Stratégies optimistes en apprentissage par renforcement

Sarah Filippi

Thèse

présentée pour obtenir le grade de docteur

de Télécom ParisTech

Spécialité : Signal et Images

Sarah Filippi

Stratégies optimistes en apprentissage par renforcement

Soutenue le 24 novembre 2010 devant le jury composé de

Jean-Yves Audibert	Rapporteurs
Rémi Munos	
Damien Ernst	Examinateurs
Frédérick Garcia	
Eric Moulines	
Fabrice Clérot	
Olivier Cappé	Directeurs de thèse
Aurélien Garivier	

Table des matières

Remerciements

Voici venue la fin de trois années de thèses, années très enrichissantes scientifiquement et également riches en rencontres. Une thèse est un travail solitaire en bien des points mais je n'aurais jamais pu en voir le bout sans un grand nombre de personnes.

Je remercie en premier lieu mes directeurs de thèse Olivier Cappé et Aurélien Garivier qui m'ont apporté en plus de leur compétences scientifiques, une disponibilité permanente et un encadrement attentionné tout au long de ces années. Ce fut un réel plaisir de travailler avec eux et de bénéficier de leurs expériences dans la réflexion scientifique et aussi de leur conseil dans la rédaction d'articles et la programmation orientée objet en Matlab. Si les compétences d'encadrement d'Olivier ont déjà été fort appréciées par d'autres de ses anciens doctorants, en particulier son goût pour le travail bien fait et ses relectures assidues, j'ai eu la très grande chance d'être la première doctorante encadrée par Aurélien que je remercie, entre autres, d'avoir parfaitement su m'encourager à chaque étape de cette thèse, qualité précieuse chez un directeur de thèse.

Je suis également très reconnaissante à Eric Moulines qui a été mon directeur de thèse initial. Je le remercie d'avoir cru en mon potentiel de chercheuse, de m'avoir convaincu de me lancer dans cette aventure et d'avoir veillé à ce que les trois années de celle-ci se déroulent au mieux pour moi. Je garderai en mémoire les après-midis passés ensemble à découvrir les processus de décision Markovien en disséquant le livre de Putterman, à formaliser les mathématiques derrière les états de croyance, et à écrire mon premier article.

Je tiens à remercier Orange Labs et en particulier Fabrice Clérot d'avoir motivé et financé cette thèse. Merci, Fabrice, pour le rôle tout particulier que tu as eu dans ma thèse : à l'origine de son sujet, tu as suivi régulièrement nos avancées en apportant des idées nouvelles et pertinentes à chacune de nos rencontres.

I would especially like to thank Csaba Szepesvári whom I have been honoured to work with. Csaba invited me to spend two months at the University of Alberta (Canada) in Autumn 2009. I discovered during my stay in Edmonton and my collaboration with Csaba another research environment. I absolutely loved the times at the whiteboard together ! I also had the opportunity to be next to great specialists in reinforcement learning. Thanks to Csaba, Rich Sutton and all the RLAI team for their exceptional welcome.

Je suis très honorée que Jean-Yves Audibert et Rémi Munos aient accepté d'être rapporteurs de ma thèse. Merci sincèrement du temps et de l'énergie que vous avez consacrés à la lecture si détaillée de mon travail. Je tiens à remercier également les examinateurs, Damien Ernst, Frédérick Garcia et Fabrice Clérot, de m'avoir fait l'honneur de venir évaluer mon travail lors de ma soutenance. Merci de votre disponibilité, de votre intérêt, et de vos questions ouvrant de nouvelles perspectives.

J'adresse également toute mon amitié et ma reconnaissance à tous les membres du groupe STA du département TSI du laboratoire LTCI qui ont rendu ces années de thèse aussi agréables qu'enrichissantes. Merci en particulier à Zaid, Julien, Tabea, Nataliya, Steffen, Jimmy, Jean-François, Marine, Olaf, Anne-Laure, Malika, Alexandre, Joffrey, Sylvain, Charanpal, Emilie, Sylvain, Amandine, Onur, David (par ordre chronologique plus ou moins fiable, mes excuses par avance pour les inévitables oublis) pour votre gaieté, complicité, soutien psychologique infaillible, amitiés toujours au rendez-vous et qui dureront au-delà de cette thèse, discussions mathématiques sur le tableau et/ou autour d'un thé en DA316, votre aide en informatique et la relecture si scrupuleuse des chapitres de ma thèse.

Je n'aurais pas été amenée à faire cette thèse si Bertrand David n'avait accepté en 2006 d'encadrer mon stage de DEA au sien du groupe AUDIO du même département. Merci à toi, Bertrand, à toute l'équipe AUDIO, ainsi qu'à Teodora, Christophe, Nancy, Valentin, Chloé, Jean-Louis, Cyril, Jean, Cyril et Aurélia de m'avoir fait découvrir un environnement de recherche si convivial. Merci également à Fabrice et Sophie-Charlotte pour l'aide informatique qui m'a souvent parue magique et à Clara pour son sourire accueillant dès l'entrée de la rue Dareau.

Cette liste de remerciements ne serait pas complète sans ceux que j'adresse à tous mes amis, dont les Balkansamblistes, et à ma famille, qui m'ont permis de garder un œil en dehors des galères scientifiques, m'évader dans la musique, et qui m'ont soutenu dans les moments les plus difficiles. Merci beaucoup.

Résumé

Cette thèse traite de méthodes « model-based » pour résoudre des problèmes d'apprentissage par renforcement. On considère un agent confronté à une suite de décisions et un environnement dont l'état varie selon les décisions prises par l'agent. Ce dernier reçoit tout au long de l'interaction des récompenses qui dépendent à la fois de l'action prise et de l'état de l'environnement. L'agent ne connaît pas le modèle d'interaction et a pour but de maximiser la somme des récompenses reçues à long terme. Nous considérons différents modèles d'interactions : les processus de décisions markoviens, les processus de décisions markoviens partiellement observés et les modèles de bandits. Pour ces différents modèles, nous proposons des algorithmes qui consistent à construire à chaque instant un ensemble de modèles permettant d'expliquer au mieux l'interaction entre l'agent et l'environnement. Les méthodes dites « model-based » que nous élaborons se veulent performantes tant en pratique que d'un point de vue théorique. La performance théorique des algorithmes est calculée en terme de regret qui mesure la différence entre la somme des récompenses reçues par un agent qui connaîtrait à l'avance le modèle d'interaction et celle des récompenses cumulées par l'algorithme. En particulier, ces algorithmes garantissent un bon équilibre entre l'acquisition de nouvelles connaissances sur la réaction de l'environnement (exploration) et le choix d'actions qui semblent mener à de fortes récompenses (exploitation). Nous proposons deux types de méthodes différentes pour contrôler ce compromis entre exploration et exploitation.

Le premier algorithme proposé dans cette thèse consiste à suivre successivement une stratégie d'exploration, durant laquelle le modèle d'interaction est estimé, puis une stratégie d'exploitation. La durée de la phase d'exploration est contrôlée de manière adaptative ce qui permet d'obtenir un regret logarithmique dans un processus de décision markovien paramétrique même si l'état de l'environnement n'est que partiellement observé. Ce type de modèle est motivé par une application d'intérêt en radio cognitive qu'est l'accès opportuniste à un réseau de communication par un utilisateur secondaire.

Les deux autres algorithmes proposés suivent des stratégies optimistes : l'agent choisit les actions optimales pour le meilleur des modèles possibles parmi l'ensemble des modèles vraisemblables. Nous construisons et analysons un tel algorithme pour un modèle de bandit paramétrique dans un cas de modèles linéaires généralisés permettant ainsi de considérer des applications telles que la gestion de publicité sur internet. Nous proposons également d'utiliser la divergence de Kullback-Leibler pour la construction de l'ensemble des modèles vraisemblables dans des algorithmes optimistes pour des processus de décision markoviens à espaces d'états et d'actions finis. L'utilisation de cette métrique améliore significativement le comportement de des algorithmes optimistes en pratique. De plus, une analyse du regret de chacun des algorithmes permet de garantir des performances théoriques similaires aux

meilleurs algorithmes de l'état de l'art.

Abstract

This thesis concerns « model-based » methods to solve reinforcement learning problems : an agent interacts with an environment by sequentially choosing actions that effect the state of the environment. The agent receives at each time point a reward which depends on the action and on the state of the environment. The aim of the agent is to maximise the cumulative rewards without knowing the model of interaction. We consider different models of interaction : Markov decision processes, partially observed Markov decision processes and bandit models. For each of these models, we provide « model-based » algorithms : These methods define a set of models which could explain the interaction between an agent and an environnement. To analyze the performance of our algorithm we study their regret which is the difference between the cumulative reward received by an agent who knows the model of the interaction and the ones received following the algorithm. Moreover, our novel algorithms perform well in practice. In particular, they guarantee a good balance of the well-known compromise between exploration and exploitation.

The first algorithm proposed in this thesis consists of following an exploration policy during which the model is estimated and then an exploitation one. The duration of the exploration phase is controled in an adaptative way. We then obtain a logarithmic regret for a parametric Markov decision problem even if the state is partially observed. This type of model is motivated by an application of interest in cognitive radio : the opportunistic access of a communication network by a secondary user.

The two other novel algorithms are optimistic ones : the agent chooses the optimal actions for the best possible model amongst a set of likely models. We construct and analyse such an algorithm in a parametric bandit model for a generalized linear model. We consider an on-line advertisement application. We then use the Kullback-Leibler divergence to construct the set of likely models in optimisic algorithms for finite Markov decision processes. This change in metric is studied in details and leads to significant improvement in practice. A theoretic analysis of the regret of those algorithms is also provided.

Notation

Notations de base

v'	Transposée du vecteur colonne v
M'	Transposée de la matrice M
$\det(M)$	Déterminant de la matrice M
I	Matrice identité
$C_{0:t}$	La suite $C_0, C_1, \ldots, C_t$
$\mathbf{e}$	Vecteur tel que, pour tout i, $\mathbf{e}(i) = 1$
$\mathbf{e}_i$	Vecteur tel que $\mathbf{e}_i(i) = 1$ et, pour tout $j \neq i$, $\mathbf{e}_i(j) = 0$

Probabilité

$\mathbb{E}_\theta^\pi$	Espérance sous une politique π et sous un paramètre θ
$\mathbb{P}_\theta^\pi$	Loi de probabilité sous une politique π et sous un paramètre θ
$\mathbb{S}^n$	Simplexe de probabilité de dimension $n-1$

Normes et distances

Soit v un vecteur de $\mathbb{R}^n$.

