcal{S})$ est la fonction de transition où $\Pi(\mathcal{S})$ désigne l'ensemble des distributions de probabilité sur $\mathcal{S}$;
- $R : \mathcal{S} \times \mathcal{A} \to (\mathbf{X}, \text{o}, \succsim)$ la fonction de récompense où $\mathbf{X}$ est l'ensemble de valuation des coûts (gain ou perte) et o une loi de composition interne de X.

Nous supposons dans le cadre de notre travail que l'ensemble des instants de décision I, des états S et des actions $\mathcal{A}$ sont finis. Ainsi notre PDMP est réduit au quadruplet $(\mathcal{S}, \mathcal{A}, T, R)$.

Remarque: 1.2.6.

1- L'ensemble des coûts X est muni d'un opérateur interne $\circ$ et d'une relation de préférence $\succsim$. De plus, pour simplifier les notations, cet opérateur $\circ$ est supposé associatif, simplifiable à gauche (c'est -à- dire $\forall x, y \in X$, $x \circ y = x \circ z \Longrightarrow y = z$).

2- Pour la loi $\circ$, on définit pour tout couple $(x, z) \in X \times X$, l'ensemble noté :
$z \bullet x = \{\, y \in X \,/\, x \circ y = z\}$; cet ensemble peut être évidemment vide.

3- En général, ($\mathbf{X}$, o, $\succsim$) = ($\mathbb{R}$, +, $\geq$), on a alors $z \bullet x = \{z - x\}$.

Dans la suite du travail, on considère, sauf mention contraire (Ω, $\mathcal{T}(\Omega)$, P) un espace probabilisé, ($\mathcal{S}$, $\mathcal{T}(\mathcal{S})$) un espace probabilisable, $(X_t)_{t\in I}$ un processus stochastique sur $\mathcal{S}$ et $(\mathcal{S}, \mathcal{A}, T, R)$ un PDMP.

1.2.6 Règle de décision et politique de décision

Définition 1.13.

*1- Une **décision** est un choix d'une action à accomplir dans un état.*

*2- Une **règle de décision** notée δ_i à l'instant i, est une spécification de la décision à prendre pour chaque état possible du système de décisions. c'est encore une application de $\mathcal{S}$ vers $\mathcal{A}$.*

Remarque: 1.2.7. Soient $(\mathcal{S}, \mathcal{A}, T, R)$ un PDMP et $i \in I$.

- ➤ Lorsque $\delta_i : \mathcal{S} \to \mathcal{A}$, le modèle de PDMP est dit en stratégie pure.
- ➤ Lorsque $\delta_i : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, où $\mathcal{P}(\mathcal{A})$ désigne l'ensemble des parties de $\mathcal{A}$, le modèle de PDMP est dit en stratégie mixte.

Dans le cadre de notre travail, nous nous intéressons à un PDMP en stratégie pure et à horizon fini (le nombre n des séquences de décision est fini).

Définition 1.14. *Une politique de décision à un horizon n est une séquence de n règles de décision. C'est encore un n-uplets de règles de décision δ_1, ..., δ_n.*

Notation 1.2.3. Soient $\mathcal{S}$ l'ensemble des états et $\mathcal{A}$ l'ensemble des actions.

- ➢ Une politique à un horizon n sera notée $\phi_n = (\delta_1, ..., \delta_n)$.
- ➢ On pose Φ_n = $\{\phi_n = (\delta_1, ..., \delta_n)$ / $\delta_i \in \Delta \ \forall i$ = 1, ..., n $\}$,

  Φ_n désigne l'ensemble des politiques à horizon n.
- ➢ Pour une politique ϕ et une règle d'action δ, on note (δ, ϕ) la politique qui consiste à appliquer la règle de décision δ à l'étape 1 et à utiliser la politique ϕ ensuite. Par extension, on écrit (a, ϕ), la règle applicable dans un état, qui consiste à exécuter l'action a dans cet état puis la politique ϕ ensuite.
- ➢ Pour un ensemble de politique Φ, on note : (a, Φ) = $\{$ (a, ϕ) ; $\phi \in \Phi$ $\}$. Par convention (a, $\emptyset$) = $\{(a)\}$.

1.2.7 Historiques

Un historique permet de définir de façon successive dans un PDMP, pour un horizon fixé, tous les états en fonction des différentes actions exécutées. Les historique dans ce modèle débutant dans l'état s correspondent aux séquences suivantes :

$$(s, a_1, s_1, a_2, s_2, ..., a_n, s_n) \text{ où } \forall i \in \{1, 2, ..., n\}, (a_i, s_i) \in \mathcal{A} \times \mathcal{S}.$$

Notation 1.2.4. On note les ensembles d'historiques débutant de l'état s par :

$$\forall n > 0, \Gamma_n^s = \{(\ s, a_1, s_1, a_2, s_2, ..., a_n, s_n\ ) / \ \forall i = 1, ..., n, (a_i, s_i) \in \mathcal{A} \times \mathcal{S}\}.$$

Remarque: 1.2.8.

La valeur ou le coût abstrait d'un historique $\gamma \in \Gamma_n^{s_0}$ débutant à l'état s_0 avec $\gamma = (s_0, a_1, s_1, a_2, s_2, ..., a_n, s_n)$ vaut : $x = x_1 \circ ... \circ x_n \in X$ où $\forall i = 1, ..., n$, $x_i = R(s_{i-1}, a_i)$.

1.2.8 Loteries

Une **loterie** est une distribution de probabilité sur X. Par conséquent une politique ϕ_n induit, pour un horizon fixé et un état initial s donné, une loterie sur X également. Nous noterons $L_s^{\phi_n}$ la loterie sur l'ensemble des coûts X induite par la politique ϕ_n à l'état s. Elle associe à tout $x \in X$ la probabilité (Weng 2006) :

$$L_s^{\phi_n}(x) = \sum_{s' \in \mathcal{S}} T(s, \delta(s))(s') L_{s'}^{\phi_{n-1}}(x \bullet R(s, \delta(s))) \text{ où } \phi_n = (\delta, \phi_{n-1}) \text{ , } \delta \in \Delta \text{ et } \phi_{n-1} \in \Phi_{n-1}.$$

Dans le cadre classique $(\mathbb{R}, +, \geq)$, cette probabilité s'écrit ainsi :

$$L_s^{\phi_n}(x) = \sum_{s' \in \mathcal{S}} T(s, \delta(s))(s') L_{s'}^{\phi_{n-1}}(x - R(s, \delta(s))).$$

Remarque: 1.2.9. Soient $s, s' \in \mathcal{S}$ et $x \in X$.

1- $L_{s'}^{\phi_{n-1}}(x - R(s, \delta(s)))$ est la probabilité de réaliser le coût $x - R(s, \delta(s))$ induite par la politique ϕ_{n-1} à l'état s pour passer à l'état s'.

2- $T(s, \delta(s))(s') = P(s'|s, \delta(s))$ est la probabilité de passer à l'état s' lorsqu'on a appliqué la règle de décision δ sur l'état s.

3- La probabilité que la politique ϕ_n à l'horizon n génère le coût x est la moyenne pondérée des probabilités aux états s' que la sous - politique ϕ_{n-1} en ces états génère le coût x réduit du coût $R(s, \delta(s))$ imputé à l'étape n. Il est donc possible d'étudier ce modèle selon les propriétés de cet ensemble X.

1.2.9 Politiques classiques

On distingue plusieurs politiques selon le modèle de PDM choisi :

- la **politique déterministe** si le modèle de PDM est déterministe.
- la **politique stochastique** si le modèle de PDM est stochastique.
- la **politique possibiliste** si le modèle de PDM est possibiliste.
- la **politique statique** ou **stationnaire** si la règle de décision δ_i est identique à tout instant de décision i.

1.3 Problématique générale d'un PDMP

1.3.1 Problématique

Le comportement d'un agent est guidé par une politique. Résoudre un PDMP consiste à déterminer une politique, indiquant pour chaque état du système, la meilleure action à exécuter, en maximisant (ou minimisant) au bout d'un horizon n une certaine fonction de récompense qui dépend des décisions prises et l'évolution du système c'est-à-dire identifier les actions à prendre dans chaque état pour maximiser l'espérance de la somme des récompenses relativement à un critère spécifique défini par l'agent décideur ; cette solution est appelée **politique optimale**. Notre travail consistera à trouver les conditions d'existence d'une telle politique.

1.3.2 Hypothèses

Dans un modèle de PDMP, plusieurs hypothèses sont préalablement prises en compte :

- ➤ l'évolution future ne dépend que de l'état actuel et de la décision (action) prise à cet instant (propriété de Markov) ;
- ➤ l'ensemble I des instants de décision est discret et généralement à pas constant c'est-à-dire la la durée qui sépare deux instants de prise de prise de décision est constante ;
- ➤ la structure $(X, o, \succsim)$ est choisie de telle sorte que $\succsim$ représente les préférences sur les séquences d'actions.

1.3.3 Exemple

Dans une usine d'impression des journaux, un agent souhaite maximiser le revenu provenant de la production de la machine ; après quatre périodes de fonctionnement, l'état de la machine dépend de son entretien. Les données du problème sont les suivantes :

- ➢ les états de la machine sont : état neuf, bon état, mauvais état, état en panne ;
- ➢ les rendements respectifs (par période) sont 30, 15, 5 et 0 (en FCFA) ;
- ➢ les actions possibles sont : entretenir, ne rien faire, rénover ;
- ➢ le revenu par objet produit est 100 FCFA.

Le **tableau 1.1** fournit les informations sur les gains associés aux différentes combinaisons d'états de décision ainsi que le rendement par période.

Dans ce tableau, Gain = Revenu par objet produit $\times$ Rendement par période.

TABLE 1.1 – Tableau des gains relatifs à chaque état.

États(s)	**Rendements par période (en FCFA)**	**Revenus par objet (en FCFA)**	**Gains (en FCFA)**
s_0	30	100	3000
s_1	15	100	1500
s_2	**5**	**100**	**500**
s_3	0	100	0

Le PDMP considéré est la donnée du quadruplet $(\mathcal{S}, \mathcal{A}, T, R)$ avec :

- l'ensemble des états $\mathcal{S} = \{s_0; s_1; s_2; s_3\}$ où :
  - s_0 signifie " état neuf" ;
  - s_1 signifie " bon état" ;
  - s_2 signifie " mauvais état" ;
  - s_3 signifie " état en panne" ;
- l'ensemble des actions $\mathcal{A} = \{a_1; a_2; a_3\}$ où :
  - a_1 signifie " entretenir " ;
  - a_2 signifie " ne rien faire" ;
  - a_3 signifie " rénover " ;
- $T : \mathcal{S} \times \mathcal{A} \to \Pi(\mathcal{S})$ la fonction de transition ; elle définit toutes les probabilités de transition entre les différents états en fonction des actions ;
- la fonction de récompense $R : \mathcal{S} \times \mathcal{A} \to (\mathbb{R}, +, \geq)$ qui caractérise les revenus (gain ou perte).

Nous allons achever la modélisation de ce problème et le résoudre au chapitre 3.

Nous avons énuméré quelques outils mathématiques nécessaires et donné la définition générale d'un PDMP. Dans le chapitre suivant, on verra à l'aide de deux approches, les conditions d'existence d'une politiques optimales dans un PDMP à horizon fini pour des préférences complètes.

⋆ ⋆ **Chapitre Deux** ⋆ ⋆

MÉTHODES DE RÉSOLUTION : APPROCHE PAR LA FONCTION DE VALEUR ET PAR LES LOTERIES

Dans certaines situations, le modèle classique des PDMP ne convient pas car les préférences à prendre en compte ne sont pas réprésentables par un critére coût scalaire additif. Il semble donc intéressant d'étudier l'extension à des PDMP exploitant les préférences non classiques. On le fera à l'aide de l'approche par les loteries développée par Weng (2006).

2.1 Approche par la fonction de valeur

C'est une méthode appropriée pour étudier le modèle classique des PDMP ; ainsi dans cette partie, on considéra $(\mathbf{X}, o, \succsim) = (\mathbb{R}, +, \geq)$. Ce modèle repose implicitement sur une structure de préférence induite par l'existence de coûts scalaires additifs et l'utilisation d'un certain critère de performance (d'évaluation) des politiques.