$\|.\|_1$	La norme L^1 : $\|v\|_1 = \sum_{i=1}^n v_i$
$\|.\|_2$	La norme L^2 : $\|v\|_2 = \sqrt{\sum_{i=1}^n v_i^2}$
$\|.\|_\infty$	La norme L^∞ : $\|v\|_\infty = \max_i v_i$
$\|.\|_M$	Pour toute matrice carrée M de taille n telle que pour tout vecteur v, $v'Mv \geq 0$, $\|v\|_M = \sqrt{v'Mv}$

$KL(.;.)$ La divergence de Kullback-Leibler entre deux vecteurs p et q de $\mathbb{S}^n$:

$$KL(p;q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}$$

$kl(.;.)$ La divergence de Kullback-Leibler entre deux lois de Bernoulli de paramètres p et q :

$$kl(p;q) = \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$$

Processus de Décision Markovien (Partiellement observés)

$\mathcal{X}$	Espace d'état
$\mathcal{A}$	Espace d'action
$\mathcal{Y}$	Espace des observations
t	temps courant
X_t	Etat à l'instant t
A_t	Action à l'instant t
Y_t	Observation à l'instant t
R_t	Récompense à l'instant t
I_t	Etat interne à l'instant t
$P(x,a;x')$	Probabilité de transition de l'état $x \in \mathcal{X}$ à l'état $x' \in \mathcal{X}$ conditionnellement à l'action $a \in \mathcal{A}$
$\mathcal{R}(x,a;.)$	Loi de probabilité de la récompense reçue dans l'état est x si l'action est a
$r(x,a)$	Espérance de la récompense reçue lorsque l'état est x et l'action est a
$G(x,a;y)$	Probabilité d'observer $y \in \mathcal{Y}$ lorsque l'état est x et l'action est a
$\mathbf{M}$	Un processus de décision Markovien $\mathbf{M} = (\mathcal{X}, \mathcal{A}, r, P)$
$R_{\max}$	Récompense maximale
π	Politique
π^*	Politique optimale
V_N^π , V_N^*	Fonction de valeur sous la politique π (resp. optimale) pour le critère fini
V_γ^π , V_γ^*	Fonction de valeur sous la politique π (resp. optimale) pour le critère γ-pondéré
γ	Facteur de pondération
η^π , η^*	Fonction de valeur sous la politique π (resp. optimale) pour le critère moyen
h^π . h^*	Un vecteur de biais sous la politique π (resp. optimale)

r^π	Vecteur de l'espérance de la récompense sous la politique π : $r^\pi(x) = r(x, \pi(x))$
P^π	Matrice de transition entre les états sous la politique π : $P^\pi(x, y) = P(x, \pi(x); y)$
$\hat{r}_t(.,.)$	Estimation des récompenses espérées à l'instant t
$\hat{P}_t(.,.;.)$	Estimation du noyau de transision à l'instant t
$\mathcal{L}_\gamma^\pi$, $\mathcal{L}_\gamma$	Opérateurs de Bellman pour le critère γ-pondéré
$\mathcal{L}^\pi$, $\mathcal{L}$	Opérateurs de Bellman pour le critère moyen

Abbréviations

MDP	Processus de décision Markovien
POMDP	Processus de décision Markovien Partiellement observé
KL	Kullback-Leibler

Introduction

Cette thèse a été effectuée au sein du *Laboratoire Traitement et Communication de l'Information* (LTCI), une Unité Mixte de Recherche du *CNRS* et de *Télécom ParisTech*. Elle a été financée par *Orange Labs* dans le cadre d'un contrat de recherche externe portant sur des travaux en apprentissage par renforcement.

L'*apprentissage par renforcement* est un domaine de l'*apprentissage automatique* (*machine learning* en anglais) permettant de résoudre des *problèmes de décisions séquentielles dans l'incertain*. On considère un *agent* (on encore, sujet, acteur, décideur) et un *environnement* (ou encore système). L'agent est confronté à une suite de prises de décisions. L'environnement, quant à lui, est caractérisé par un *état* qui évolue dans le temps de manière aléatoire en fonction des décisions prises par l'agent. Une prise de décision consiste en un choix d'*action*. Ce choix dépend lui-même de l'état actuel de l'environnement. Ainsi, la décision prise à chaque instant induit un changement d'état et, par conséquent, influence la prise de décision suivante. Le déroulement dans le temps de ces événements successifs, qui caractérisent l'interaction entre l'agent et l'environnement, est essentiel.

Cette dénomination d'apprentissage par *renforcement* trouve son origine dans le fait qu'à chaque instant l'agent reçoit des *récompenses* (ou encore bénéfices, gains) selon l'état de l'environnement et l'action qu'il a choisie. Ces récompenses le guident dans ses prises de décisions. Ainsi, le but de l'agent est de choisir ses actions de manière à maximiser les récompenses reçues. En général, l'agent a pour objectif de maximiser, non pas les récompenses reçues à chaque instant, mais celles reçues sur le long terme. L'action permettant de maximiser la récompense immédiate étant donné l'état de l'environnement est a priori différente de celle aboutissant à un grand bénéfice à long terme. En effet, la récompense reçue par l'agent dépendant à la fois de l'action choisie et de l'état de l'environnement, il peut être nécessaire de choisir des actions peu rentables à court terme afin d'atteindre un état dans lequel la récompense est très grande.

L'apprentissage par renforcement est utilisé dans de nombreux domaines comme la robotique, la théorie des jeux, l'économie, la gestion de capteur, les communications numériques, etc ... Citons deux applications d'intérêt pour des entreprises des Nouvelles Technologies de l'Information et de la Communication (NTIC) comme *Orange*, l'une dans le domaine des télécommunications sans fil et l'autre en marketing.

La première application concerne l'accès opportuniste aux ressources spectrales pour les radios cognitives. Dans la radio à bandes licenciées, les ressources spectrales sont divisées en bandes de fréquences attribuées par licence à des utilisateurs dits primaires. Chaque bande de fréquence est associée à un utilisateur fixe et peut donc être occupée ou non par l'utilisateur.

Le nombre d'utilisateurs des ressources spectrales étant en constante croissance ces dernières années, un besoin de nouveau système d'allocation des ressources se fait ressentir. L'idée de la radio cognitive est celle d'un système de communication intelligent qui détecterait les besoins des utilisateurs et fournirait les ressources radio et les services sans fil les plus appropriés en fonction des ressources disponibles. Le but est de partager les bandes de fréquences attribuées à des utilisateurs primaires avec des utilisateurs qui ne possèdent pas de licence. Ces utilisateurs dits secondaires identifient prudemment les ressources spectrales disponibles afin de communiquer en évitant de perturber le réseau primaire. Dans cette situation, l'environnement est le réseau primaire et l'état de l'environnement désigne la disponibilité des ressources. L'agent est un utilisateur secondaire et les récompenses sont proportionnelles aux nombres de transmissions qu'il a pu effectuer. Pour cause de limitations techniques et étant donné le coût énergétique de la surveillance du spectre, l'utilisateur secondaire ne peut pas observer l'état de toutes les ressources spectrales simultanément. Dans ce modèle, la principale décision de l'agent est de choisir quelles bandes de fréquences écouter afin de déterminer si ces ressources spectrales sont actuellement utilisées ou non par les utilisateurs primaires et de pouvoir ensuite maximiser son débit de transmission.

Une application qui a suscité un grand intérêt ces dernières années concerne l'optimisation des ressources publicitaires sur internet. On considère un gestionnaire de site qui cherche à déterminer à chaque instant l'annonce publicitaire à afficher sur une des pages de son site internet. Dans ce modèle, l'état de l'environnement peut être vu comme étant la page demandée par un utilisateur. L'agent est le gestionnaire de site et l'action est l'annonce publicitaire qu'il choisit d'afficher. La récompense est la réaction binaire du visiteur qui clique ou non sur l'annonce. La décision du choix de l'annonce publicitaire n'a aucune influence sur la page demandée à l'instant suivant mais permet la réception d'une récompense qui diffère selon la publicité choisie par le gestionnaire de site. Ainsi, dans ce modèle, la seule conséquence de l'action concerne la récompense reçue. Cela correspond à un *modèle de bandit*. Pour chaque annonce publicitaire, l'agent dispose de certaines caractéristiques des annonces. Cette information peut être, par exemple, un thème sémantique : « sport », « cinéma », « informatique », « voyage », etc... Le taux moyen de clic d'une annonce est supposé être lié à cette information. Ainsi deux annonces publicitaires d'une même catégorie ont vraisemblablement le même taux de clic.

Les *processus de décision markoviens* (*Markov Decision Process* -MDP- en anglais) permettent de modéliser l'interaction entre l'agent et l'environnement. Cette modélisation suppose qu'à chaque instant, l'état de l'environnement ne dépend des états et des actions passés qu'à travers l'état et l'action à l'instant précédent. Cette hypothèse markovienne de l'évolution des états convient à un grand nombre d'applications. Le *modèle d'interaction* est ainsi déterminé par la loi des récompenses reçues et le modèle d'évolution des états. Dans certaines applications, comme, par exemple, l'accès opportuniste aux ressources spectrales pour les radios cognitives présenté ci-dessus, l'agent ne peut pas observer entièrement l'état de l'environnement à chaque instant. Il prend alors ses décisions en fonction de l'observation partielle qu'il perçoit. Ces situations sont modélisées par des *processus de décisions markoviens* dits *partiellement observés* (*Partially Observed Markov Decision Process* -POMDP- en anglais).

Dans cette thèse, nous nous intéressons à la fois à des *processus de décisions markoviens*, à des *processus de décisions markoviens partiellement observés* et à des *modèles de bandits*. Pour ces différents modèles, nous proposons des méthodes d'apprentissage par renforcement qui se veulent performantes tant en pratique que d'un point de vue théorique. Nous nous sommes intéressés à des modèles liés aux deux applications évoquées ci-dessus dans les chapitres 2 et 3 de cette thèse.

Pour choisir ses actions dans le but de maximiser ses récompenses reçues à long terme, l'agent suit ce qui est appelé une *politique* (ou stratégie) *optimale*. Lorsque le modèle d'interaction est connu de l'agent, la recherche d'une politique optimale consiste à résoudre la *tâche de planification*. Dans un problème d'apprentissage par renforcement, l'agent ne connaît initialement pas les conséquences de ses décisions sur l'évolution des états ni sur la distribution des récompenses reçues. Il doit donc les apprendre tout au long de l'interaction avec l'environnement afin de choisir au mieux les décisions à prendre. Cet apprentissage nécessite d'essayer un certain nombre de fois les différentes actions possibles dans chacun des états de l'environnement. C'est ce que l'on appelle l'*exploration*. Ce terme d'exploration est souvent opposé à l'*exploitation* qui consiste à prendre des décisions dans le seul but de maximiser les récompenses reçues. A chaque instant, l'agent est donc confronté au dilemme entre sélectionner ses actions afin d'acquérir de nouvelles connaissances sur la réaction de l'environnement ou choisir celles qui lui semblent mener à de fortes récompenses. Un algorithme d'apprentissage par renforcement se doit de gérer ce *compromis entre exploration et exploitation*.