2.1.1 Critères de performance

Un critère de performance pour un PDMP sert à évaluer une politique sur la base d'une mesure du cumul espéré des récompenses instantanées le long d'une trajectoire.
On définit quelques critères classiques :

- le critère fini : $E[R_0 + R_1 + R_2 + ... + R_{n-1}|s_0]$,
- le critère γ-pondéré : $E[R_0 + \gamma R_1 + \gamma^2 R_2 + ... + \gamma^t R_t + ...|s_0]$;
- le critère total : $E[R_0 + R_1 + R_2 + ... + R_t + ...|s_0]$;
- le critère moyen : $\lim\limits_{n\to+\infty} \frac{1}{n} E[R_0 + R_1 + R_2 + ... + R_{n-1}|s_0]$.

$E[.|s_0]$ est l'espérance mathématique lorsque que le système débute à l'état s_0. Elle permet d'évaluer l'espérance mathématique de la somme des récompenses instantanées.

Remarque: 2.1.1.

1- $\gamma \in [0, 1[$ est le coefficient d'actualisation (ou facteur d'atténuation) et est interprété comme la probabilité de survie de l'agent décideur ($1 - \gamma$ est la probabilité d'arrêt du processus).

2- $R_t = R(s_t, a_t)$ est la récompense obtenue par l'agent à l'instant $t+1$ lorsque le système se trouve à l'état s_t après l'exécution de l'action a_t.

3- s_0 est l'état initial du système.

2.1.2 Fonctions de valeur

Une **Fonction de valeur** est une fonction d'utilité qui sert à évaluer le degré d'importance de la politique choisie par l'agent décideur selon le critère de performance considéré.

$\forall \phi \in \Phi$, on définit les fonctions de valeur classiques $V^\phi : \mathcal{S} \to (\mathbb{R}, +, \geq)$ par :

- $V^\phi(s) = \sum_{t=0}^{n-1} R_t | s_0 = s$ en modèle déterministe ;
- $V^\phi(s) = E^\phi \left[\sum_{t=0}^{n-1} R_t | s_0 = s \right]$ pour le critère fini ;
- $V^\phi(s) = E^\phi \left[\sum_{t=0}^{\infty} \gamma^t R_t | s_0 = s \right]$ pour le critère γ-pondéré ;
- $V^\phi(s) = E^\phi \left[\sum_{t=0}^{\infty} R_t | s_0 = s \right]$ pour le critère total ;
- $V^\phi(s) = \lim_{n \to +\infty} E^\phi \left[\frac{1}{n} \sum_{t=0}^{n-1} R_t | s_0 = s \right]$ pour le critère moyen.

Remarque: 2.1.2.

1- $E^\phi[.]$ désigne l'espérance mathématique relative à la politique ϕ.

2- $\forall s \in \mathcal{S}, \forall \phi \in \Phi$, $V^\phi(s)$ est l'espérance du cumul des récompenses que l'on peut obtenir à partir de l'état s en suivant la politique ϕ.

3- Les quatre derniers cas sont en modèles stochastiques.

4- Il n'existe pas encore des travaux sur les fonctions de valeur non classiques.

2.1.3 Relation de préférence sur les politiques

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

$$\forall V_1, V_2 \in \mathcal{V},\ \ (\ \ V_1 \succsim_{\mathcal{V}} V_2 \Longleftrightarrow \forall s \in \mathcal{S},\ \ V_1(s) \geq V_2(s)\ \ ).$$

$\succsim_{\mathcal{V}}$ induit un ordre partiel sur l'ensemble des politiques Φ pour tout $V \in \mathcal{V}$, et défini par :

$$\forall \phi_1, \phi_2 \in \Phi,\ \ (\ \phi_1 \succsim_{\Phi} \phi_2 \Longleftrightarrow \forall s \in \mathcal{S},\ \ V^{\phi_1}(s) \geq V^{\phi_2}(s)\ )$$

où $\forall s \in \mathcal{S}, \forall i \in \{1, 2\}$, $V^{\phi_i}(s)$ est la fonction de valeur relative à la politique appliquée ϕ_i lorsque l'état initial était s.

2.1.4 Équation de BELLMAN pour le critère fini

Une propriété fondamentale de la fonction de valeur d'une politique ϕ est le fait qu'elle vérifie une équation récursive, l'équation de Bellman pour le critère fini (Bellman 1957) :

$$\forall s \in \mathcal{S}, \forall a \in \mathcal{A}, V^{\phi}(s) = R(s, a) + \sum_{s' \in \mathcal{S}} P(s'|s, a)V^{\phi}(s') \qquad (1).$$

Dans le cadre des PDMP, l'objectif est de déterminer une politique optimale.
On note V^* la fonction de valeur optimale, qui à chaque état initial s associe $V^*(s) = \max_{\phi \in \Phi} V^{\phi}(s)$.
Il peut exister plusieurs politiques optimales, qui partagent cette fonction de valeur optimale.
Le théorème suivant permet de déterminer les fonctions de valeurs optimales et les politiques optimales .

Théorème 2.1. ***(Bellman 1957)***

Soit $n \in \mathbb{N}^*$. Les n-uplets fonctions de valeur optimales $(V_n^*, V_{n-1}^*, ...V_1^*) = V^*$ sont solutions uniques du système d'équations :

$$\forall s \in \mathcal{S},\ V_{t+1}^*(s) = \max_{a \in \mathcal{A}} \left\{ R_{n-1-t}(s, a) + \sum_{s' \in \mathcal{S}} P_{n-1-t}(s'|s, a)V_t^*(s') \right\}$$

avec $t \in \{0, 1, 2, ..., n-1\}$ et $V_0 = 0$.

Les politiques optimales pour le critère fini $\phi^* = (\delta_0^*, \delta_1^*, ..., \delta_{n-1}^*)$ sont déterminées par :

$$\forall s \in \mathcal{S},\ \delta_t^*(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ R_t(s, a) + \sum_{s' \in \mathcal{S}} P_t(s'|a, s)V_{n-1-t}^*(s') \right\}$$

avec $t \in \{0, 1, 2, ..., n-1\}$.

Remarque: 2.1.3. argmax est la fonction réciproque de la fonction **maximun**.

Preuve.

Supposons que l'agent se trouve dans l'état s lors de la dernière étape de décision, confronté au choix de la meilleure décision à prendre. Il est clair que c'est celle qui maximise la récompense instantanée à venir, qui viendra s'ajouter aux récompenses déjà perçues. On a ainsi :

$$\delta^*_{n-1}(s) \in \underset{a\in\mathcal{A}}{\text{argmax}}\{R_{n-1}(s,a)\} \text{ et } V^*_1(s) = \max_{a\in\mathcal{A}} R_{n-1}(s,a)$$

où δ^*_{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 que l'agent se trouve 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'amener de manière aléatoire vers un nouvel état s' à l'étape $n-1$. Ainsi, il sait qu'en suivant la politique optimale avec δ^*_{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 récompense :

$$R_{n-2}(s,a) + \sum_{s'\in\mathcal{S}} P_{n-2}(s'|a,s)V^*_1(s').$$

Ainsi, le problème de l'agent à l'étape $n-2$ se ramène simplement à rechercher l'action qui maximise cette somme, soit :

$$\delta^*_{n-2}(s) \in \underset{a\in\mathcal{A}}{\text{argmax}} \left\{ R_{n-2}(s,a) + \sum_{s'\in\mathcal{S}} P_{n-2}(s'|a,s)V^*_1(s') \right\}$$

et

$$V^*_2(s) = \max_{a\in\mathcal{A}} \left\{ R_{n-2}(s,a) + \sum_{s'\in\mathcal{S}} P_{n-2}(s'|a,s)V^*_1(s') \right\}.$$

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

$$\delta^*_0(s) \in \underset{a\in\mathcal{A}}{\text{argmax}} \left\{ R_0(s,a) + \sum_{s'\in\mathcal{S}} P_0(s'|a,s)V^*_{n-1}(s') \right\}$$

et

$$V^*_n(s) = \max_{a\in\mathcal{A}} \left\{ R_0(s,a) + \sum_{s'\in\mathcal{S}} P_0(s'|a,s)V^*_{n-1}(s') \right\}.$$

On a ainsi construit par itération n-uplets fonctions de valeur optimales :

$$(V_n^*, V_{n-1}^*, ...V_1^*) = V^*$$

associées aux n-uplets politiques optimales :

$$(\delta_0^*, \delta_1^*, ..., \delta_{n-1}^*) = \phi^*.$$

■

Remarque: 2.1.4. Nous constatons les résultats suivants :

- une fonction de valeur optimale, solution de l'équation de Bellman, est un unique vecteur dont les composantes sont des fonctions de valeur optimales, relatives à chaque instant (ou étape) ;
- étant donné que la fonction argmax est définie sur $\mathcal{A}$ et pouvant avoir plusieurs éléments qui déterminent une règle de décision optimale, on peut conclure que les politiques optimales ne sont pas toujours uniques ;
- la complétude de la relation de préférence sur X (cas classique) et l'additivité de la loi $\circ$ sont les conditions d'existence des fonctions de valeurs optimales, par conséquent des politiques optimales.

2.1.5 Résolution de l'équation d'optimalité de BELLMAN à horizon fini

L'algorithme suivant de programmation dynamique à horizon fini permet de déterminer rapidement les fonctions de valeur optimales, les règles de décision optimales par conséquent les politiques optimales.

1 : $t \leftarrow n-1$

2 : $V_0 \leftarrow 0$

3 : **repeat**

4 : $t \leftarrow t-1$

5 : **for all** $s \in \mathcal{S}$ **do**

6 :
$$\begin{cases} V_{t+1}^*(s) \leftarrow V_{t+1}^*(s) + \max\limits_{a\in\mathcal{A}} \left\{ R_{n-1-t}(s,a) + \sum\limits_{s'\in\mathcal{S}} P_{n-1-t}(s'|a,s) V_t^*(s') \right\} \\ \delta_{n-1-t}^*(s) \in \operatorname{argmax}\limits_{a\in\mathcal{A}} \left\{ R_{n-1-t}(s,a) + \sum\limits_{s'\in\mathcal{S}} P_{n-1-t}(s'|a,s) V_t^*(s') \right\} \end{cases}$$

7 : **end for**

8 : **retourner** V^*, δ^*

9 : **until** $t = 0$

Remarque: 2.1.5. Pour chaque fonction de valeur optimale obtenue à l'étape précédente, les opérations suivantes sont effectuées :

- dans chaque état, l'algorithme calcule la fonction de valeur optimale à l'horizon t (ligne 6) ;
- puis construit la ou les meilleures règles de décision pour l'horizon t (ligne 6) en sélectionnant une action parmi la ou les meilleurs actions calculées dans chaque état ;
- ces opérations sont effectuées pour chaque fonction de valeur optimale calculée à l'étape précédente ;
- l'algorithme calcule V^*, δ^* à chaque étape en reposant sur le **théorème 2.1** et retourne le résultat final (ligne 8) ;
- pour obtenir une seule politique optimale ϕ_n à horizon n, il suffit de considérer les n meilleures règles de décision.

2.2 Approche par les loteries

C'est une approche de résolution de deux types de PMDP (PMDP classique et PMDP non classique). Elle est plus avantageuse que celle précédente car elle peut permettre de résoudre les PMDP où l'ensemble des valuations X est qualitatif. Les travaux les plus récents de cette approche sont ceux de Paul WENG (2006).

2.2.1 Définitions et notations

➤ Pour un ensemble Y et une relation de préférence $\succsim$ sur cet ensemble, on définit l'ensemble des éléments maximaux par $M(Y, \succsim) = \{y \in Y | \forall z \in Y, \neg(z \succ y)\}$ = $M(Y)$.

➤ Si $\succsim$ est complète, $M(Y)$ est noté max(Y) et devient simplement l'ensemble des éléments optimaux définis par max$(Y) = \{y^* \in Y / \forall y \in Y, y^* \succsim y\}$.

➤ Si l'on note $\succsim_\Phi$ la relation de préférence sur les politiques alors l'ensemble des politiques optimales pour un horizon fixé n donné est noté $\Phi_n^* = M(\Phi_n, \succsim_\Phi)$.

De plus, on définit $\forall n > 0$, Φ_n^+ par :

$$\begin{cases} \Phi_1^+ = \Phi_1^* \\ \forall n \geq 1, \Phi_{n+1}^+ = \bigcup\limits_{\phi_n \in \Phi_n^+} M\left(\{(\delta, \phi_n) | \delta \in \Delta\}, \succsim_\Phi\right). \end{cases}$$

On remarquera que l'algorithme de recherche arrière construit ces ensembles. Pour une politique calculée à l'étape précédente, on calcule la ou les meilleures (au sens de $\succsim_\Phi$) règles de décision à lui ajouter à la première étape.

➤ Enfin on définit $\forall n > 0$, Φ_n^{+M} par :

$$\begin{cases} \Phi_1^{+M} = \Phi_1^* \\ \forall n \geq 1, \Phi_{n+1}^{+M} = M\left(\bigcup\limits_{\phi_n \in \Phi_n^{+M}} \{(\delta, \phi_n) | \delta \in \Delta\}, \succsim_\Phi\right) \end{cases} \quad (3).$$

➤ Notons une réécriture intéressante de Φ_n^{+M} qui nous servira dans la définition des algorithmes :

$$\forall n \geq 1, \Phi_{n+1}^{+M} = M\left(\left(\bigcup\limits_{\phi_n \in \Phi_n^{+M}} (\{(\delta, \phi_n) | \delta \in \Delta\}, \succsim_\Phi\right), \succsim_\Phi\right) \quad (4).$$

Remarque: 2.2.1. Ces ensembles sont définis de manière récursive. On constate que :

- pour une étape donnée, on considère dans la relation (4), les meilleures politiques parmi l'ensemble des politiques déterminées précédemment auxquelles on adjoint une règle de décision ;
- la différence entre la relation (3) et la relation (4) est la portée de l'opérateur de maximisation,
- dans la relation (4), l'opération de maximisation est définie une fois sur tout un ensemble contrairement à la relation (3) où la maximisation est définie sur plusieurs petits ensembles ; on peut donc soupçonner un coût de calcul plus important pour cette dernière relation ;
- de plus, pour déterminer un élément de Φ_{n+1}^{+M}, il est nécessaire de calculer entièrement Φ_n^{+M} ; par contre, pour obtenir un élément de Φ_{n+1}^+, il suffit de déterminer un seul élément de Φ_n^+.

Définition 2.1. *La relation de préférence $\succsim$ sur l'ensemble $(X, \circ)$ est dite **pré-additive** si et seulement si*

$$\forall \gamma, \gamma' \in X, \forall x \in X, (\gamma \succsim \gamma' \iff x \circ \gamma \succsim x \circ \gamma').$$

Définition 2.2. *Soit $L(X)$ l'ensemble des loteries probabilistes sur X.*
*Une relation de préférence $\succsim_L$ sur les loteries définies sur $(X, \circ)$ est **invariante par translation** si et seulement si* $\forall L_1, L_2 \in L(X), \forall c \in X, \; (\; L_1 \succsim_L L_2 \Longrightarrow L_1^{\rightarrow c} \succsim_L L_2^{\rightarrow c} \;)$ *où*

$$\forall i \in \{1, 2\}, \forall x, c \in X, \; L_i^{\rightarrow c}(c \circ x) = L_i(x) \, .$$

Interprétation 2.1. Cette définition permet d'affirmer qu'une préférence entre deux loteries est conservée même si tous les éléments sur lesquels sont définies les loteries sont translatés d'une même quantité. Cette définition peut être considérée comme la version probabiliste de la préaddivité.

Définition 2.3. *Une relation de préférence $\succsim_L$ sur les loteries définies sur $(X, \circ)$ vérifie la **propriété d'indépendance** si et seulement si* $\forall \; L_1, \; L_2, \; L_3 \in L(X), \forall \lambda \in]0, 1[,$

$$(\; L_1 \succsim_L L_2 \Longrightarrow \lambda L_1 + (1 - \lambda) L_3 \succsim_L \lambda L_2 + (1 - \lambda) L_3 \;) \, .$$

Interprétation 2.2. Cette définition correspond en fait à une version affaiblie de la propriété d'indépendance de Von Neumann et al.1944 formulée par Fishburn 1970. Elle dit en substance que les préférences sur deux loteries ne peuvent s'inverser si on combine ces deux loteries à une troisième loterie, c'est-à-dire, de manière intuitive que l'ajout de conséquences identiques (avec les mêmes probabilités) à deux loteries ne peut inverser le sens de préférence.

Définition 2.4. *Une relation de préférence $\succsim_\Phi$ sur les politiques sera dite **stable** si et seulement si* $\forall \phi, \phi' \in \Phi, \forall \delta \in \Delta, \; (\; \phi \succsim_\Phi \phi' \Longrightarrow (\delta, \phi) \succsim_\Phi (\delta, \phi') \;)$.

Interprétation 2.3. Intuitivement, cette définition signifie simplement que si une politique ϕ est préférée à une politique ϕ' alors le fait de retarder l'application de ces deux politiques par l'utilisation d'une même règle de décision δ conserve le sens de la préférence. Cette propriété est cruciale pour permettre le calcul itératif de politiques préférées.

2.2.2 Relation de préférence sur les loteries

Dans le modèle des PDMP, il est possible de distinguer trois niveaux de relation de préférence (Weng 2006). Une première relation $\succsim$ est définie sur les historiques ou de manière équivalente sur l'ensemble des coûts X. Comme une politique pour un horizon fixé et un état initial donné définit une loterie sur l'ensemble X, comparer deux politiques à un horizon donné et dans un certain état initial équivaut à comparer leurs loteries respectives. C'est pourquoi à

partir de la première relation de préférence, il est nécessaire de définir une relation de préférence $\succsim_L$ sur les loteries. Enfin cette dernière induit une troisième relation de préférence $\succsim_\Phi$ sur les politiques permettant de définir la notion d'optimalité ou de maximalité sur l'ensemble des politiques. La relation $\succsim_\Phi$ est définie par :

$$\forall(\phi, \phi') \in \Phi \times \Phi, \phi \succsim_\Phi \phi' \Leftrightarrow \forall s \in \mathcal{S}, L_s^\phi \succsim_L L_s^{\phi'}. \tag{5}$$

Rappelons que dans cette relation, L_s^ϕ est la loterie sur l'ensemble des coûts X induite par la politique ϕ à l'état s.

Voici deux lemmes qui serviront ultérieurement.

Lemme 2.1. *(Weng 2006)*

*Si $\succsim_L$ est **transitive** alors $\succsim_\Phi$ est **transitive**. De plus, si $\succsim_\Phi$ est **stable** alors la relation $\sim_\Phi$ est **stable** également.*

Le lemme suivant indique que sous les conditions d'indépendance et de transitivité de la relation de préférence sur les loteries, la combinaison d'un nombre quelconque de loteries conserve le sens de préférence.

Lemme 2.2. *Si $\succsim_L$ est **indépendante** et **transitive** alors si $(L_i)_{i=1,...,n}$ et $(L'_i)_{i=1..n}$ représentent deux familles finies de loteries telles que $\forall i = 1, ..., n$, $L_i \succsim L'_i$, on a :*

$\forall i = 1, ..., n, \lambda_i \in [0, 1]$, *tels que* $\sum_{i=1}^{n} \lambda_i = 1$, $\sum_{i=1}^{n} \lambda_i L_i \succsim_L \sum_{i=1}^{n} \lambda_i L'_i$.

Preuve.

La preuve se fait par récurrence sur n.

Pour $n = 2$, prenons deux couples de loteries (L_1, L_2) et (L'_1, L'_2) telles que $L_1 \succsim_L L'_1$ et $L_2 \succsim_L L'_2$. En appliquant la propriété d'indépendance sur la première relation et L_2, on a :

$$\forall\lambda \in [0, 1], \lambda L_1 + (1 - \lambda)L_2 \succsim_L \lambda L'_1 + (1 - \lambda)L_2.$$

Puis en appliquant la propriété d'indépendance sur la seconde relation et L'_1, on a :

$$\forall\lambda \in [0, 1], \lambda L'_1 + (1 - \lambda)L_2 \succsim_L \lambda L'_1 + (1 - \lambda)L'_2$$

Enfin par transitivité, on obtient bien :

$$\forall\lambda \in [0, 1], \lambda L_1 + (1 - \lambda)L_2 \succsim_L \lambda L'_1 + (1 - \lambda)L'_2.$$

Supposons que la relation est vraie avec n loteries.

Considérons deux familles de loteries $(L_i)_{i=1,...,n+1}$ et $(L'_i)_{i=1,...,n+1}$ telles que $\forall i = 1, ..., n+1$, $L_i \succsim_L L'_i$. Soit une séquence $(\lambda_i)_{i=1,...,n+1} \in [0,1]$ telle que $\sum_{i=1,...,n+1} \lambda_i = 1$.

Cas 1 : $\lambda_{n+1} = 1$: La propriété est démontrée.

Cas 2 : $\lambda_{n+1} \neq 1$: Posons $L = \sum_{i=1,...,n} \frac{\lambda_i}{1-\lambda_{n+1}} L_i$ et $L' = \sum_{i=1,...,n} \frac{\lambda_i}{1-\lambda_{n+1}} L'_i$.

Ce sont deux loteries. Et d'après l'hypothèse de récurrence, $L \succsim_L L'$.

En appliquant la propriété démontrée pour $n = 2$, en prenant $\lambda = \lambda_{n+1}$, on obtient :

$$\lambda_{n+1} L_{n+1} + (1 - \lambda_{n+1}) L \succsim_L \lambda L'_{n+1} + (1 - \lambda) L'.$$

En développant L et L', on obtient bien :

$$\sum_{i=1,...,n+1} \lambda_i L_i \succsim_L \sum_{i=1,...,n+1} \lambda_i L'_i.$$

■

La proposition suivante donne des conditions suffisantes pour garantir la stabilité de la relation de préférence sur les politiques.

Proposition 2.1. *Si $\succsim_L$ (respectivement $\succ_L$) est **invariante par translation, transitive** et **indépendante** alors $\succsim_\Phi$ (respectivement $\succ_\Phi$) est **stable**.*

Preuve.

Soient deux politiques ϕ, ϕ' telles que $\phi \succsim_\Phi \phi$. Soit une règle de décision δ.

Par définition, on a :

$$\phi \succsim_\Phi \phi' \iff \forall s \in \mathcal{S},\ L^{\phi}_{s'} \succsim_L L^{\phi'}_{s'} \text{ (D'après la relation (1)).}$$

Considérons un état initial s quelconque. Par définition, la loterie induite par (δ, ϕ) en s vaut :

$$\forall x \in X,\ L_s^{(\delta,\phi)}(x) = \sum_{s' \in S} T(s, \delta(s))(s') L^{\phi}_{s'}(x \bullet R(s, \delta(s))).$$

De même, pour (δ, ϕ'), on obtient :

$$\forall x \in X,\ L_s^{(\delta,\phi)}(x) = \sum_{s' \in S} T(s, \delta(s))(s') L^{\phi'}_{s}(x \bullet R(s, \delta(s))).$$

En posant $\forall s' \in \mathcal{S}, \forall x \in X,\ L_{s'}(x) = L^{\phi}_{s'}(x \bullet R(s, \delta(s)))$ et $L'_{s'}(x) = L^{\phi'}_{s'}(x \bullet R(s, \delta(s)))$,

on peut réécrire les loteries :

$$L_s^{(\delta,\phi)} = \sum_{s' \in \mathcal{S}} T(s, \delta(s))(s') L_{s'} \text{ et } L_s^{(\delta,\phi')} = \sum_{s' \in \mathcal{S}} T(s, \delta(s))(s') L'_{s'}.$$

D'après l'hypothèse de simplifiabilité à gauche, les ensembles $x \bullet R(s, \delta(s))$ sont les singletons ou sont vides. Pour les x tels que $x \bullet R(s, \delta(s))$ est vide,

$$L_{s'}(x) = L_{s'}^{\phi}(\emptyset) = 0.$$

Pour les x tels que $(x \bullet R(s, \delta(s)))$ est un singleton,

$$\text{il existe un } y \text{ tel que } x \bullet R(s, \delta(s)) = y, \text{ autrement dit, } x = R(s, \delta(s)) \circ y.$$

Alors, par définition :

$$L_{s'}(x) = L_{s'}(R(s, \delta(s)) \circ y) = L_{s'}^{\phi}(y).$$

Plus simplement, on peut écrire :

$$\forall x \in X,\ L_{s'}^{\phi}(x) = L_{s'}(R(s, \delta(s)) \circ x) \text{ et } L_{s'}^{\phi'}(x) = L'_{s'}(R(s, \delta(s)) \circ x).$$

Donc en vertu de l'hypothèse d'invariance par translation :

$$\forall s' \in \mathcal{S},\ L_{s'} \succsim_L L'_{s'}\ .$$

D'après le **lemme 2.2**, On a :

$$\sum_{s' \in \mathcal{S}} T(s, \delta(s))(s') L_{s'} \succsim_L \sum_{s' \in S} T(s, \delta(s))(s') L'_{s'}.$$

On a bien :

$$L_s^{(\delta,\phi)} \succsim_L L_s^{(\delta,\phi')}.$$

Par conséquent $\succsim_\Phi$ est stable. De manière similaire, on démontre que si $\succ_L$ est transitive, invariante par translation et indépendante alors la relation $\succ_\Phi$ est stable.