Il existe deux grandes familles d'algorithmes d'apprentissage par renforcement. Les méthodes dites *« model-free »* et *« model-based »* diffèrent selon que l'agent utilise ou non une estimation explicite du modèle d'interaction. Dans les méthodes « model-free », l'agent apprend directement les conséquences de ses actions en terme de récompenses reçues tandis que, dans les méthodes dites « model-based », l'agent estime le modèle tout au long de l'interaction avec l'environnement et utilise des algorithmes de *planification* pour déterminer la stratégie à suivre.

Nous nous intéressons dans cette thèse à des approches « model-based ». C'est donc en utilisant le modèle estimé que l'agent prend ses décisions à chaque instant. Dans ce type d'approche, l'exploration permet l'estimation du modèle d'interaction. Le compromis entre exploration et exploitation est alors manifeste : plus l'agent choisit d'explorer, moins il reçoit de récompenses mais plus grandes sont ses garanties sur le modèle estimé. De plus, si le modèle est suffisamment bien estimé, l'agent peut déterminer de manière précise les actions à choisir pour maximiser les récompenses. Les instants dédiés à l'exploration et à l'exploitation peuvent être dissociés en deux phases successives ou répartis tout au long de l'interaction, voire même indiscernables. Dans ce dernier cas, la stratégie de l'algorithme induit une exploration de tous les couples état-action mais celle-ci n'est pas explicite. Lorsque les deux phases sont successives, toute la difficulté d'un algorithme d'apprentissage par renforcement est de déterminer la longueur de la phase d'exploration, qui est bien évidemment la première des deux phases. Un des algorithmes que nous proposons permet d'adapter cette longueur au cours de la phase d'exploration, en décidant à chaque instant si suffisamment d'informations sur le modèle ont été accumulées pour passer à la phase d'exploitation.

Un principe « model-based » connu pour équilibrer exploration et exploitation de manière à assurer de bonnes performances théoriques consiste à *être optimiste face à l'incertain*. Cette belle philosophie de vie a été introduite en apprentissage par renforcement pour des modèles de bandit puis étendue pour des processus de décision markoviens plusieurs décennies plus tard. Il s'agit de répertorier, à chaque instant, l'ensemble des modèles d'interaction compatibles avec les informations accumulées durant l'interaction : la suite des états observés, actions prises et récompenses reçues. La récompense à long terme reçue si l'agent se trouvait dans chacun de ces modèles possibles est ensuite calculée. L'algorithme optimiste sélectionne le modèle qui lui permettrait de recevoir la plus grande des récompenses. Ce modèle est appelé le *modèle optimiste*. L'agent choisit alors ses actions comme s'il se trouvait dans ce modèle. Cette approche permet d'explorer suffisamment tout en maximisant les récompenses reçues. Notons qu'ici l'exploration a lieu tout au long de l'interaction et non pas seulement durant une phase initiale. La taille de l'ensemble des modèles considérés à chaque instant est contrôlée par ce que nous appellerons un *bonus d'exploration*. Nous proposons deux algorithmes optimistes,

l'un pour un modèle de bandit paramétrique (chapitre 3) et l'autre dans un MDP à espaces d'états et d'actions finis (chapitre 4).

Au delà des performances pratiques, nous nous intéressons aux garanties théoriques associées aux algorithmes proposés. Il existe différentes mesures théoriques de la performance d'un algorithme d'apprentissage par renforcement. Celle qui nous semble la plus intéressante prend en compte les récompenses reçues tout au long de l'interaction, que ce soit lorsque l'agent suit une stratégie d'exploration ou d'exploitation. Il s'agit du *regret cumulé* défini comme étant la différence entre les récompenses en suivant l'algorithme et celles reçues par un agent *oracle*, qui connaîtrait à l'avance les meilleures décisions à prendre. Nous cherchons en particulier à fournir des bornes supérieures du regret en temps fini (non asymptotiques) pour chacun des algorithmes d'apprentissage par renforcement que nous proposons.

Ce document est divisé en quatre chapitres, les trois derniers exposent les contributions de cette thèse.

Le premier chapitre introduit les notions mathématiques utiles pour la lecture de cette thèse. Le formalisme des processus de décisions markoviens est présenté et les notions de politiques (ou stratégies) de décision ainsi que les différents critères de maximisation des récompenses couramment utilisés sont définis. Après avoir exposé des algorithmes courants de planification, nous décrivons les deux grandes familles d'algorithmes d'apprentissage par renforcement que sont les algorithmes *« model-free »* et *« model-based »*, et citons les approches importantes pour chacune d'entre elles. Nous présentons ensuite les processus de décisions markoviens partiellement observés dans lesquels l'agent n'observe pas l'état de l'environnement mais perçoit une observation aléatoire qui dépend de ce dernier. Des méthodes permettant de résoudre la tâche de planification ainsi que l'apprentissage par renforcement dans ce modèle plus complexe sont également exposées.

Le deuxième chapitre traite d'un POMDP particulier, aussi appelé *« restless bandit »*, dans lequel l'agent peut choisir une action lui permettant d'observer une partie de l'état de l'environnement. L'accès opportuniste aux ressources spectrales pour les radios cognitives est un exemple typique de ce modèle. Dans un premier temps, nous proposons un algorithme pour déterminer une stratégie permettant de maximiser les récompenses reçues à long terme dans le cas où l'agent connaît le modèle d'interaction. Dans un deuxième temps, nous simplifions le modèle et fournissons, dans ce cas, un algorithme d'apprentissage par renforcement original appelé *algorithme de pavage* basé sur l'idée du découpage de l'espace des paramètres en différentes zones associées aux politiques optimales. Cet algorithme exploite le fait que le modèle simplifié ne dépend que d'un paramètre de dimension assez petite. Une analyse théorique du regret et des résultats numériques décrivent les performances de cet algorithme.

Dans le troisième chapitre, nous nous intéressons à des modèles de bandits paramétriques où le nombre d'actions est très grand et le paramètre est de dimension beaucoup plus petite. Nous supposons que l'agent dispose au préalable d'informations caractéristiques pour chacune des actions. Nous étendons les travaux menés par la communauté en considérant des modèles linéaires généralisés pour la loi des récompenses. Nous proposons un algorithme optimiste pour ce modèle et analysons ses performances en terme de regret. Le regret obtenu est similaire à ceux existant dans la littérature pour des modèles de récompenses linéaires. Il ne dépend pas du nombre d'actions possibles mais de la dimension de l'espace des paramètres. La plupart des algorithmes optimistes pour lesquels il existe des bornes de regret théoriques sont rarement accompagnés d'une étude de leur performance pratique. Nous proposons une nouvelle manière de régler le bonus d'exploration de l'algorithme et illustrons les performances de l'algorithme à la fois sur des données simulées et sur des données réelles concernant l'optimisation de publicités sur internet.

Le quatrième chapitre s'intéresse à un changement de métrique dans le calcul des modèles

optimistes. Les méthodes d'apprentissage par renforcement optimistes garantissant des bornes de regret non-asymptotiques utilisent des inégalités de concentration pour définir le modèle optimiste. Dans la plupart des algorithmes proposés dans la littérature, les bornes de déviation utilisées majorent l'écart entre le vrai modèle et le modèle estimé à l'aide de la métrique L^1. Nous proposons de remplacer cette métrique par l'utilisation de la divergence de Kullback-Leibler (KL), cette mesure de proximité étant plus adaptée pour les espaces de probabilité que la métrique L^1. Nous exploitons des inégalités de concentration récentes utilisant la divergence de KL et fournissons deux algorithmes optimistes utilisant cette mesure respectivement pour un problème de bandit à récompenses binaires et pour un processus de décision markovien à espace d'états et d'actions finis. Une analyse théorique garantit des regrets logarithmiques pour ces deux algorithmes et la performance pratique est illustrée sur plusieurs exemples de référence dans les MDP. Nous concluons sur une discussion concernant les principales différences entre l'utilisation de ces deux métriques qui explique la bonne performance du recours à la divergence de KL dans le calcul de modèles optimistes.

CHAPITRE 1

Processus de décision markoviens et apprentissage par renforcement

Les Processus de Décisions Markoviens (MDP) constituent un formalisme mathématique permettant de modéliser l'interaction entre un agent et son environnement. A la base de ce formalisme, se trouvent les concepts d'*état* qui décrit la situation dans laquelle se trouve l'environnement, d'*action* qui résume la décision prise par l'agent et de *récompense*. L'action influence l'évolution des états et permet la réception d'un gain (ou récompense) aléatoire qui dépend de l'action choisie et de l'évolution des états. Dans ce modèle, on admet que la transition entre les états est une fonction aléatoire de l'action courante qui ne dépend pas de la suite des états et des actions passés. La séquence des états forme alors une chaîne de Markov contrôlée par les actions d'où le nom de *processus de décision markovien*. Le graphe ci-dessous illustre cette interaction.

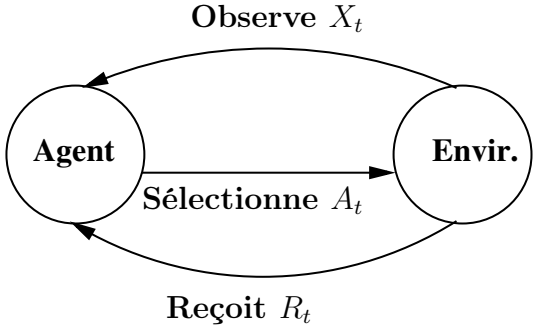

FIGURE 1.1 – Graphe de l'interaction entre un agent et un environnement du point de vue de l'agent.

Les récompenses reçues par l'agent le guident dans ses choix d'action. Ainsi, il prend ses décisions dans le but de maximiser ses récompenses à long terme. Différents critères de maximisation existent selon le problème étudié. L'agent peut viser à maximiser les récompenses reçues jusqu'à un instant fini fixé, ou sur un horizon infini durant lequel les récompenses sont potentiellement pondérées en fonction de l'instant où elles sont perçues. Chacun de ces critères de performance conduit à une stratégie de prise de décision différente.