■

2.2.3 Cadre des préférences complètes

Le cadre des préférences complètes se caractérise par la donnée d'une relation d'ordre total sur X, d'une relation de préférence complète, transitive sur les loteries et d'une relation de préférence stable sur les politiques. Grâce à cette hypothèse de complétude, une politique maximale est une politique optimale. Sous ces conditions, nous démontrons qu'il existe au moins une politique optimale.

Proposition 2.2. *Si $\succsim_L$ est complète, transitive et $\succsim_\Phi$ est stable alors pour tout $n > 0$, les ensembles Φ_n^+, Φ_n^* ne sont pas vides et $\Phi_n^+ \subseteq \Phi_n^*$.*

Preuve.

D'après le **lemme 2.1**, $\succsim_\Phi$ est transitive.

La démonstration se fait par récurrence sur n.

Pour $n = 1$, pour chaque état, on peut sélectionner une meilleure action car $\succsim_L$ est complète.

On peut définir par conséquent une meilleure règle de décision.

On a alors $\Phi_1^+ = \Phi_1^*$, qui sont non vides.

Soit $n \geq 1$. On suppose que les ensembles Φ_n^+, Φ_n^* sont non vides et que $\Phi_n^+ \subseteq \Phi_n^*$.

Démontrons-le pour $n + 1$.

Par construction, Φ_{n+1}^+ est non vide également.

Soit $\phi_{n+1}^+ = (\delta^+, \Phi_n^*) \in \Phi_{n+1}^+$ avec $\delta^+ \in \Delta$ et $\phi_n^* \in \Phi_n^+ \subseteq \Phi_n^*$.

Montrons que $\Phi_{n+1}^+ \in \Phi_{n+1}^*$.

Par hypothèse $\forall \phi_n \in \Phi_n$, $\phi_n^* \succsim_\Phi \phi_n$.

Par stabilité $\forall \phi_n \in \Phi_n$, $\forall \delta \in \Delta$, $(\delta, \phi_n^*) \succ_\Phi (\delta, \phi_n)$.

Or, comme $\succsim_L$ est complète par définition de ϕ_{n+1}^+, $\forall \delta \in \Delta$, $(\delta^+, \phi_n^*) \succ_\Phi (\delta, \phi_n)$.

Donc par transitivité, $\forall \phi_n \in \Phi$, $\forall \delta \in \Delta$, $(\delta^+, \phi_n^*) \succ_\Phi (\delta, \phi_n)$.

Par conséquent, $\phi_{n+1}^+ \in \Phi_{n+1}^*$ et cet ensemble est non vide. ■

Le corollaire suivant montre que dans le cadre des préférences complètes, quand la relation de préférence stricte sur les politiques est stable également, il est possible de construire itérativement toutes les politiques optimales.

Corollaire 2.1. *Weng(2006)*

*Si $\succsim_L$ est **complète, transitive** et les relations $\succsim_\Phi$ et $\succ_\Phi$ sont **stable** alors*

$$\forall n > 0, \textit{ l'ensemble } \Phi_n^* \textit{ n'est pas vide et } \Phi_n^+ = \Phi_n^*.$$

Proposition 2.3. *Si $\succsim_L$ est **complète, transitive** et $\succsim_\Phi$ est **stable** alors l'égalité suivante est vérifiée :*

$$\forall n > 0,\ \Phi_n^+ = \Phi_n^{+M}.$$

Preuve.

La démonstration se fait par récurrence sur n.

L'égalité est vraie pour $n = 1$.

Soit $n \geq 1$. Supposons l'égalité vérifiée.

Démontrons - la pour $n+1$.

Par définition,

$$\begin{aligned}\Phi_{n+1}^{+} &= M\left(\bigcup_{\phi_n \in \Phi_n^{+M}} \{(\delta, \phi_n) | \delta \in \Delta\}, \succsim_{\Phi}\right) \\ &= M\left(\bigcup_{\phi_n \in \Phi_n^{+M}} M\left(\{(\delta, \phi_n) | \delta \in \Delta\}, \succsim_{\Phi}\right), \succsim_{\Phi}\right) \\ &= M\left(\bigcup_{\phi_n \in \Phi_n^{+}} M\left(\{(\delta, \phi_n) | \delta \in \Delta\}, \succsim_{\Phi}\right), \succsim_{\Phi}\right) \\ &= M(\Phi_{n+1}^{+}, \succsim_{\Phi}).\end{aligned}$$

D'après la **proposition 2.2**,

$$\Phi_n^{+} \subset \Phi_{n+1}^{*}.$$

Par conséquent

$$M(\Phi_{n+1}^{+}, \succsim_{\Phi}) = \Phi_{n+1}^{+}.$$

Finalement,

$$\Phi_{n+1}^{+M} = \Phi_{n+1}^{+}.$$

■

2.2.4 Algorithme de recherche arrière généralisé (Weng 2006)

L'algorithme de recherche arrière généralisé s'écrit :

1 : $t \leftarrow n$

2 : $\Phi_n^{*} \leftarrow \{()\}$

3 : **repeat**

4 : $t \leftarrow t-1$

5 : **for all** $\phi \in \Phi_{t+1}^{*}$ **do**

6 : **for all** $s \in \mathcal{S}$ **do**

7 : $\Phi_t^{*}(s) \leftarrow \Phi_t^{*}(s) \bigcup \max\{(a, \phi) : a \in A\}$

8 : **end for**

9 : ajoute dans Φ_t^{*} des politiques obtenues à partir de $\Phi_t^{*}(s)$

10 : **end for**

11 : **until** $t = 0$

Remarque: 2.2.2. Pour chaque politique obtenue à l'étape précédente, les opérations suivantes sont effectuées :

- dans chaque état, l'algorithme calcule les meilleures actions à effectuer à l'horizon t (ligne 7) ;
- puis construit la ou les meilleures règles de décision pour l'horizon t (ligne 9) en sélectionnant une action parmi la ou les meilleurs actions calculées dans chaque état ;
- ces opérations sont effectuées pour chaque politique maximale calculée à l'étape précédente ;
- l'algorithme calcule Φ_t^+ à chaque étape en reposant sur la propriété suivante :

$$\forall t > 0, \Phi_t^* = \Phi_t^+$$

(Cette propriété est très intéressante quand on veut rapidement une politique optimale sans les avoir toutes) ;

- pour obtenir une seule politique optimale, il est possible de ne calculer qu'une seule sous-politique optimale à chaque étape.

Nous avons présentés deux approches de résolution d'un PDMP :

➢ l'approche par la fonction de valeur, utilisée dans le cadre classique des PDMP et basée sur l'équation d'optimalité de Bellman pour le critère fini. On constate que la complétude de la relation de préférence sur X et l'additivité de la loi $\circ$ sont les conditions d'existence d'une politique optimale ;

➢ l'approche par les loteries, en plus du cadre classique des PDMP, peut être utilisée dans le cadre des préférences non classiques. On constate que les conditions d'existence d'une politique optimale sont : l'invariance par translation, la transitivité, l'indépendance de la relation de préférence sur les loteries et la complétude, la stabilité de la relation de préférence sur les politiques.

⋆ ⋆ Chapitre Trois ⋆ ⋆

UN EXEMPLE PRATIQUE DE PDMP

Dans ce chapitre, nous considérons le problème de **l'exemple 1.3.3** que subit un agent dans une entreprise. Nous allons le modéliser par un PDMP et le résoudre manuellement par l'approche par la fonction de valeur pour le critère fini.

3.1 Modélisation du problème

On rappelle que le PDMP considéré est la donnée du quadruplet $(\mathcal{S}, \mathcal{A}, T, R)$ avec :

- l'ensemble des états $\mathcal{S} = \{s_0; s_1; s_2; s_3\}$;
- l'ensemble des actions $\mathcal{A} = \{a_1; a_2; a_3\}$;
- $T : \mathcal{S} \times \mathcal{A} \to \Pi(\mathcal{S})$ la fonction de transition ;
- la fonction de récompense $R : \mathcal{S} \times \mathcal{A} \to (\mathbb{R}, +, \geq)$.

Le **tableau 3.1** fournit les informations sur les revenus associés aux différentes combinaisons d'états de décision ainsi que les probabilités de transitions.

TABLE 3.1 – Tableau de modélisation du PDMP

États(s)	**Actions**(a)	**Coûts**	**Gains**	$P(s'\|s,a)$	**nouvel état**(s)	$R(s,a)$
s_0	a_1	500	3000	$\frac{3}{4}$	s_0	2500
s_0	a_1	500	3000	$\frac{1}{4}$	s_1	2500
s_0	a_1	500	-	0	s_2	-
s_0	a_1	500	-	0	s_3	-
s_0	a_2	0	3000	$\frac{4}{5}$	s_1	3000
s_0	a_2	0	3000	$\frac{1}{5}$	s_3	3000
s_0	a_2	0	-	0	s_0	-
s_0	a_2	0	-	0	s_2	-
s_0	a_3	3000	3000	1	s_0	0
s_0	a_3	3000	3000	0	s_1	0
s_0	a_3	3000	3000	0	s_2	0
s_0	a_3	3000	3000	0	s_3	0
s_1	a_1	1000	1500	$\frac{4}{7}$	s_1	500
s_1	a_1	1000	1500	$\frac{2}{7}$	s_2	500
s_1	a_1	1000	1500	$\frac{1}{7}$	s_3	500
s_1	a_1	1000	-	0	s_0	-
s_1	a_2	0	1500	$\frac{4}{5}$	s_2	1500
s_1	a_2	0	1500	$\frac{1}{5}$	s_3	1500
s_1	a_2	0	-	0	s_0	-
s_1	a_2	0	-	0	s_1	-
s_1	a_3	3000	1500	1	s_0	-1500
s_1	a_3	3000	-	0	s_1	-
s_1	a_3	3000	-	0	s_2	-
s_1	a_3	3000	-	0	s_3	-
s_2	a_1	**1000**	**500**	$\frac{3}{4}$	s_2	**-500**
s_2	a_1	**1000**	**500**	$\frac{1}{4}$	s_3	**-500**
s_2	a_1	**1000**	**-**	0	s_1	**-**
s_2	a_1	**1000**	**-**	0	s_0	**-**

TABLE 3.2 – Tableau de modélisation du PDMP (suite)

États(s)	**Actions**(a)	**Coûts**	**Gains**	$P(s'\|s,a)$	**nouvel état**(s')	$R(s,a)$
s_2	a_2	**0**	**500**	$\frac{1}{2}$	s_2	**500**
s_2	a_2	**0**	**500**	$\frac{1}{2}$	s_3	**500**
s_2	a_2	**0**	**-**	0	s_1	**-**
s_2	a_2	**0**	**-**	0	s_0	**-**
s_2	a_3	**3000**	**500**	1	s_0	**-2500**
s_2	a_3	**3000**	**-**	0	s_1	**-**
s_2	a_3	**3000**	**-**	0	s_2	**-**
s_2	a_3	**3000**	**-**	0	s_3	**-**
s_3	a_3	3000	0	1	s_0	-3000
s_3	a_3	3000	-	0	s_1	-
s_3	a_3	3000	-	0	s_2	-
s_3	a_3	3000	-	0	s_3	-
s_3	a_1	2000	-	0	s_0	-
s_3	a_1	2000	0	$\frac{1}{7}$	s_1	-2000
s_3	a_1	2000	0	$\frac{2}{7}$	s_2	-2000
s_3	a_1	2000	0	$\frac{4}{7}$	s_3	-2000
s_3	a_2	0	-	0	s_0	-
s_3	a_2	0	-	0	s_1	-
s_3	a_2	0	-	0	s_2	-
s_3	a_2	0	0	1	s_3	0

Remarque: 3.1.1.