En apprentissage par renforcement, l'agent ne dispose d'aucune information a priori sur le

modèle d'interaction, c'est-à-dire sur la loi des récompenses reçues et sur l'évolution des états de l'environnement conditionnellement aux décisions prises. Il ne connaît donc pas initialement les conséquences de ses actes tant au niveau des récompenses reçues qu'au niveau des transitions entre les états. Mais, avant de s'intéresser à ce cadre, il est important d'analyser comment l'agent doit agir pour maximiser ses récompenses lorsque il connaît le modèle de l'interaction. Cette question est évidemment centrale dans les problèmes de décisions dans l'incertain et sa solution est parfois loin d'être évidente. C'est ce que l'on appelle la tâche de *planification*.

Dans ce chapitre, nous présenterons initialement (section 1.1), le formalisme mathématique des MDP et définirons les notions de *politique*, soit une stratégie qui permet de choisir les actions à effectuer, et de *politique optimale* pour des critères de performances fixés. Dans un deuxième temps (section 1.2), nous nous intéresserons à résoudre le problème de *planification* pour des MDP à espaces d'états et d'actions finis. Nous présenterons, dans la troisième partie de ce chapitre (section 1.3), différentes approches possibles lorsque l'agent ne dispose d'aucune information sur le modèle d'interaction. La dernière section de ce chapitre (section 1.4) concernera une extension des MDP que sont les processus de décision markoviens partiellement observés (POMDP). Dans ces modèles, l'agent n'observe pas l'état de l'environnement mais perçoit une observation qui dépend de manière plus ou moins bruitée de cet état caché. Nous présenterons le modèle des POMDP et quelques méthodes de planification et d'apprentissage par renforcement dans ce contexte.

Ce chapitre se présente comme une introduction des notions nécessaires à la compréhension de la suite du déroulement de la thèse. On propose au lecteur qui souhaiterait en connaître plus sur ces sujets de consulter un des ouvrages suivants [Puterman, 1994; Sutton and Barto, 1998; Bertsekas, 1995] ou l'ouvrage francophone paru récemment [Sigaud and Buffet, 2008].

1.1 Processus de décision markoviens

Soit $t \in \mathbb{N}$ un *instant de décision*. On appelle X_t l'état du système (ou de l'environnement) à l'instant t et A_t la nouvelle action effectuée par l'agent décisionnel. Cette action engendre un changement de l'environnement : l'état devient alors X_{t+1}. Et de ce fait, l'agent reçoit une récompense R_t. Un processus de décision Markovien $\mathbf{M}$ est défini par le quintuplet $(\mathcal{X}, \mathcal{A}, P, \mathcal{R})$ suivant :

- $\mathcal{X}$ est l'espace d'état,
- $\mathcal{A}$ est l'espace des actions,
- P est le noyau de transition des états contrôlés par les actions,
- $\mathcal{R}$ est la loi des récompenses reçues.

Dans ce document, nous supposons que les espaces d'états $\mathcal{X}$ et d'actions $\mathcal{A}$ sont finis. Dans un processus de décision markovien, on suppose que l'état de l'environnement est une fonction aléatoire de l'état de l'environnement à l'instant précédent et de l'action effectuée. L'état de l'environnement X_{t+1} à l'instant $t+1$ vaut $x' \in \mathcal{X}$ avec probabilité

$$\mathbb{P}\left[X_{t+1} = x' \,\middle|\, X_t = x, A_t = a \right] \stackrel{\text{def}}{=} P(x, a; x') \,,$$

si, à l'instant t, l'état est x et l'action choisie est a. La chaîne des états $(X_t)_{t\geq 0}$ est une chaîne de Markov conditionnellement à la séquence des actions $(A_t)_{t\geq 0}$. Notons que le processus $(X_t)_{t\geq 0}$ n'est, en toute généralité, pas markovien. Tout dépend du choix des actions. Par exemple, si les actions sont choisies en fonction de l'histoire des actions et des états passés, alors $(X_t)_t$ n'est pas markovien tandis que, si le choix de l'action au temps t ne dépend que de l'état X_t alors le processus des états est markovien.

La récompense R_t, reçue à l'instant t, est un nombre réel qui dépend de l'état du système X_t et de l'action A_t de manière déterministe ou probabiliste. Pour tout sous-ensemble C de $\mathbb{R}$, on note

$$\mathcal{R}(x, a; C) \stackrel{\text{def}}{=} \mathbb{P}\left[R_t \in C \,|\, X_t = x, A_t = a\right]$$

la probabilité que la récompense reçue appartienne à C étant donné que l'état du système est x et l'action effectuée est a. Nous verrons dans la suite que l'espérance des récompenses reçues joue un rôle important. On définit la fonction $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ telle que, pour tout $x \in \mathcal{X}$, $a \in \mathcal{A}$:

$$\mathbb{E}\left[R_t \,|\, X_t = x, A_t = a\right] \stackrel{\text{def}}{=} r(x, a) \ .$$

Lorsque la récompense reçue dépend de manière déterministe de l'état et de l'action courants, on a $R_t = r(X_t, A_t)$. On supposera dans la suite que les récompenses R_t sont positives et bornées par $R_{\max}$.

Notons que les variables d'état X_t et d'action A_t sont des variables aléatoires discrètes définies sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ respectivement à valeurs dans $\mathcal{X}$ et $\mathcal{A}$. La récompense R_t est une variable aléatoire à valeur dans $\mathbb{R}$. Un *processus de décision markovien* (MDP) est alors un modèle stochastique tel que $(X_t)_{t\geq 0}$ est un processus de Markov contrôlé par le processus des actions $(A_t)_{t\geq 0}$. Plus précisément, il existe un noyau de transition P tel que pour tout $t \geq 0$, pour tout $x \in \mathcal{X}$,

$$\mathbb{P}\left[X_{t+1} = x \,|\, X_{0:t}, A_{0:t}\right] \stackrel{\text{def}}{=} P(X_t, A_t; x).$$

1.1.1 Exemples

Avant de poursuivre l'étude des MDP, donnons quelques exemples. Les processus de décision markoviens sont largement utilisés dans de nombreux domaines d'applications comme la robotique, les jeux, l'économie, la gestion de capteur, le domaine des communications numériques (voir entre autre [Sutton and Barto, 1998; Bertsekas, 1995; Bertsekas and Tsitsiklis, 1996; Feinberg and Shwartz, 2002; Hero et al., 2008] pour quelques exemples d'application). Nous proposons ici trois exemples simples qui permettront d'illustrer notre discours par la suite.

Rivière

Ce premier exemple est un modèle proposé par [Strehl and Littman, 2008] (appelé *River Swim*). Il s'agit d'un MDP composé de 6 états représentant une rivière dont le courant coule de

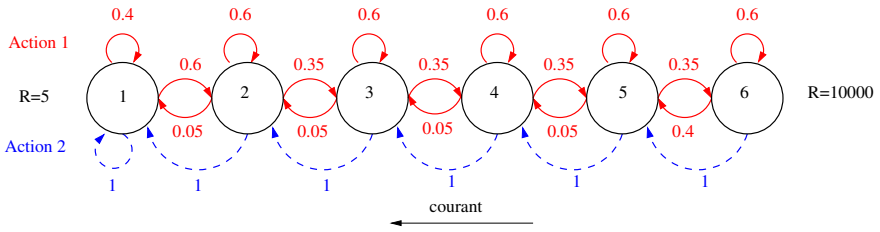

FIGURE 1.2 – Modèle de transition dans la *Rivière* : les flèches continues (resp. en pointillés) représentent les transitions si l'action 1 (resp. 2) a été choisie.

la droite vers la gauche. Les états 1 et 6 sont deux rives de cette rivière. L'agent est un nageur qui part du milieu de la rivière, assez près de la rive gauche. Dans chaque état, il peut nager soit vers la gauche soit vers la droite. L'espace $\mathcal{A}$ contient donc deux éléments et l'on note $A_t = 1$ si le nageur nage vers la droite et $A_t = 0$ s'il nage vers la gauche. Nager vers la droite

(en remontant le courant) permet d'atteindre l'état immédiatement à droite avec probabilité 0.35, laisse l'agent dans le même état avec une grande probabilité égale à 0.6, et peut même dévier l'agent vers la gauche avec une probabilité 0.05 (voir Figure 1.2). Au contraire, nager vers la gauche (avec le courant de la rivière) permet d'atteindre l'état immédiatement à gauche avec probabilité 1. L'agent reçoit une petite récompense quand il atteint la rive de gauche et une récompense beaucoup plus grande quand il atteint celle de droite. Lorsqu'il est dans la rivière, il ne reçoit aucune récompense. Dans cet exemple, l'hypothèse markovienne est assez intuitive : l'état à un instant donné dépend de l'endroit où se trouvait le nageur à l'instant précédent et de la direction vers laquelle il a nagé mais pas de ce qu'il a fait précédemment.

Gestion de stock

Ce deuxième exemple est un peu plus compliqué en terme de transition entre les états. Imaginons un vendeur de fruit qui a une capacité de stockage de 3 cageots de fruits. Chaque jour, la demande est de 0 à 3 cageots selon une probabilité uniforme sur $\{0, 1, 2, 3\}$. En début de journée, le vendeur peut commander entre zéro et 2 cageots qui arriveront le lendemain matin. De plus, les fruits peuvent pourrir (avec une probabilité p) auquel cas tous les fruits sont à jeter. On notera X_t le nombre de cageots en boutique le matin du t-ième jour et $A_t \in \{0, 1, 2\}$ le nombre de cageots achetés.

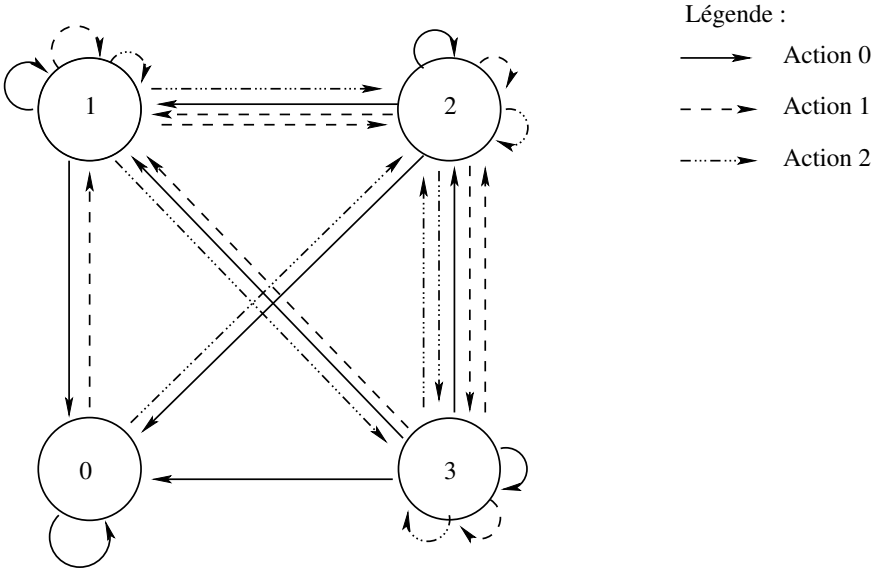

FIGURE 1.3 – Modèle de transition dans la *gestion de stock*.