1- $P(s'|a,s)$ est la probabilité que la machine passe à l'état s' à un instant ou une étape $t+1$ lorsque l'agent a exécuté l'action a à l'état s à un instant (ou une étape) t.

2- $R(s,a)$ = Gain - Coût (exprimé en FCFA).

3- Les cases vides de la colonne des $R(s,a)$ représentent des récompenses qui n'existent pas et les cases vides de la colonne des gains, représentent ceux qui n'existent pas.

Fonction de valeur

Nous prenons pour horizon $n = 3$ et le problème modélisé en stratégie pure, a comme fonction de valeur :

$$\forall \phi \in \Phi, \forall s \in \mathcal{S}, V^{\phi}(s) = E^{\phi}\left[\sum_{t=0}^{2} R_t | s_0 = s\right]$$

où $\forall t \in \{0, 1, 2\}$, $R_t|s_0 = s$ est la récompense marginale obtenue à l'étape t lorsque l'état initial du processus est $s_0 = s$.

Équation de BELLMAN pour le critère fini

Soient $s \in \mathcal{S}$, $a \in \mathcal{A}$. D'après le **théorème 2.2**, il existe une unique fonction de valeur optimale $V^* = (V_3^*, V_2^*, V_1^*)$ solution des systèmes d'équations :

$$\forall s \in \mathcal{S}, V_{t+1}^*(s) = \max_{a \in \mathcal{A}} \left\{ R_{2-t}(s,a) + \sum_{s' \in \mathcal{S}} P_{2-t}(s'|a,s) V_t^*(s') \right\} \tag{3.1}$$

avec $t \in \{0, 1, 2\}$ et $V_0 = 0$.
D'après ce théorème, il existe au moins une politique optimale $\phi^* = (\delta_0^*, \delta_1^*, \delta_2^*)$ déterminée par :

$$\forall s \in \mathcal{S}, \delta_t^*(s) = \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ R_{2-t}(s,a) + \sum_{s' \in \mathcal{S}} P_{2-t}(s'|a,s) V_{2-t}^*(s') \right\} \tag{3.2}$$

avec $t \in \{0, 1, 2\}$.

Remarque: 3.1.2. Soit $i \in \{0, 1, 2\}$.

1- $R_i(s,a)$ est la récompense ou le revenu (gain ou perte) perçue par l'agent lorsqu'il exécute l'action a à un état s et à l'instant ou une étape i.

2- δ_i^* est une règle de décision optimale à l'instant (ou étape) i définie de $\mathcal{S}$ vers $\mathcal{A}$.

3- V_i^* est la fonction de valeur optimale à l'instant (ou étape) i définie de $\mathcal{S}$ vers $\mathbb{R}$ qui à un état s' associe $V_i^*(s')$ qui est la récompense moyenne obtenue à l'état s' lorsqu'il suit la règle de décision optimale δ_{2-i}^*.

3.2 Détermination de la fonction de valeur optimale

Nous allons procéder manuellement, en fonction des différentes valeurs d'éléments de $\mathcal{S}$, à la résolution des systèmes d'équations définis par la relation (3.1) pour déterminer la fonction

de valeur optimale $V^* = (V_3^*, V_2^*, V_1^*)$. Cette détermination se fera de façon progressive.

Détermination de V_1^*

• Pour $s = s_0$, on a :

$$\begin{aligned} V_1^*(s_0) &= \max_{a\in\mathcal{A}} \left\{ R_2(s_0,a) + \sum_{s'\in\mathcal{S}} P_2(s'|a,s_0)V_0^*(s') \right\} \\ &= \max \{R_2(s_0,a_1)\ ; R_2(s_0,a_2)\ ; R_2(s_0,a_3)\} \text{ car } V_0^* = 0 \\ &= \max \{2500 \ \ (a_1)\ ;\ 3000 \ \ (a_2)\ ;\ 0 \ \ (a_3)\} \\ &= 3000. \end{aligned}$$

• Pour $s = s_1$, on a :

$$\begin{aligned} V_1^*(s_1) &= \max_{a\in\mathcal{A}} \left\{ R_2(s_1,a) + \sum_{s'\in\mathcal{S}} P_2(s'|a,s_1)V_0^*(s') \right\} \\ &= \max \{R_2(s_1,a_1)\ ;\ R_2(s_1,a_2)\ ;\ R_2(s_1,a_3)\} \text{ car } V_0^* = 0 \\ &= \max \{500 \ \ (a_1)\ ; 1500 \ \ (a_2)\ ;\ -1500 \ \ (a_3)\} \\ &= 1500. \end{aligned}$$

• Pour $s = s_2$, on a :

$$\begin{aligned} V_1^*(s_2) &= \max_{a\in\mathcal{A}} \left\{ R_2(s_2,a) + \sum_{s'\in\mathcal{S}} P_2(s'|a,s_2)V_0^*(s') \right\} \\ &= \max \{R_2(s_2,a_1)\ ;\ R_2(s_2,a_2)\ ;\ R_2(s_2,a_3)\} \text{ car } V_0^* = 0 \\ &= \max \{-500 \ \ (a_1)\ ;\ 500 \ \ (a_2)\ ;\ -2500 \ \ (a_3)\} \\ &= 500. \end{aligned}$$

• Pour $s = s_3$, on a :

$$\begin{aligned} V_1^*(s_3) &= \max_{a\in\mathcal{A}} \left\{ R_2(s_3,a) + \sum_{s'\in\mathcal{S}} P_2(s'|a,s_3)V_0^*(s') \right\} \\ &= \max \{R_2(s_3,a_1)\ ;\ R_2(s_3,a_2)\ ;\ R_2(s_3,a_3)\} \text{ car } V_0^* = 0 \\ &= \max \{-2000 \ \ (a_1)\ ;\ 0\ (a_2)\ ;\ -3000 \ \ (a_3)\} \\ &= 0. \end{aligned}$$

On a ainsi déterminé la fonction de valeur optimale à l'instant 1 :

$V_1^* : \mathcal{S} \longrightarrow (\mathbb{R}, +, \geq)$ qui à $s \in \mathcal{S}$ associe $V_1^*(s) = \begin{cases} V_1^*(s_0) = 3000 \\ V_1^*(s_1) = 1500 \\ V_1^*(s_2) = 500 \\ V_1^*(s_3) = 0 \end{cases}$

Détermination de V_2^*.

• Pour $s = s_0$, on a :

$$V_2^*(s_0) = \max_{a\in\mathcal{A}} \left\{ R_1(s_0,a) + \sum_{s'\in\mathcal{S}} P_1(s'|a,s_0)V_1^*(s') \right\}$$

$$= \max\{R_1(s_0,a_1) + [P_1(s_0|a_1,s_0)V_1^*(s_0) + P_1(s_1|a_1,s_0)V_1^*(s_1) + P_1(s_2|a_1,s_0)V_1^*(s_2) + P_1(s_3|a_1,s_0)V_1^*(s_3)]\,;\, R_1(s_0,a_2) + [P_1(s_0|a_2,s_0)V_1^*(s_0) + P_1(s_1|a_2,s_0)V_1^*(s_1) + P(s_2|a_2,s_0)V_1^*(s_2) + P_1(s_3|a_2,s_0)V_1^*(s_3)]\,;\, R_1(s_0,a_3) + [P_1(s_0|a_3,s_0)V_1^*(s_0) + P_1(s_1|a_3,s_0)V_1^*(s_1) + P_1(s_2|a_3,s_0)V_1^*(s_2) + P_1(s_3|a_3,s_0)V_1^*(s_3)]\}$$

$$= \max\{2500 + [\tfrac{3}{4}\times 3000 + \tfrac{1}{4}\times 1500 + 0\times 500 + 0\times 0](a_1); 3000 + [0\times 3000 + \tfrac{4}{5}\times 1500 + 0\times 500 + \tfrac{1}{5}\times 0](a_2); 0 + [1\times 3000 + 0\times 1500 + 0\times 500 + 0\times 0](a_3)\}$$

$$= \max\ \{5125\ (a_1)\ ;\ 4200\ (a_2)\ ;\ 3000\ (a_3)\}$$

$$= 5125\ .$$

• Pour $s = s_1$, on a :

$$V_2^*(s_1) = \max_{a\in\mathcal{A}} \left\{ R_1(s_1,a) + \sum_{s'\in\mathcal{S}} P_1(s'|a,s_1)V_1^*(s') \right\}$$

$$= \max\{R_1(s_1,a_1) + [P_1(s_0|a_1,s_1)V_1^*(s_0) + P_1(s_1|a_1,s_1)V_1^*(s_1) + P_1(s_2|a_1,s_1)V_1^*(s_2) + P_1(s_3|a_1,s_1)V_1^*(s_3)]\,;\, R_1(s_1,a_2) + [P_1(s_0|a_2,s_1)V_1^*(s_0) + P_1(s_1|a_2,s_1)V_1^*(s_1) + P_1(s_2|a_2,s_1)V_1^*(s_2) + P_1(s_3|a_2,s_1)V_1^*(s_3)]\,;\, R_1(s_1,a_3) + [P_1(s_0|a_3,s_1)V_1^*(s_0) + P_1(s_1|a_3,s_1)V_1^*(s_1) + P_1(s_2|a_3,s_1)V_1^*(s_2) + P_1(s_3|a_3,s_1)V_1^*(s_3)]\}$$

$$= \max\{500 + [0\times 3000 + \tfrac{4}{7}\times 1500 + \tfrac{2}{7}\times 500 + \tfrac{1}{7}\times 0](a_1); 1500 + [0\times 3000 + 0\times 1500 + \tfrac{4}{5}\times 500 + \tfrac{1}{5}\times 0](a_2); -1500 + [1\times 3000 + 0\times 1500 + 0\times 500 + 0\times 0](a_3)\}$$

$$= max\ \ \{1500\ \ (a_1)\ ;\ 1900\ \ (a_2)\ ;\ 1500\ \ (a_3)\}$$

$$= 1900\ .$$

• Pour $s = s_2$, on a :

$$V_2^*(s_2) = \max_{a\in\mathcal{A}} \left\{ R_1(s_2,a) + \sum_{s'\in\mathcal{S}} P_1(s'|a,s_2)V_1^*(s') \right\}$$

$$\begin{aligned}
&= \max\{R_1(s_2,a_1) + [P_1(s_0|a_1,s_2)V_1^*(s_0) + P_1(s_1|a_1,s_2)V_1^*(s_1) + P_1(s_2|a_1,s_2)V_1^*(s_2) \\
&\quad + P_1(s_3|a_1,s_2)V_1^*(s_3)]\,; R_1(s_2,a_2) + [P_1(s_0|a_2,s_2)V_1^*(s_0) + P_1(s_1|a_2,s_2)V_1^*(s_1) \\
&\quad + P_1(s_2|a_2,s_2)V_1^*(s_2) + P_1(s_3|a_2,s_2)V_1^*(s_3)]\,; R_1(s_2,a_3) + [P_1(s_0|a_3,s_2)V_1^*(s_0) \\
&\quad + P_1(s_1|a_3,s_2)V_1^*(s_1) + P_1(s_2|a_3,s_2)V_1^*(s_2) + P_1(s_3|a_3,s_2)V_1^*(s_3)]\}
\end{aligned}$$

$$\begin{aligned}
&= \max\{-500 + [0\times 3000 + 0\times 1500 + \tfrac{3}{4}\times 500 + \tfrac{1}{4}\times 0](a_1); 500 + [0\times 3000 \\
&\quad + 0\times 1500 + \tfrac{1}{2}\times 500 + \tfrac{1}{2}\times 0](a_2); -2500 + [1\times 3000 + 0\times 1500 + 0\times 500 \\
&\quad + 0\times 0](a_3)\}
\end{aligned}$$

$$= \max \ \{-125 \ \ (a_1);\ 750\ (a_2);\ 500 \ \ (a_3)\}$$

$$= 750 \quad .$$

• Pour $s = s_3$, on a :

$$V_2^*(s_3) = \max_{a\in\mathcal{A}} \left\{ R_1(s_3,a) + \sum_{s'\in\mathcal{S}} P_1(s'|a,s_3)V_1^*(s') \right\}$$

$$\begin{aligned}
&= \max\{R_1(s_3,a_1) + [P_1(s_0|a_1,s_3)V_1^*(s_0) + P_1(s_1|a_1,s_3)V_1^*(s_1) + P_1(s_2|a_1,s_3)V_1^*(s_3) \\
&\quad + P_1(s_3|a_1,s_3)V_1^*(s_3)]\,; R_1(s_3,a_2) + [P_1(s_0|a_2,s_3)V_1^*(s_0) + P_1(s_1|a_2,s_3)V_1^*(s_1) \\
&\quad + P_1(s_2|a_2,s_3)V_1^*(s_2) + P_1(s_3|a_2,s_3)V_1^*(s_3)]\,; R_1(s_3,a_3) + [P_1(s_0|a_3,s_3)V_1^*(s_0) \\
&\quad + P_1(s_1|a_3,s_3)V_1^*(s_1) + P_1(s_2|a_3,s_3)V_1^*(s_2) + P_1(s_3|a_3,s_3)V_1^*(s_3)]\}
\end{aligned}$$

$$\begin{aligned}
&= \max\{-2000 + [0\times 3000 + \tfrac{1}{7}\times 1500 + \tfrac{2}{7}\times 500 + \tfrac{4}{7}\times 0](a_1); 0 + [0\times 3000 \\
&\quad + 0\times 1500 + 0\times 500 + 1\times 0](a_2); -3000 + [1\times 3000 + 0\times 1500 + 0\times 500 \\
&\quad + 0\times 0](a_3)\}
\end{aligned}$$

$$= \max \ \{-1642.85 \ \ (a_1)\ ;\ 0 \ \ (a_2)\ ;\ 0 \ \ (a_3)\}$$

$= 0.$

On a ainsi déterminé la fonction de valeur optimale à l'instant 2 :

$$V_2^* : \mathcal{S} \longrightarrow (\mathbb{R}, +, \geq) \text{ qui à } s \in \mathcal{S} \text{ associe } V_2^*(s) = \begin{cases} V_2^*(s_0) &= 5125 \\ V_2^*(s_1) &= 1900 \\ V_2^*(s_2) &= 750 \\ V_2^*(s_3) &= 0 \end{cases}$$

Détermination de V_3^*

• Pour $s = s_0$, on a :

$$V_3^*(s_0) = \max_{a \in \mathcal{A}} \left\{ R_0(s_0, a) + \sum_{s' \in \mathcal{S}} P_0(s'|a, s_0) V_2^*(s') \right\}$$

$$\begin{aligned} &= \max\{R_0(s_0, a_1) + [P_0(s_0|a_1, s_0)V_2^*(s_0) + P_0(s_1|a_1, s_0)V_2^*(s_1) + P_0(s_2|a_1, s_0)V_1^*(s_2) \\ &\quad + P_0(s_3|a_1, s_0)V_1^*(s_3)]\,; R_0(s_2, a_2) + [P_0(s_0|a_2, s_0)V_2^*(s_0) + P_0(s_1|a_2, s_0)V_2^*(s_1) \\ &\quad + P_0(s_2|a_2, s_0)V_1^*(s_2) + P_0(s_3|a_2, s_0)V_2^*(s_3)]\,; R_0(s_2, a_3) + [P_0(s_0|a_3, s_0)V_2^*(s_0) \\ &\quad + P_0(s_1|a_3, s_0)V_2^*(s_1) + P_0(s_2|a_3, s_0)V_2^*(s_2) + P_0(s_3|a_3, s_0)V_2^*(s_3)]\} \end{aligned}$$

$$\begin{aligned} &= \max\{2500 + [\tfrac{3}{4} \times 5125 + \tfrac{1}{4} \times 1900 + 0 \times 750 + 0 \times 0](a_1); 3000 + [0 \times 5125 \\ &\quad + \tfrac{4}{5} \times 1900 + 0 \times 750 + \tfrac{1}{5} \times 0](a_2); 0 + [1 \times 5125 + 0 \times 1900 \\ &\quad + 0 \times 750 + 0 \times 0](a_3)\} \end{aligned}$$

$$= \max \quad \{6818,75 \quad (a_1);\ 4520 \quad (a_2); 5125 \quad (a_3)\}$$

$$= 6818,75\ .$$

• Pour $s = s_1$, on a :

$$V_3^*(s_1) = \max_{a \in \mathcal{A}} \left\{ R_0(s_1, a) + \sum_{s' \in \mathcal{S}} P_0(s'|a, s_1) V_2^*(s') \right\}$$

$$\begin{aligned} &= \max\{R_0(s_1, a_1) + [P_0(s_0|a_1, s_1)V_2^*(s_0) + P_0(s_1|a_1, s_1)V_2^*(s_1) + P_0(s_2|a_1, s_1)V_2^*(s_2) \\ &\quad + P_0(s_3|a_1, s_1)V_2^*(s_3)]\,; R_0(s_2, a_2) + [P_0(s_0|a_2, s_1)V_2^*(s_0) + P_0(s_1|a_2, s_1)V_2^*(s_1) \\ &\quad + P_0(s_2|a_2, s_1)V_2^*(s_2) + P_0(s_3|a_2, s_1)V_2^*(s_3)]\,; R_0(s_2, a_3) + [P_0(s_0|a_3, s_1)V_2^*(s_0) \end{aligned}$$

$$+ P_0(s_1|a_3, s_1)V_2^*(s_1) + P_0(s_2|a_3, s_1)V_2^*(s_2) + P_0(s_3|a_3, s_1)V_2^*(s_3)]\}$$

$$= \max\{500 + [0 \times 5125 + \tfrac{4}{7} \times 1900 + \tfrac{2}{7} \times 750 + \tfrac{1}{7} \times 0](a_1); 1500 + [0 \times 5125$$
$$+ 0 \times 1900 + \tfrac{4}{5} \times 750 + \tfrac{1}{5} \times 0](a_2); -1500 + [1 \times 5125 + 0 \times 1900$$
$$+ 0 \times 750 + 0 \times 0](a_3)\}$$

$$= \max \ \{1800 \ \ (a_1); 2100 \ \ (a_2); 3625 \ \ (a_3)\}$$

$$= 3625 \ \ .$$

• Pour $s = s_2$, on a :

$$V_3^*(s_2) = \max_{a \in \mathcal{A}} \left\{ R_0(s_2, a) + \sum_{s' \in \mathcal{S}} P_0(s'|a, s_2)V_2^*(s') \right\}$$

$$= \max\{R_0(s_2, a_1) + [P_0(s_0|a_1, s_2)V_2^*(s_0) + P_0(s_1|a_1, s_2)V_2^*(s_1) + P_0(s_2|a_1, s_2)V_2^*(s_2)$$
$$+ P_0(s_3|a_1, s_2)V_2^*(s_3)]\,; R_0(s_2, a_2) + [P_0(s_0|a_2, s_2)V_2^*(s_0) + P_0(s_1|a_2, s_2)V_2^*(s_1)$$
$$+ P_0(s_2|a_2, s_2)V_2^*(s_2) + P_0(s_3|a_2, s_2)V_2^*(s_3)]\,; R_0(s_2, a_3) + [P_0(s_0|a_3, s_2)V_2^*(s_0)$$
$$+ P_0(s_1|a_3, s_2)V_2^*(s_1) + P_0(s_2|a_3, s_2)V_2^*(s_2) + P_0(s_3|a_3, s_2)V_2^*(s_3)]\}$$

$$= \max\{-500 + [0 \times 5125 + 0 \times 1900 + \tfrac{3}{4} \times 750 + \tfrac{1}{4} \times 0](a_1); 500 + [0 \times 5125$$
$$+ 0 \times 1900 + \tfrac{1}{2} \times 750 + \tfrac{1}{2} \times 0](a_2); -2500 + [1 \times 5125 + 0 \times 1900$$
$$+ 0 \times 750 + 0 \times 0](a_3)\}$$

$$= \max \ \{62,5 \ \ (a_1); 875 \ \ (a_2); 2625 \ \ (a_3)\}$$

$$= 2625 \ \ .$$

• Pour $s = s_3$, on a :

$$V_3^*(s_3) = \max_{a \in \mathcal{A}} \ \left\{ R_0(s_3, a) + \sum_{s' \in \mathcal{S}} P_0(s'|a, s_3)V_2^*(s') \right\}$$

$$= \max\{R_0(s_3, a_1) + [P_0(s_0|a_1, s_3)V_2^*(s_0) + P_0(s_1|a_1, s_3)V_2^*(s_1) + P_0(s_2|a_1, s_3)V_2^*(s_3)$$
$$+ P_0(s_3|a_1, s_3)V_2^*(s_3)]\,; R_0(s_3, a_2) + [P_0(s_0|a_2, s_3)V_2^*(s_0) + P_0(s_1|a_2, s_3)V_2^*(s_1)$$

$$+ P_0(s_2|a_2,s_3)V_2^*(s_2) + P_0(s_3|a_2,s_3)V_2^*(s_3)]\,;\, R_0(s_3,a_3) + [P_0(s_0|a_3,s_3)V_2^*(s_0)$$
$$+ P_0(s_1|a_3,s_3)V_2^*(s_1) + P_0(s_2|a_3,s_3)V_2^*(s_2) + P_0(s_3|a_3,s_3)V_2^*(s_3)]\}$$

$$= \max\{-2000 + [0\times 5125 + \tfrac{1}{7}\times 1900 + \tfrac{2}{7}\times 750 + \tfrac{4}{7}\times 0](a_1); 0 + [0\times 5125$$
$$+ 0\times 1900 + 0\times 750 + 1\times 0](a_2); -3000 + [1\times 5125 + 0\times 1900 + 0\times 750$$
$$+ 0\times 0](a_3)\}$$

$$= \max\{-1514,28 \quad (a_1);\ 0 \quad (a_2); 2125 \quad (a_3)\}$$

$$= 2125 \quad .$$

On a ainsi déterminé la fonction de valeur optimale à l'instant 3 :

$$V_3^* : \mathcal{S} \longrightarrow (\mathbb{R}, +, \geq) \text{ qui à } s \in \mathcal{S} \text{ associe } V_3^*(s) = \begin{cases} V_3^*(s_0) &= 6818,75 \\ V_3^*(s_1) &= 3625 \\ V_3^*(s_2) &= 2625 \\ V_3^*(s_3) &= 2125 \end{cases}$$

D'après les calculs précédents, $V^* = (V_3^*, V_2^*, V_1^*)$ est la fonction de valeur optimale de notre modèle.

3.3 Recherche d'une politique optimale

On cherche une politique optimale de la forme $\phi^* = (\delta_0^*, \delta_1^*, \delta_2^*)$ de notre modèle associée à la fonction de valeur optimale $V^* = (V_3^*, V_2^*, V_1^*)$. Nous allons procéder en fonction des différentes valeurs d'éléments de $\mathcal{S}$, à la résolution des systèmes d'équations définis par la relation (3.2) .