Comme on peut l'observer sur la figure 1.3, le schéma de transition entre les états conditionnellement aux actions est plus compliqué que pour le modèle de la rivière. Supposons, par exemple, qu'il y ait deux cageots en magasin le matin du t-ième jour ($X_t = 2$) et que le vendeur achète un cageot ($A_t = 1$). Le lendemain matin, il y aura au moins un cageot dans le magasin -celui acheté la veille- donc $P(2, 1; 0) = 0$. De plus, la probabilité qu'il y ait 2 cageots en stock le lendemain matin, tout comme celle qu'il y en ait 3, est égale à la probabilité que les fruits n'aient pas pourri, et que respectivement 0 ou 1 cageot ait été vendu :

$$P(2, 1; 2) = P(2, 1; 3) = (1 - p)/4 \, .$$

Il ne reste le lendemain qu'un seul cageot soit si tous les fruits ont été vendus soit si ils ont tous pourri d'où $P(2, 1; 1) = \frac{1-p}{2} + p$. De manière générale, si il n'y a aucun cageot en stock

un matin ($x = 0$), alors il y en aura, le lendemain en début de journée, le nombre de cageots achetés

$$P(0, a; a) = 1 \quad \forall a \in \mathcal{A} .$$

S'il y a x cageots un matin et que le vendeur en achète a, alors le lendemain il y en aura entre a et $\max\{x + a, 3\}$. Les probabilités de transitions vérifient

$$P(x, a; a) = 1 - x\frac{1-p}{4}$$

$$P(x, a; x') = \begin{cases} \frac{1-p}{4} & \text{si } a < x' \leq x + a \leq 3 \\ \frac{1-p}{4} & \text{si } x + a > 3 \text{ et } a < x' < 3 \\ (x + a - 2)\frac{1-p}{4} & \text{si } x + a > 3 \text{ et } x' = 3 \\ 0 & \text{si } x' < a \end{cases}$$

La récompense moyenne reçue pour le jour t dépend du nombre de cageots vendus dans la t-ième journée et du nombre de cageots achetés. Supposons que chaque cageot vendu rapporte une récompense de 1, que l'achat d'un cageot coûte 0.6 et que la perte des fruits due à de la pourriture coûte 1 :

$$r(X_t, A_t) = X_t\frac{1-p}{4} - p - A_t \times 0.6 .$$

Modèle de bandit

Ce troisième exemple est un modèle important en apprentissage par renforcement auquel nous ferons très souvent référence. Il s'agit du *modèle de bandit* qui peut être formalisé comme un MDP avec un seul état et $|\mathcal{A}|$ actions différentes. Le nom de ce modèle vient des machines à sous du casino, aussi appelées bandits manchots : le joueur choisit un parmi $|\mathcal{A}|$ bras et reçoit une récompense aléatoire associée au bras. La moyenne de la récompense reçue est supposée être différente pour chacun des bras. Notons $r(a)$ la récompense moyenne reçue en jouant le bras a. Le but de l'agent est de détecter puis de jouer le meilleur bras, c'est-à-dire celui ayant la plus grande récompense moyenne :

$$a^* = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, r(a) .$$

Si l'agent (ou le joueur dans le casino) connaît la loi des récompenses associées à chacun des bras, en particulier, s'il connaît les récompenses moyennes, il suffit qu'il détermine le meilleur bras et le joue en permanence pour maximiser ses récompenses. Tandis que, en apprentissage par renforcement, l'agent ne connaissant pas ces lois, le jeu devient beaucoup plus compliqué : l'agent doit essayer différents bras et apprendre au plus vite quel est le meilleur.

1.1.2 Règles de décision et politiques

A chaque instant, l'agent est confronté à un problème de décision dans l'incertain qui consiste à sélectionner une parmi $|\mathcal{A}|$ actions différentes. La procédure suivie par l'agent pour sélectionner l'action à effectuer à un instant donné t est appelée *règle de décision* et est notée π_t. Celle-ci peut-être déterministe ou aléatoire. Le processus des états étant markovien conditionnellement aux actions, on considérera uniquement des règles de décision dites *markoviennes*, c'est-à-dire qui dépendent de l'état courant et pas de la séquence des états et actions passés. Une règle de décision est dite *déterministe et markovienne*, si π_t est une fonction qui à un état associe une action :

$$\pi_t : \mathcal{X} \to \mathcal{A} .$$

Elle est dite *aléatoire et markovienne*, si π_t est, conditionnellement à l'état, une mesure de probabilité sur l'espace des actions :

$$\pi_t : \mathcal{X} \times \mathcal{A} \to [0,1] \ .$$

Une *politique* est la séquence des règles de décision utilisées par l'agent à chaque instant de décision :

$$\pi = (\pi_t)_{t \geq 0} \ .$$

On appelle Π^D l'ensemble des politiques markoviennes déterministes, et Π^A celui des politiques markoviennes aléatoires. De la même manière, on note D^D (resp. D^A) l'ensemble des règles de décision markoviennes et déterministes (resp. markoviennes et aléatoires).

Dans une politique, les règles de décision sont a priori différentes à chaque instant. Une politique markovienne est dite *stationnaire* si la règle de décision est la même quelle que soit l'instant de décision c'est-à-dire si

$$\forall t \geq 0 \ , \pi_t = \pi_0 \ .$$

Dans ce cas, π désignera à la fois la politique et la règle de décision. Appelons Π^{SD} (resp. Π^{SA}) l'ensemble des politiques stationnaires déterministes (resp. aléatoires).

Les différentes classes de politiques vérifient les relations d'inclusion suivantes :

$$\begin{array}{ccc} \Pi^{SD} & \subset & \Pi^{SA} \\ \cap & & \cap \\ \Pi^{D} & \subset & \Pi^{A} \end{array}$$

1.1.3 Critère de performance, fonction de valeur

Le but de l'agent est de trouver, parmi une famille de politiques donnée, celle qui permet de maximiser les récompenses reçues. Différents critères sont couramment utilisés selon que l'on cherche à maximiser les récompenses reçues à un horizon temporel fini ou infini et que l'on souhaite atténuer ou non la contribution des récompenses associées à des actions faites à des instants éloignés de l'instant courant. Chacun de ces critères permet de définir une *fonction de valeur*. Étant donnée une politique π fixée, cette fonction associe, à tout état initial $x \in \mathcal{X}$, la valeur du critère considéré en suivant la politique π. Pour toute politique π, nous notons $V^\pi : \mathcal{X} \to \mathbb{R}$ la fonction de valeur associée. Nous présentons dans la suite trois critères de performance. Les politiques considérées dans les définitions ci-dessous peuvent être déterministes ou aléatoires.

Critère à horizon fini

Le problème décisionnel est dit être à *horizon temporel fini* n si l'interaction entre l'agent et l'environnement se termine à un instant $n < \infty$ fixé à l'avance. Dans ce cas, le critère de performance naturel est l'espérance de la somme des récompenses jusqu'à cet instant. Ainsi, on définit la fonction de valeur d'une politique $\pi = (\pi_t)_{t \geq 0}$ pour le *critère fini* :

$$\forall x \in \mathcal{X} \ , \quad V_n^\pi(x) \stackrel{\text{def}}{=} \mathbb{E}^\pi \left[\sum_{t=0}^{n-1} r(X_t, A_t) \,\middle|\, X_0 = x \right] \ ,$$

où l'on note $\mathbb{E}^\pi$ l'espérance lorsque l'agent suit la politique π.

Une variante de ce critère fini est le *critère du plus court chemin stochastique*. Il est utilisé dans des problèmes à horizon temporel aléatoire fini non borné dans lequel se trouve un état

absorbant de récompense nulle. Par convention, on appelle 0 cet état absorbant. Une fois que le système a atteint cet état, il y reste quelque soit l'action choisie : pour toute action $a \in \mathcal{A}$,

$$P(0, a; 0) = 1 \quad \text{et} \quad r(0, a) = 0 \, .$$

La fonction de valeur associée à ce critère est

$$\forall x \in \mathcal{X} \, , \quad V_C^\pi(x) \stackrel{\text{def}}{=} \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} r(X_t, A_t) \, \middle| \, X_0 = x \right] \, . \tag{1.1}$$

Ce critère est bien défini et dès lors que l'état absorbant 0 est atteignable avec un temps d'atteinte d'espérance finie quelque soit la politique jouée puisqu'il s'agit alors d'une somme sur un nombre fini de termes.

Critère γ-pondéré ou actualisé

Le *critère γ-pondéré* ou *actualisé* est le critère le plus courant pour des problèmes à horizon infini. Il s'agit de la somme des récompenses reçues, pondérées par un facteur $\gamma \in [0; 1[$. La fonction de valeur d'une politique $\pi = (\pi_t)_{t \geq 0}$ est alors :

$$\forall x \in \mathcal{X} \, , \quad V_\gamma^\pi(x) \stackrel{\text{def}}{=} \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(X_t, A_t) \, \middle| \, X_0 = x \right] \, .$$

Notons que le facteur de pondération γ permet notamment d'assurer la convergence de la série. La pondération des récompenses reçues par un facteur γ^t traduit le fait que les récompenses reçues à très long terme sont moins valorisées que les récompenses immédiates. Plus γ est proche de 1, plus les récompenses à long terme sont aussi importantes que celles à court terme.

La fonction de valeur V_γ^π est proportionnelle à V_C^π pour un modèle de plus court chemin stochastique avec un temps d'arrêt géométrique de paramètre $1 - \gamma$ et dans lequel la récompense reçue est toujours nulle sauf quand l'état absorbant est atteint [Toussaint and Storkey, 2006].

Critère moyen

Contrairement au critère γ-pondéré, le *critère moyen* attribue la même importance à toutes les récompenses, indépendamment de leur éloignement de l'instant courant. La fonction de valeur d'une politique $\pi = (\pi_t)_{t \geq 0}$ pour ce critère, appelée *fonction de valeur moyenne*, est alors :

$$\forall x \in \mathcal{X} \, , \quad \eta^\pi(x) \stackrel{\text{def}}{=} \lim_{n \to \infty} \frac{1}{n} \mathbb{E}^\pi \left[\sum_{t=0}^{n-1} r(X_t, A_t) \, \middle| \, X_0 = x \right] \, .$$

Dans un MDP à espaces d'état et d'action finis, cette limite est bien définie si la politique π est stationnaire (voir proposition 8.1.1 de [Puterman, 1994]). Ce critère est plus compliqué à analyser que le précédent. Nous parlerons dans la suite du lien entre le critère moyen et le critère γ-pondéré. On expliquera, de plus, que, sous certaines conditions sur la chaîne de Markov induite par la politique π, la fonction de valeur associée au critère moyen est indépendante de l'état de départ.

1.1.4 Politique optimale

L'agent a pour but de déterminer une politique qui lui permette de gagner le plus de récompense possible en maximisant un des critères présentés précédemment. On dit qu'une politique $\pi^* \in \Pi^A$ est *optimale* pour un des critères précédents si

$$\forall \pi \in \Pi^A,\ \forall x \in \mathcal{X}\ , \quad V^{\pi}(x) \leq V^{\pi^*}(x) \tag{1.2}$$

où V^π est la fonction de valeur de la politique π pour le critère choisi ($V^\pi = V_n^\pi$ ou V_γ^π ou η^π). La *fonction de valeur optimale* V^* est définie par :

$$\forall x \in \mathcal{X}\ , \quad V^*(x) = \sup_{\pi \in \Pi^A} V^\pi(x)\ . \tag{1.3}$$

Une politique $\pi^* \in \Pi^A$ optimale vérifie :

$$\forall x \in \mathcal{X}\ , \quad V^{\pi^*}(x) = V^*(x)\ .$$

Nous verrons dans la suite que, pour chacun des trois critères définis précédemment, et étant donné que les espaces d'états et d'actions sont finis, il existe une politique optimale déterministe markovienne. Pour les critères γ-pondérés et moyen, cette politique optimale est stationnaire.

1.1.5 Table état-action et politique gloutonne

Nous venons d'introduire la fonction de valeur liée à une politique pour des critères donnés ainsi que la fonction de valeur optimale. Une autre fonction jouant un rôle important en apprentissage par renforcement est la *table état-action*. Nous définissons dans ce paragraphe la table état-action ainsi que la politique gloutonne dans le cas du critère γ-pondéré ; pour le critère moyen, des définitions similaires peuvent être déduites, les liens entre la fonction de valeur et la table état-action étant semblables.

Pour une politique π fixée de fonction de valeur V_γ^π, définissons la *table état-action* Q_γ^π par :

$$\forall x \in \mathcal{X}\ ,\ \forall a \in \mathcal{A}\ ,\ Q_\gamma^\pi(x,a) = r(x,a) + \gamma \sum_{x' \in \mathcal{X}} P(x,a;x') V_\gamma^\pi(x)$$

C'est l'espérance de la somme pondérée des récompenses reçues à partir de l'état x si l'agent exécute tout d'abord l'action a puis suit la politique π :

$$Q_\gamma^\pi(x,a) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t r(X_t, A_t) \,\middle|\, X_0 = x, A_0 = a \right]$$

Notons Q_γ^* la *table état-action optimale* : $Q_\gamma^* = \max_\pi Q_\gamma^\pi$. Cette table et la fonction de valeur optimale sont liées par :

$$\forall x \in \mathcal{X}\ , \quad V_\gamma^*(x) = \max_a Q_\gamma^*(x,a)\ .$$

De plus, une politique optimale stationnaire déterministe peut être définie en fonction de Q_γ^* par

$$\forall x \in \mathcal{X}\ , \quad \pi^*(x) = \operatorname*{argmax}_a Q_\gamma^*(x,a).$$

On écrit aussi

$$Q_\gamma^*(x,a) = \mathbb{E}\left[r(X_t, A_t) + \gamma V_\gamma^*(X_{t+1}) \,\middle|\, X_t = x, A_t = a \right]\ .$$

Soit V une fonction de $\mathcal{X}$ dans $\mathbb{R}$ et Q la table état-action associée. On dit qu'une politique stationnaire π est *gloutonne par rapport à V (ou par rapport à Q)* si, pour tout $x \in \mathcal{X}$:

$$\begin{aligned}\pi(x) &\in \operatorname*{argmax}_{a\in\mathcal{A}} \mathbb{E}\left[r(X_t, A_t) + \gamma V(X_{t+1}) \,|\, X_t = x, A_t = a\right] \\ &= \operatorname*{argmax}_{a\in\mathcal{A}} \left\{ r(x,a) + \gamma \sum_{x'\in\mathcal{X}} P(x,a;x')V(x') \right\} \\ &= \operatorname*{argmax}_{a\in\mathcal{A}} Q(x,a) \ . \end{aligned} \tag{1.4}$$

Une politique gloutonne associe donc à chaque état x l'action qui maximise le vecteur $Q(x,.)$.

1.1.6 Notations vectorielles

Nous introduisons ici quelques notations que nous utiliserons régulièrement dans la suite de cette thèse.

Conditionnellement à toute politique markovienne aléatoire π, le processus des états $(X_t)_{t\geq 0}$ est une chaîne de Markov. L'état à l'instant $t+1$ ne dépend que de l'état à l'instant t : pour tous $x, x' \in \mathcal{X}$,

$$\begin{aligned}\mathbb{P}\left[X_{t+1} = x' \,\middle|\, X_t = x\right] &= \sum_{a\in\mathcal{A}} \mathbb{P}\left[X_{t+1} = x' \,\middle|\, X_t = x, A_t = a\right] \mathbb{P}\left[A_t = a \,|\, X_t = x\right] \\ &= \sum_{a\in\mathcal{A}} P(x,a;x')\pi_t(x,a) \ .\end{aligned}$$

Si π est une politique stationnaire aléatoire, notons P^π le noyau de transition de la chaîne de Markov $(X_t)_{t\geq 0}$. Pour tous états $x \in \mathcal{X}$ et $x' \in \mathcal{X}$:

$$P^\pi(x;x') \stackrel{\text{def}}{=} \sum_{a\in\mathcal{A}} \pi(x,a)P(x,a;x') \ .$$

De plus, on définit le vecteur r^π tel que, pour tous $x, x' \in \mathcal{X}$,

$$r^\pi(x) \stackrel{\text{def}}{=} \sum_{a\in\mathcal{A}} \pi(x,a)r(x,a) \ .$$

Pour une politique stationnaire déterministe, on définit de même pour tout $t \geq 0$ et pour tous $x, x' \in \mathcal{X}$

$$r^\pi(x) \stackrel{\text{def}}{=} r(x,\pi(x)) \quad \text{et} \quad P^\pi(x,x') \stackrel{\text{def}}{=} P(x,\pi(x);x') \ .$$

1.2 Planification dans des MDP discrets à horizon infini

Après avoir exposé le formalisme mathématique des processus de décision markoviens, nous présentons maintenant des méthodes permettant de déterminer une politique optimale pour un MDP donné. On suppose que l'agent connaît le modèle d'interaction, c'est-à-dire qu'il connaît l'espérance des récompenses $r(.,.)$ et les probabilités de transition $P(.,.;.)$. C'est ce que l'on appelle résoudre le problème de *planification*.

Revenons un instant aux exemples du paragraphe 1.1.1. Dans le cas du bandit, connaître le modèle signifie connaître l'espérance de la récompense de chacun des bras. La politique optimale dans ce cas est très simple : elle consiste à jouer à chaque instant le bras $a^* = \operatorname{argmax}_{a\in\mathcal{A}} r(a)$. Dans le modèle de la rivière, la récompense reçue lorsque le nageur se trouve sur la rive droite étant considérablement plus grande que celle reçue sur la rive gauche, si

l'horizon temporel est suffisamment grand, le but de l'agent est de rejoindre la rive droite pour obtenir un maximum de récompenses. Pour cela, il doit nager continuellement vers la droite : pour tout $x \in \mathcal{X}$, $\pi^*(x) = 1$. Dans l'exemple de la gestion de stock, il est beaucoup plus difficile de déterminer intuitivement la politique optimale. Il est donc nécessaire de connaître des méthodes génériques permettant de trouver la solution du problème de planification.

Dans un MDP à horizon fini, les politiques optimales sont a priori non stationnaires : les actions optimales ne sont pas les mêmes si l'on est proche de la fin de l'interaction ou s'il reste du temps pour essayer d'accumuler des récompenses. Par exemple, pour le modèle de la rivière, si il ne reste plus beaucoup de temps au nageur et qu'il se trouve près de la rive gauche, il a tout intérêt à aller nager jusqu'à la rive gauche pour gagner une petite récompense plutôt que de continuer à nager vers la rive droite sans espoir de recevoir une récompense. La politique optimale dans de tels problèmes à horizon fini peut alors être déterminée en calculant la fonction de valeur optimale à chaque pas de temps de proche en proche à partir de l'instant $t = n$ jusque l'instant $t = 0$ à l'aide des méthodes de la *programmation dynamique* [Puterman, 1994].

Nous nous intéressons dans la suite à des problèmes de décisions à horizons infinis. Nous présentons successivement les méthodes de planification pour le critère actualisé et pour le critère moyen. Dans ces deux cas, les politiques optimales sont stationnaires. Pour chacun des deux critères, nous énonçons les équations d'optimalité caractérisant les politiques optimales puis exposons des algorithmes pratiques permettant de résoudre ces équations. Tous les résultats présentés dans cette section sont démontrés dans le livre de référence de [Puterman, 1994].

1.2.1 Le critère γ-pondéré

Commençons par énoncer quelques propriétés fondamentales des processus de décision markovien à horizon infini pour le critère actualisé.