Recherche de la règle de décision optimale δ_0^*

• Pour $s = s_0$, on a :

$$\delta_0^*(s_0) \in \underset{a\in\mathcal{A}}{\text{argmax}} \left\{ R_0(s_0,a) + \sum_{s'\in\mathcal{S}} P_0(s'|a,s_0)V_2^*(s') \right\}$$

$$\Longleftrightarrow \delta_0^*(s_0) \in \underset{a\in\mathcal{A}}{\text{argmax}} \ \{6818,75 \ \ (a_1)\,; 4520 \ \ (a_2)\,;\ 5125 \ \ (a_3)\}$$

$$\Longleftrightarrow \delta_0^*(s_0) \in \{a_1\}$$

$$\Longleftrightarrow \delta_0^*(s_0) = a_1 \ .$$

• Pour $s = s_1$, on a :

$$\delta_0^*(s_1) \in \underset{a\in\mathcal{A}}{\text{argmax}} \left\{ R_0(s_1, a) + \sum_{s^{'}\in\mathcal{S}} P_0(s^{'}|a, s_1) V_2^*(s^{'}) \right\}$$

$$\Longleftrightarrow \delta_0^*(s_1) \in \underset{a\in\mathcal{A}}{\text{argmax}} \ \{1800 \ (a_1) \ ; \ 2100 \ (a_2) \ ; 3625 \ (a_3)\}$$

$$\Longleftrightarrow \delta_0^*(s_1) \in \{a_3\}$$

$$\Longleftrightarrow \delta_0^*(s_1) = a_3 \ .$$

• Pour $s = s_2$, on a :

$$\delta_0^*(s_2) \in \underset{a\in\mathcal{A}}{\text{argmax}} \left\{ R_0(s_2, a) + \sum_{s^{'}\in\mathcal{S}} P_0(s^{'}|a, s_2) V_2^*(s^{'}) \right\}$$

$$\Longleftrightarrow \delta_0^*(s_2) \in \underset{a\in\mathcal{A}}{\text{argmax}} \ \{62,5 \ (a_1) \ ; \ 875 \ (a_2) \ ; \ 2625 \ (a_3)\}$$

$$\Longleftrightarrow \delta_0^*(s_2) \in \{a_3\}$$

$$\Longleftrightarrow \delta_0^*(s_2) = a_3 \ .$$

• Pour $s = s_3$, on a :

$$\delta_0^*(s_3) \in \underset{a\in\mathcal{A}}{\text{argmax}} \left\{ R_0(s_3, a) + \sum_{s^{'}\in\mathcal{S}} P_0(s^{'}|a, s_3) V_2^*(s^{'}) \right\}$$

$$\Longleftrightarrow \delta_0^*(s_3) \in \underset{a\in\mathcal{A}}{\text{argmax}} \ \{-1514,28 \ (a_1) \ ; \ 0 \ (a_2) \ ; \ 2125 \ (a_3)\}$$

$$\Longleftrightarrow \delta_0^*(s_3) \in \{a_3\}$$

$$\Longleftrightarrow \delta_0^*(s_3) = a_3 \ .$$

On a ainsi construit la règle de décision optimale δ_0^* définie par :

$$\delta_0^* : \mathcal{S} \longrightarrow \mathcal{A} \text{ qui à } s \in \mathcal{S} \text{ associe } \delta_0^*(s) = \begin{cases} \delta_0^*(s_0) &= a_1 \\ \delta_0^*(s_1) &= a_3 \\ \delta_0^*(s_2) &= a_3 \\ \delta_0^*(s_3) &= a_3 \end{cases}$$

Interprétation 3.1. La règle de décision optimale δ_0^* signifie qu'à l'étape ou l'instant 0, l'agent doit :

- entretenir la machine si elle est à l'état neuf ($\delta_0^*(s_0) = a_1$) ;
- rénover la machine si elle est en bon état ($\delta_0^*(s_1) = a_3$) ;
- rénover la machine si elle est en mauvais état ($\delta_0^*(s_2) = a_3$) ;
- rénover la machine si elle est à l'état en panne ($\delta_0^*(s_3) = a_3$).

Recherche de la règle de décision optimale δ_1^*

• Pour $s = s_0$, on a :

$$\delta_1^*(s_0) \in \underset{a\in\mathcal{A}}{\text{argmax}} \left\{ R_1(s_0, a) + \sum_{s'\in\mathcal{S}} P_1(s'|a, s_0) V_1^*(s') \right\}$$

$$\iff \delta_1^*(s_0) \in \underset{a\in\mathcal{A}}{\text{argmax}} \ \{5125 \ (a_1) \ ; \ 4200 \ (a_2) \ ; \ 3000 \ (a_3)\}$$

$$\iff \delta_1^*(s_0) \in \{a_1\}$$

$$\iff \delta_1^*(s_0) = a_1 \ .$$

• Pour $s = s_1$, on a :

$$\delta_1^*(s_1) \in \underset{a\in\mathcal{A}}{\text{argmax}} \left\{ R_0(s_1, a) + \sum_{s'\in\mathcal{S}} P_0(s'|a, s_1) V_1^*(s') \right\}$$

$$\iff \delta_1^*(s_1) \in \underset{a\in\mathcal{A}}{\text{argmax}} \ \{1500 \ (a_1) \ ; \ 1900 \ (a_2); \ 1500 \ (a_3)\}$$

$$\iff \delta_1^*(s_1) \in \{a_2\}$$

$$\Longleftrightarrow \delta_1^*(s_1) = a_2 \ .$$

• Pour $s = s_2$, on a :

$$\delta_1^*(s_2) \in \underset{a\in\mathcal{A}}{\text{argmax}} \left\{ R_0(s_2,a) + \sum_{s^{'}\in\mathcal{S}} P_0(s^{'}|a,s_2)V_1^*(s^{'}) \right\}$$

$$\Longleftrightarrow \delta_1^*(s_2) \in \underset{a\in\mathcal{A}}{\text{argmax}} \ \{-125 \ \ (a_1) \ ; \ 750 \ \ (a_2) \ ; \ 500 \ \ (a_3)\}$$

$$\Longleftrightarrow \delta_1^*(s_2) \in \{a_2\}$$

$$\Longleftrightarrow \delta_1^*(s_2) = a_2 \ .$$

• Pour $s = s_3$, on a :

$$\delta_1^*(s_3) \in \underset{a\in\mathcal{A}}{\text{argmax}} \left\{ R_0(s_3,a) + \sum_{s^{'}\in\mathcal{S}} P_0(s^{'}|a,s_3)V_1^*(s^{'}) \right\}$$

$$\Longleftrightarrow \delta_1^*(s_3) \in \underset{a\in\mathcal{A}}{\text{argmax}} \ \{-1642,85 \ \ (a_1) \ ; \ 0 \ \ (a_2) \ ; \ 0 \ (a_3)\}$$

$$\Longleftrightarrow \delta_1^*(s_3) \in \{a_2, a_3\}$$

$$\Longleftrightarrow \delta_1^*(s_3) = a_2 \text{ ou } \delta_1^*(s_3) = a_3 \ .$$

On a ainsi construit deux règles de décisions optimales δ_1^* et $\delta_1^{'*}$ définies respectivement par :

$$\delta_1^* : \mathcal{S} \longrightarrow \mathcal{A} \text{ qui à } s \in \mathcal{S} \text{ associe } \delta_1^*(s) = \begin{cases} \delta_1^*(s_0) = a_1 \\ \delta_1^*(s_1) = a_2 \\ \delta_1^*(s_2) = a_2 \\ \delta_1^*(s_3) = a_2 \end{cases}$$

et

$$\delta_1^{'*} : \mathcal{S} \longrightarrow \mathcal{A} \text{ qui à } s \in \mathcal{S} \text{ associe } \delta_1^{'*}(s) = \begin{cases} \delta_1^{'*}(s_0) &= a_1 \\ \delta_1^{'*}(s_1) &= a_2 \\ \delta_1^{'*}(s_2) &= a_2 \\ \delta_1^{'*}(s_3) &= a_3 \end{cases} .$$

Interprétation 3.2. La règle de décision optimale δ_1^* signifie qu'à l'étape ou l'instant 1, l'agent :

- doit entretenir la machine si elle est à l'état neuf ($\delta_1^*(s_0) = a_1$) ;
- ne doit rien faire si la machine est en bon état ($\delta_1^*(s_1) = a_2$) ;
- ne doit rien faire si la machine est en mauvais état ($\delta_1^*(s_2) = a_2$) ;
- ne doit rien faire si la machine est à l'état en panne ($\delta_1^*(s_3) = a_2$).

Interprétation 3.3. La règle de décision optimale $\delta_1'^*$ signifie qu'à l'étape ou l'instant 1, l'agent :

- doit entretenir la machine si elle est à l'état neuf ($\delta_1'^*(s_0) = a_1$) ;
- ne doit rien faire si la machine est en bon état ($\delta_1'^*(s_1) = a_2$) ;
- ne doit rien faire si la machine est en mauvais état ($\delta_1'^*(s_2) = a_2$) ;
- doit rénover la machine si elle est à l'état en panne ($\delta_1'^*(s_3) = a_3$).

Recherche de la règle de décision optimale δ_2^*

• Pour $s = s_0$, on a :

$$\delta_2^*(s_0) \in \underset{a\in\mathcal{A}}{\text{argmax}} \left\{ R_1(s_0, a) + \sum_{s'\in\mathcal{S}} P_1(s'|a, s_0) V_0^*(s') \right\}$$

$$\Longleftrightarrow \delta_2^*(s_0) \in \underset{a\in\mathcal{A}}{\text{argmax}} \ \{R_2(s_0, a_1) \ ; \ R_2(s_0, a_2) \ ; \ R_2(s_0, a_3)\} \text{ car } V_0^* = 0$$

$$\Longleftrightarrow \delta_2^*(s_0) \in \{2500 \ \ (a_1); \ 3000 \ \ (a_2) \, ; 0 \ \ (a_3)\}$$

$$\Longleftrightarrow \delta_2^*(s_0) = a_2,$$

• Pour $s = s_1$, on a :

$$\delta_2^*(s_1) \in \underset{a\in\mathcal{A}}{\text{argmax}} \left\{ R_1(s_1, a) + \sum_{s'\in\mathcal{S}} P_1(s'|a, s_1) V_0^*(s') \right\}$$

$$\Longleftrightarrow \delta_2^*(s_1) \in \underset{a\in\mathcal{A}}{\text{argmax}} \ \{R_2(s_1, a_1) \ ; \ R_2(s_1, a_2) \ ; \ R_2(s_1, a_3)\} \text{ car } V_0^* = 0$$

$$\Longleftrightarrow \delta_2^*(s_1) \in \{500\ \ (a_1)\ ;\ 1500\ \ (a_2)\ ; -1500\ \ (a_3)\}$$

$$\Longleftrightarrow \delta_2^*(s_1) = a_2.$$

• Pour $s = s_2$, on a :

$$\delta_2^*(s_2) \in \underset{a\in\mathcal{A}}{\text{argmax}} \left\{ R_1(s_2,a) + \sum_{s^{'}\in\mathcal{S}} P_1(s^{'}|a,s_2)V_0^*(s^{'}) \right\}$$

$$\Longleftrightarrow \delta_2^*(s_2) \in \underset{a\in\mathcal{A}}{\text{argmax}}\ \{R_2(s_2,a_1)\ ;\ R_2(s_2,a_2)\ ;\ R_2(s_2,a_3)\} \text{ car } V_0^* = 0$$

$$\Longleftrightarrow \delta_2^*(s_2) \in \{-500\ \ (a_1)\ ;\ 500\ \ (a_2)\ ; -2500\ \ (a_3)\}$$

$$\Longleftrightarrow \delta_2^*(s_2) = a_2\ .$$

• Pour $s = s_3$, on a :

$$\delta_2^*(s_3) \in \underset{a\in\mathcal{A}}{\text{argmax}} \left\{ R_1(s_3,a) + \sum_{s^{'}\in\mathcal{S}} P_1(s^{'}|a,s_3)V_0^*(s^{'}) \right\}$$

$$\Longleftrightarrow \delta_2^*(s_3) \in \underset{a\in\mathcal{A}}{\text{argmax}}\ \{R_2(s_3,a_1)\ ;\ R_2(s_3,a_2)\ ;\ R_2(s_3,a_3)\} \text{ car } V_0^* = 0$$

$$\Longleftrightarrow \delta_2^*(s_3) \in \{-2000\ \ (a_1);\ 0\ \ (a_2); -3000\ \ (a_3)\}$$

$$\Longleftrightarrow \delta_2^*(s_3) = a_2\ .$$

On a ainsi construit la règle de décision optimale :

$$\delta_2^* : \mathcal{S} \longrightarrow \mathcal{A} \text{ qui à } s \in \mathcal{S} \text{ associe } \delta_2^*(s) = \begin{cases} \delta_2^*(s_0) &= a_2 \\ \delta_2^*(s_1) &= a_2 \\ \delta_2^*(s_2) &= a_2 \\ \delta_2^*(s_3) &= a_2 \end{cases} ;$$

Interprétation 3.4. Cette règle de décision est stationnaire car elle est constante, ce qui signifie qu'à l'étape ou l'instant 2, l'agent ne doit rien faire pour n'importe quel état de la machine.