Évaluation d'une politique

On s'intéresse tout d'abord à trouver la fonction de valeur associée à une politique donnée. Pour toute politique stationnaire π, on définit l'opérateur $\mathcal{L}_\gamma^\pi$ sur l'espace $\mathcal{V}$ des fonctions de $\mathcal{X}$ dans $\mathbb{R}$ par la relation

$$\begin{aligned}\forall W \in \mathcal{V}\,, \quad \mathcal{L}_\gamma^\pi W(x) &\stackrel{\text{def}}{=} \mathbb{E}^\pi \left[r(X_t, A_t) + \gamma W(X_{t+1}) \,|\, X_t = x \right] \\ &= \sum_{a \in \mathcal{A}} \pi(x,a) \left(r(x,a) + \gamma \sum_{x' \in \mathcal{X}} P(x,a;x') W(x') \right) .\end{aligned}$$

En utilisant les notations présentées paragraphe 1.1.6, cette définition s'écrit :

$$\forall W \in \mathcal{V}\,, \quad \mathcal{L}_\gamma^\pi W \stackrel{\text{def}}{=} r^\pi + \gamma P^\pi W\,. \tag{1.5}$$

Soit π une politique stationnaire. La fonction de valeur V_γ^π associée à cette politique est solution de l'équation du point fixe pour l'opérateur $\mathcal{L}_\gamma^\pi$. De plus, l'opérateur $\mathcal{L}_\gamma^\pi$ défini par (1.5) est une contraction de $\mathcal{V}$ dans lui même : pour toutes fonctions V et W de $\mathcal{V}$,

$$\left\| \mathcal{L}_\gamma^\pi(V) - \mathcal{L}_\gamma^\pi(W) \right\|_\infty \leq \gamma \|V - W\|_\infty\,.$$

D'après le théorème du point fixe, cela implique qu'il existe une unique solution à l'équation du point fixe de cet opérateur. Donc, V_γ^π est l'unique fonction de $\mathcal{V}$ telle que

$$\forall x \in \mathcal{X}\,, \quad V_\gamma^\pi(x) = \sum_{a \in \mathcal{A}} \pi(x,a) \left(r(x,a) + \gamma \sum_{x' \in \mathcal{X}} P(x,a;x') V_\gamma^\pi(x') \right) .$$

Cette équation peut se réécrire sous la forme vectorielle

$$V_\gamma^\pi = (I - \gamma P^\pi)^{-1} r^\pi \ . \tag{1.6}$$

La fonction de valeur V_γ^π est alors obtenue en résolvant un système d'équations linéaires.

Équation d'optimalité

Nous exposons maintenant des équations caractérisant la fonction de valeur optimale pour le critère γ-pondéré :

$$V_\gamma^*(x) = \sup_{\pi \in \Pi^A} V_\gamma^\pi(x) \ .$$

On définit *l'opérateur de Bellman* $\mathcal{L}_\gamma : \mathcal{V} \to \mathcal{V}$ tel que

$$\forall W \in \mathcal{V} \ , \quad \forall x \in \mathcal{X} \ , \quad \mathcal{L}_\gamma W(x) \stackrel{\text{def}}{=} \max_{a \in \mathcal{A}} \mathbb{E}\left[r(X_t, A_t) + \gamma W(X_{t+1}) \,|\, X_t = x, A_t = a \right] \tag{1.7}$$

$$= \max_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \sum_{x' \in \mathcal{X}} P(x,a;x') W(x') \right\} \ . \tag{1.8}$$

La fonction de valeur optimale pour le critère γ-pondéré vérifie l'équation dite de *Bellman* : pour tout $x \in \mathcal{X}$

$$V_\gamma^*(x) = \max_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \sum_{x' \in \mathcal{X}} P(x,a;x') V_\gamma^*(x') \right\} \ . \tag{1.9}$$

De plus, l'opérateur $\mathcal{L}_\gamma$ étant une contraction de $\mathcal{V}$ dans lui même, la fonction V_γ^* est l'unique solution de l'équation du point fixe de cet opérateur. Ainsi, pour déterminer V_γ^*, il suffit de trouver la solution de cette équation. Lorsque les espaces d'état $\mathcal{X}$ et d'action $\mathcal{A}$ sont discrets, le maximum de l'équation (1.9) est atteint, et il existe alors une politique stationnaire déterministe π^* optimale définie comme étant la politique gloutonne par rapport à V_γ^*.

Recherche de la politique optimale

Deux algorithmes très utilisés pour déterminer la politique optimale peuvent être déduits des équations présentées ci-dessus. L'*algorithme d'itération sur les valeurs* (voir algorithme 1.1) consiste à chercher par itérations successives le point fixe de l'opérateur $\mathcal{L}_\gamma$ [Bellman, 1956; Puterman, 1994]. Pour toute fonction de valeur initiale $V_0 \in \mathcal{V}$, la suite des $(V_n)_{n \geq 1}$ déterminée par $V_{n+1} = \mathcal{L}_\gamma V_n$ converge en norme infinie vers V_γ^*. De plus, pour tout $\epsilon > 0$, une fois que la condition d'arrêt de l'algorithme donnée par ϵ est vérifiée, la politique stationnaire π, définie par l'équation (1.11), est *ϵ-optimale* :

$$\forall x \in \mathcal{X} \ , \quad V_\gamma^\pi(x) \geq V_\gamma^*(x) - \epsilon \ .$$

En utilisant une méthode du type élimination d'actions (voir paragraphe 6.7 de [Puterman, 1994]), il est possible d'introduire un autre critère d'arrêt garantissant que la politique obtenue à la fin de l'algorithme est bien l'optimale.

Le deuxième algorithme appelé *algorithme d'itération sur les politiques* (voir algorithme 1.2) utilise le fait que toute politique π_1 peut être améliorée en appliquant l'opérateur $\mathcal{L}_\gamma$. En effet, toute politique π_2 gloutonne par rapport à $V_\gamma^{\pi_1}$ vérifie $V_\gamma^{\pi_2} \geq V_\gamma^{\pi_1}$. L'égalité a lieu lorsque les politiques π_1 et π_2 sont optimales. L'algorithme d'itération sur les politiques se divise en deux phases :

Algorithme 1.1 Algorithme d'itération sur les valeurs

Initialisation : $V_0 \in \mathcal{V}$, $\epsilon > 0$, $n = 0$.

Pour tout $n \geq 0$,

pour tout $x \in \mathcal{X}$,

$$V_{n+1}(x) = \max_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \sum_{x' \in \mathcal{X}} P(x,a;x') V_n(x') \right\}$$

jusqu'à ce que

$$\|V_{n+1} - V_n\|_\infty \leq \epsilon \frac{1-\gamma}{2\gamma} \; . \tag{1.10}$$

Pour tout $x \in \mathcal{X}$,

$$\pi(x) \in \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \sum_{x' \in \mathcal{X}} P(x,a;x') V_n(x') \right\} \; . \tag{1.11}$$

– *évaluation de la politique* : il s'agit de calculer la fonction de valeur V_γ^π d'une politique π donnée en résolvant le système d'équations linéaire (1.6) ou en cherchant de proche en proche le point fixe de l'opérateur $\mathcal{L}_\gamma^\pi$;

– *amélioration de la politique* : cette phase permet d'améliorer la politique précédente en utilisant la politique gloutonne par rapport à la fonction de valeur calculée précédemment.

Cet algorithme termine en un nombre fini d'itérations. La politique stationnaire π^* ainsi obtenue est solution de l'équation du point fixe de l'opérateur $\mathcal{L}_\gamma$ et est donc une politique optimale. Notons P^π la matrice de transition entre les états sous la politique déterministe stationnaire π : pour tout états x et y

$$P^\pi(x,y) = P(x,\pi(x);y) \; .$$

Algorithme 1.2 Algorithme d'itération sur les politiques

Initialisation : $\pi_0 : \mathcal{X} \to \mathcal{A}$ quelconque, $n = 0$.

Pour tout $n \geq 0$

Évaluation de la règle de décision π_n

Calcul de V_n, solution de

$$(I - \gamma P^{\pi_n}) V_n = r^{\pi_n}$$

Amélioration de la règle de décision

$$\forall x \in \mathcal{X} \; , \quad \pi_{n+1}(x) \in \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ r(x,a) + \gamma \sum_{x' \in \mathcal{X}} P(x,a;x') V_n(x') \right\} \tag{1.12}$$

jusqu'à ce que $\pi_{n+1} = \pi_n$.

$\pi^* = \pi_n$.

1.2.2 Le critère moyen

Nous nous intéressons maintenant au critère moyen. Tout comme dans la partie précédente, les politiques considérées ici sont déterministes, stationnaires et markoviennes. Le

critère moyen s'écrit, pour tout $x \in \mathcal{X}$,

$$\eta^\pi(x) = \lim_{n\to\infty} \frac{1}{n} V_n^\pi(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} ((P^\pi)^t r^\pi)(x) \ . \tag{1.13}$$

Pour toute politique stationnaire π et toute matrice de transition P, la matrice limite

$$\bar{P}^\pi \stackrel{\text{def}}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} (P^\pi)^t$$

existe si l'espace d'état $\mathcal{X}$ est fini [Puterman, 1994].

Lien avec le critère pondéré

Pour toute politique stationnaire π et pour tout $\gamma \in]0,1[$, le développement en série de Laurent de la fonction V_γ^π permet d'écrire

$$V_\gamma^\pi = \frac{1}{1-\gamma} \eta^\pi + h^\pi + O(|1-\gamma|) \ , \tag{1.14}$$

où η^π et h^π, appelés respectivement le *gain* et le *biais* du MDP sous la politique π, sont définis comme suit

$$\eta^\pi = \bar{P}^\pi r^\pi \ , \qquad h^\pi = H^\pi r^\pi \ ,$$

où $H^\pi \stackrel{\text{def}}{=} (I - P^\pi + \bar{P}^\pi)^{-1} - \bar{P}^\pi$. On remarque que la fonction η^π satisfaisant l'équation (1.14) n'est rien d'autre que la fonction de valeur moyenne associée à la politique π. On a alors

$$\eta^\pi = P^\pi \eta^\pi$$

et que

$$\eta^\pi + h^\pi = r^\pi + P^\pi h^\pi \ .$$

Une politique π est dite *Blackwell optimale* si elle est optimale pour tous les problèmes γ-pondérés avec $\gamma \in [\bar{\gamma}, 1]$ pour un réel $\bar{\gamma} < 1$. Les politiques Blackwell optimales permettent de faire le lien entre les politiques optimales pour le critère moyen et le critère actualisé. En effet, pour deux politiques π et π' Blackwell optimales, les critères moyens et biais sont égaux : $\eta^\pi = \eta^{\pi'} = \eta^*$ et $h^\pi = h^{\pi'} = h^*$ (voir proposition 4.1.4 de [Bertsekas, 1995], volume 2). De plus, pour tout état $x \in \mathcal{X}$, le couple (η^*, h^*) vérifie les deux équations suivantes :

$$\eta^*(x) = \max_{a \in \mathcal{A}} \ P(x,a;x')\eta^*(x') \tag{1.15}$$

et

$$\eta^*(x) + h^*(x) = \max_{a \in \bar{\mathcal{A}}} \left\{ r(x,a) + \sum_{x' \in \mathcal{X}} P(x,a;x') h^*(x') \right\} \tag{1.16}$$

où $\bar{\mathcal{A}}$ est l'ensemble des actions qui maximisent l'équation (1.15) ; et toute politique Blackwell optimale π est telle que $\pi(x)$ est l'action qui maximise le membre de droite de l'équation (1.16). Ces différents résultats impliquent qu'une politique Blackwell optimale est également optimale pour le critère moyen. On remarque que h^* est défini à une constante près. En effet, pour tout entier $k \in \mathbb{N}$, $h^* = H^{\pi^*} r^{\pi^*} + k\mathbf{e}$ satisfait aussi l'équation (1.16) où $\mathbf{e}$ est le vecteur dont toutes les coordonnées sont égales à 1.