D'après les résultats précédents, le modèle PDMP admet deux politiques optimales à 3 horizons $\phi = (\delta_0^*, \delta_1^*, \delta_2^*)$ et $\phi' = (\delta_0'^*, \delta_1'^*, \delta_2'^*)$ associées à la fonction de valeur optimale $V^* = (V_3^*, V_2^*, V_1^*)$.

3.4 Interprétation des résultats

Interprétation 3.5. Cas de la politique ϕ

La politique ϕ signifie à :

- l'étape ou l'instant 0, l'agent doit entretenir la machine si elle est à l'état neuf pour un gain moyen futur de 6818,75 FCFA, rénover la machine si elle est en bon état pour un gain moyen futur de 3625 FCFA, rénover la machine si elle est en mauvais état pour un gain moyen futur de 2625 FCFA, rénover la machine si elle est à l'état en panne pour un gain moyen futur de 2125 FCFA ;
- l'étape ou l'instant 1, l'agent doit entretenir la machine si elle est à l'état neuf pour un gain moyen futur de 5125 FCFA, ne doit rien faire si la machine est en bon état pour un gain moyen futur de 1900 FCFA, ne doit rien faire si la machine est en mauvais état pour un gain moyen futur de 750 FCFA, ne doit rien faire si la machine est à l'état en panne pour un gain moyen futur de 0 FCFA ;
- l'étape ou l'instant 2, l'agent ne doit rien faire si la machine est à l'état neuf pour un gain moyen futur de 3000 FCFA, ne doit rien faire si la machine est en bon état pour un gain moyen futur de 1500 FCFA, ne doit rien faire si la machine est en mauvais état pour un gain moyen futur de 500 FCFA, ne doit rien faire si la machine est à l'état en panne pour un gain moyen futur de 0 FCFA.

Interprétation 3.6. Cas de la politique ϕ'

La politique ϕ' signifie à :

- l'étape ou l'instant 0, l'agent doit entretenir la machine si elle est à l'état neuf, pour un gain moyen futur de 6818,75 FCFA, rénover la machine si elle est en bon état, pour un gain moyen futur de 3625 FCFA, rénover la machine si elle est en mauvais état, pour un gain moyen futur de 2625 FCFA, rénover la machine si elle est à l'état en panne, pour un gain moyen futur de 2125 FCFA.

- ➤ l'étape ou l'instant 1, l'agent doit entretenir la machine si elle est à l'état neuf, pour un gain moyen futur de 5125 FCFA, ne doit rien faire si la machine est en bon état, pour un gain moyen futur de 1900 FCFA, ne doit rien faire si la machine est en mauvais état, pour un gain moyen futur de 750 FCFA, doit rénover la machine si elle est à l'état en panne, pour un gain moyen futur de 0 FCFA ;
- ➤ l'étape ou l'instant 2, l'agent ne doit rien faire si la machine est à l'état neuf, pour un gain moyen futur de 3000 FCFA, ne doit rien faire si la machine est en bon état, pour un gain moyen futur de 1500 FCFA, ne doit rien faire si la machine est en mauvais état, pour un gain moyen futur de 500 FCFA, ne doit rien faire si la machine est à l'état en panne pour un gain moyen futur de 0 FCFA.

On remarque que les politiques optimales ϕ et ϕ' diffèrent seulement à l'étape 1 mais conduisent aux mêmes gains moyens futurs en fonction des états et actions. On remarque aussi que les gains moyens futurs sont disposés de façon décroissante.

La mise en pratique de l'une de ces politiques optimales par l'agent permettra d'éviter une faillite et de faire une bonne planification financière.

⋆ ⋆ Chapitre Quatre ⋆ ⋆

IMPLICATION PÉDAGOGIQUE

La réalisation de ce mémoire constitue une étape importante et indispensable dans le cadre de notre formation. Nous avons rassemblé et exploité l'essentiel des articles qui ont traits à notre thème. Cependant, en tant que futur enseignant, ce travail est la source de profondes satisfactions : découvertes intellectuelles, enrichissement personnel et l'utilisation des nouvelles technologies de l'information et de la communication.

4.1 Réalisation d'une expérience de travail intellectuel, approfondie et autonome

La réalisation de ce travail est une activité qui nous a permis d'apprendre plusieurs choses telles que :

- délimiter un problème ;
- découvrir et rassembler une documentation à son propos ;
- ordonner des matériaux ;
- conduire une réflexion personnelle sur le problème choisi ;
- souvent, à établir des contacts directs avec des personnes, des institutions, des champs d'activités ;
- apprendre à ordonner ses propres idées et à les formuler d'une manière compréhensible par autrui (un élève par exemple).

Apport dans l'enseignement de la notion de probabilité en terminales scientifiques

Nous avons étudié la notion de chaîne de Markov, énoncé la notion de probabilité conditionnelle et de processus stochastique. Ce sont des notions qui sont souvent enseignées dans les classes de terminales scientifiques sous des autres formes.

- La notion de chaîne de Markov est souvent présentée sous la forme des arbres de choix, pour représenter les différentes probabilités d'un problème et pour faciliter la compréhension chez les élèves.
- La notion de processus stochastique est encore appelée variable aléatoire réelle lorsqu'il comporte une seule variable aléatoire et l'ensemble des états est un sous ensemble de $\mathbb{R}$.

4.2 Utilisation des nouvelles technologies de l'information et de la communication

Les outils informatiques (calculatrices, logiciels de géométrie, ...)

L'objectif de l'enseignement des mathématiques est de développer conjointement et progressivement les capacités d'expérimentation et de raisonnement, d'imagination et d'analyse critique. À travers la résolution de problèmes, la modélisation de quelques situations et l'apprentissage progressif de la démonstration, les élèves peuvent prendre conscience petit à petit de ce qu'est une véritable activité mathématique, identifier un problème, expérimenter sur des exemples, conjecturer un résultat, bâtir une argumentation, mettre en forme une solution, contrôler les résultats obtenus et évaluer leur pertinence en fonction du problème étudié. Par ses spécificités, l'outil informatique complète les moyens à la disposition des enseignants et des élèves pour mettre en œuvre ces différents aspects d'une véritable activité mathématique. En effet, il permet notamment :

- d'obtenir rapidement une représentation d'un problème, d'un concept afin de lui donner du sens et de favoriser son appropriation par l'élève ;
- de relier différents aspects (algébrique, géométrique, ...) d'un même concept ou d'une même situation ;
- d'explorer des situations en faisant apparaître de façon dynamique différentes configurations ;
- d'émettre des conjectures à partir d'une expérimentation interactive lors de l'étude d'un problème comportant des questions ouvertes ou d'une certaine complexité, et de procéder à des premières vérifications ;
- de se consacrer à la résolution de problèmes issus de situations courantes, alors que les calculs sont longs ou complexes ;

- de procéder rapidement à la vérification de certains résultats obtenus.

Internet

L'usage de l'internet (ou d'un intranet) en mathématiques en est aux balbutiements, mais déjà certaines applications méritent d'être développées dans le cadre d'une utilisation généralisée dans l'ensemble des disciplines :

- la recherche documentaire sur la toile concerne aussi les mathématiques : c'est particulièrement le cas dans le cadre de la pédagogie de projet au collège et au lycée. De plus de nombreux sites (académiques ou autres) proposent des exercices, des tests, des énigmes parfois sous forme de concours ;
- l'utilisation de logiciels en ligne commence à être proposée grâce au développement de versions Java ou Active X de certains logiciels (Cabri, Geoplan, Geospace) ;
- le courrier électronique permet des échanges personnalisés entre élèves ou entre le professeur et des élèves. Il peut être aussi le prétexte à des exercices spécifiques (description de figure, mise en forme de démonstration, passage d'un langage codé au langage courant, etc).

En somme, la réalisation de ce travail nous a permis de s'initier à la recherche, de développer notre capacité à exploiter des articles scientifiques, d'utiliser la notion de probabilité conditionnelle (enseignée au lycée) dans le modèle PDMP et de s'imprégner à l'usage des nouvelles technologiques de l'information et de la communication. Enfin nous pensons que le modèle de gestion à travers les PDMP peut être appliqué dans les institutions éducatives et nous envisageons y faire des implémentations dès lors que les données seront disponibles.

♣ Conclusion générale et perspectives ♣

Nous pouvons retenir qu'un PDMP est un formalisme puissant pour représenter les problèmes séquentiels et stochastiques, afin de pouvoir le résoudre et trouver une politique optimale. Cependant, selon l'approche par la fonction de valeur, basée sur l'équation de Bellman pour le critère fini, une telle politique existe si la relation de préférence sur X est complète et la loi $\circ$ est additive . Par contre, selon l'approche par les loteries, nous avons proposé des propriétés simples et suffisantes sur la relation de préférence sur celles-ci garantissant l'admissibilité de la recherche arrière. Il en résulte que l'invariance par translation, la transitivité, la complétude, l'indépendance de la relation de préférence sur les loteries, la stabilité de la relation de préférence sur les politiques sont des conditions d'existence d'une politique optimale. Dans la pratique, les algorithmes généraux sont directement utilisés en programmation dynamique.

Étant donné que nous avons énoncé nos résultats dans un cadre probabiliste, ils pourraient probablement être transposés au cadre possibiliste ou à d'autres types d'incertain. Le cadre des processus décisionnels de Markov s'est largement imposé comme modèle pour la planification dans l'incertain en intelligence artificielle ces dernières années. Néanmoins, ce modèle présente certaines insuffisances pour résoudre certains problèmes de planification, par exemple :

- l'hypothèse d'observabilité complète de l'état du monde à chaque instant.
- l'hypothèse de connaissance parfaite du modèle (transitions, récompenses).

Pour pallier ces différentes limitations, plusieurs pistes ont été suivies. Il sera envisageable de :

- redéfinir et résoudre l'équation d'optimalité dans le cadre des MDP non classiques,
- modéliser les systèmes qui s'exécutent à l'horizon infini (par exemple le fonctionnement des feux rouges, l'exploitation des gisements fossiles) ?.

♣ Bibliographie ♣

[1] Bellman R.E. (1957). *Dynamic Programming*. Princeton University Press.

[2] Bonet B., Pearl J. (2002). *Qualitative MDPs and POMDPs : An order-of-magnitude approximation*. UAI vol.18. P.61 - 68.

[3] Cavados-Cadenas R., de Oca R.M. (2000). *Nearly optimal policies in risk-sensitive positive dynamic programming on discrete spaces*. Mathematical Methods of Operations Research, vol.52, p.133-167.

[4] Dubois D., Prade H. (1995). *Possibility Theory as a basis of Qualitative Decision Theory*. IJCAI. vol.14, p.1925-1930.

[5] Fishburn P.(1970). *Utility theory for decision making*. Wiley.

[6] Puterman M.L (1994). *Markovien Decision processes - Discrete stochastique programming*. Wiley-Interscience. New York.

[7] Sabbadin Regis (2009). *Modèles et algorithmes pour la décision séquentielle dans l'incertain*. HDR en intelligence artificielle. Université Paul SABATIER De Toulouse. 156 pages.

[8] Sabbadin R., Fargier H., Lang J (1998). *Towards qualitative approaches to multi-stage decision making*. International journal of Approximate Reasoning. vol.19. p.441 - 471.

[9] Sobel M.(1975). *Ordinal dynamic programming*. Management science. vol.21 p.967 - 975.

[10] Viet A.-F., Jean Pierre L., Bouzid M., Mouaddib A. -I. (2013). *Utilisation des processus décisionnels de Markov pour l'aide à la maîtrise d'une maladie animale*. Revue d'Intelligence Artificielle 27. 471 - 492. DOI :10.3166/ria.27.471- 492.

[11] Von Neumann J., Morgenstern O. (1994). *Theory of games and economic behavior*. Princeton university press.

Weng Paul (2006). *Processus de décision markoviens et préférences non classiques*. Revue d'intelligence artificielle. Volume $20 - n°2 - 3$. 22 pages.