Equations d'optimalité

Si η et h sont deux vecteurs qui satisfont les *équations de Bellman* (1.15) et (1.16) alors η est égale à la récompense moyenne optimale η^*. De plus, si π est une politique gloutonne par rapport à h^*, c'est-à-dire telle que, pour tout $x \in \mathcal{X}$, $\pi(x)$ maximise le membre de droite de l'équation (1.16), alors π est une politique stationnaire optimale. Ainsi, la politique optimale pour le critère moyen est optimale pour un critère γ-pondéré pour une valeur de γ assez proche de 1.

Comme dans l'étude du critère pondéré, on définit un opérateur de Bellman $\mathcal{L}$ tel que, pour toute fonction $h \in \mathcal{V}$ et tout état $x \in \mathcal{X}$,

$$(\mathcal{L}h)(x) = \max_{a \in \bar{\mathcal{A}}} \left\{ r(x,a) + \sum_{x' \in \mathcal{X}} P(x,a;x')h(x') \right\} .$$

De même, définissons l'opérateur $\mathcal{L}^\pi$ associé à une politique déterministe stationnaire π : pour toute fonction $h \in \mathcal{V}$ et tout état $x \in \mathcal{X}$,

$$(\mathcal{L}^\pi h)(x) = r(x,\pi(x)) + \sum_{x' \in \mathcal{X}} P(x,\pi(x);x')h(x') .$$

Critère moyen indépendant de l'état

Sous certaines conditions sur l'espace d'état du processus de décision markovien, la récompense moyenne optimale η^* est indépendante de l'état initial. En particulier, c'est le cas si l'espace d'état $\mathcal{X}$ peut être partitionné en deux sous-espaces $\mathcal{X}_t$ et $\mathcal{X}_r$ ($\mathcal{X} = \mathcal{X}_t \cup \mathcal{X}_r$) tels que

- tous les états de $\mathcal{X}_t$ sont transients sous toute politique stationnaire ;
- pour tous les états x et x' dans $\mathcal{X}_c$, x' est accessible à partir de x : il existe une politique stationnaire π et un entier k tels que $\mathbb{P}^\pi [X_k = x' \,|\, X_0 = x, \pi] > 0$.

Un MDP vérifiant ces conditions est appelé *communiquant*. On rappelle qu'un état d'une chaîne de Markov est dit transient (resp. récurrent) s'il n'est visité, avec probabilité 1, qu'un nombre fini de fois (resp. un nombre infini de fois) quelque soit l'état de départ. Dans ce cas, pour une politique π stationnaire fixée, et pour tous $x, x' \in \mathcal{X}$ $\bar{P}^\pi(x,x') = \nu^\pi(x')$ où ν^π est la loi stationnaire (ou loi invariante) de la chaîne de Markov des états lorsque la politique π est suivie. On définit ν^π comme un vecteur ligne qui vérifie $\nu^\pi P^\pi = \nu^\pi$. La récompense moyenne η^π reçue en suivant une politique stationnaire π étant égale à $\bar{P}^\pi r^\pi$, celle ci est également indépendante de l'état lorsque le MDP est communiquant. Elle vérifie alors l'équation suivante

$$\eta^\pi + h^\pi(x) = r(x,\pi(x)) + \sum_{x' \in \mathcal{X}} P(x,\pi(x);x')h^\pi(x') ,$$

où h^π est un vecteur de biais associé à la politique π. La récompense moyenne optimale vérifie l'équation d'optimalité

$$\eta^* + h^*(x) = \max_{a \in \mathcal{A}} \left\{ r(x,a) + \sum_{x' \in \mathcal{X}} P(x,a;x')h^*(x') \right\} \tag{1.17}$$

Tout comme précédemment les vecteurs h^π et h^* ne sont pas uniques. La politique gloutonne par rapport à un vecteur h^* satisfaisant l'équation (1.17) est une politique optimale.

Dans le reste de cette partie, on restreint la présentation aux MDP communiquants.

Recherche d'une politique optimale

De manière similaire au cas du critère actualisé, il existe deux familles de méthodes permettant de déterminer une politique optimale maximisant le critère moyen : la première se base sur l'itération sur les valeurs alors que la deuxième est liée à l'algorithme d'itération sur les politiques. Toute politique optimale pour le critère γ-pondéré, pour une valeur de γ assez proche de 1, étant également optimale pour le critère moyen, il est possible de déterminer une politique optimale pour le critère moyen en utilisant un des algorithme 1.1 ou 1.2 avec une valeur de γ assez grande. Cependant, il n'existe pas de résultat permettant de dire quelle est la valeur de γ à utiliser pour assurer que la politique ainsi obtenue soit bien égale à la politique optimale pour le critère moyen.

L'algorithme d'itération sur les valeurs dans le cas du critère moyen est présenté ci-dessous. Soit

$$osc : \mathcal{V} \to \mathbb{R}$$

la fonction *oscillation* définie pour tout vecteur $V \in \mathcal{V}$ par $osc(V) = \max_x V(x) - \min_x V(x)$. Sous certaines conditions sur les matrices de transitions du MDP, que nous n'expliciterons pas ici (voir [Puterman, 1994]), pour tout $\epsilon > 0$, il existe un entier n tel que $osc(V_{n+1} - V_n) < \epsilon$. La condition d'arrêt (1.18) est alors vérifiée au bout d'un certain temps et la politique obtenue est ϵ-optimale.

Algorithme 1.3 Itération sur les valeurs

Initialisation : $V_0 \in \mathcal{V}$, $\epsilon > 0$ et $n = 0$.

Pour tout $n \geq 0$

Pour tout $x \in \mathcal{X}$,

$$V_{n+1}(x) = \max_{a \in \mathcal{A}} \left\{ r(x,a) + \sum_{x' \in \mathcal{X}} P(x,a;x') V_n(x') \right\} ,$$

jusqu'à ce que

$$osc(V_{n+1} - V_n) \leq \epsilon \, . \tag{1.18}$$

Pour tout $x \in \mathcal{X}$

$$\pi(x) \in \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ r(x,a) + \sum_{x' \in \mathcal{X}} P(x,a;x') V_n(x') \right\} .$$

La convergence de cet algorithme peut être très lente. Un algorithme renormalisant V_n à chaque itération permet d'éviter de tels problèmes numériques (voir algorithme 1.4). L'algorithme d'itération sur les politiques dans le cas du critère moyen est présenté dans l'algorithme 1.5.

1.3 Apprentissage par renforcement

Dans la section précédente, nous avons présenté les équations caractérisant les politiques optimales dans un problème de décision markovien ainsi que des algorithmes permettant d'estimer ces politiques en pratique. Ces méthodes pour résoudre le problème de planification supposent que les probabilités de transition $P(\cdot,\cdot;\cdot)$ ainsi que la loi des récompenses $\mathcal{R}(\cdot,\cdot;\cdot)$ soient connues. Les méthodes d'apprentissage par renforcement permettent, quant à elles, de considérer des situations dans lesquelles le modèle est inconnu de l'agent. Ce domaine de l'apprentissage automatique a suscité un grand intérêt ces dernières années et de nombreux

Algorithme 1.4 Itération sur les valeurs relatives

Initialisation : $V_0 \in \mathcal{V}$, $x^* \in \mathcal{X}$ et $\epsilon > 0$.
$w_0 = V_0 - V_0(x^*)\mathbf{e}$ et $n = 0$.
Pour tout $n \geq 0$
  $V_{n+1} = \mathcal{L}w_n(x)$
  $w_{n+1} = V_{n+1} - V_{n+1}(x^*)\mathbf{e}$
jusqu'à ce que $osc(V_{n+1} - V_n) \leq \epsilon$.
Pour tout $x \in \mathcal{X}$,

$$\pi(x) \in \underset{a \in \mathcal{A}}{\operatorname{argmax}} \left\{ r(x,a) + \sum_{x' \in \mathcal{X}} P(x,a;x')V_n(x') \right\}$$

Algorithme 1.5 Algorithme d'itération sur les politiques

Intialisation : $\pi_0 : \mathcal{X} \to \mathcal{A}$ quelconque. $n = 0$.
Pour tout $n \geq 0$
  Évaluation de la règle de décision π_n
  Calcul de $\eta_n \in \mathbb{R}$ et $h_n \in \mathcal{V}$ tels que

$$r^{\pi_n} - \eta_n e + (P^{\pi_n} - I)h_n = 0$$

où $min_x h_n(x) = 0$.
  Amélioration de la règle de décision

$$\forall x \in \mathcal{X}\ , \quad \pi_{n+1} \in \underset{\pi}{\operatorname{argmax}} \{r^\pi + P^\pi h_n\}$$

jusqu'à ce que $\pi_{n+1} = \pi_n$.
$\pi^* = \pi_n$.

ouvrages lui sont dédiés. On notera en particulier les livres [Sutton and Barto, 1998; Bertsekas and Tsitsiklis, 1996; Sigaud and Buffet, 2008].

L'agent n'ayant initialement aucune connaissance sur le modèle d'interaction, il doit apprendre les conséquences de ses actions tout en interagissant avec l'environnement. Son but reste de maximiser les récompenses reçues à long terme, mais, en plus de vouloir jouer les actions qui lui semblent les plus favorables, il doit également sélectionner des actions peut-être sous-optimales au regard des informations déjà collectées, mais qui lui permettent d'acquérir des informations sur la réaction de l'environnement. Dans l'exemple de la *rivière* (voir le paragraphe 1.1.1), le nageur a besoin de jouer suffisamment de fois les deux actions (nager à gauche et à droite) dans chaque état pour découvrir qu'il peut gagner une grande récompense en arrivant sur la rive de droite. De même, dans le modèle de bandit, le joueur doit jouer suffisamment chacun des bras afin de déterminer quel est l'optimal. Néanmoins, le joueur évite de jouer trop souvent des bras qui ne permettent de recevoir qu'une récompense très faible. Ce compromis entre jouer des actions qui semblent être les meilleures (*exploitation*) et acquérir plus de connaissances sur la dynamique de l'environnement en choisissant des actions potentiellement sous-optimales (*exploration*) est une problématique centrale de l'apprentissage par renforcement.

Nous nous intéressons dans cette thèse aux méthodes d'apprentissage par renforcement dites *« on-line »*. Dans ces méthodes, la recherche de la politique optimale s'accomplit durant l'interaction avec l'environnement. Certains travaux de recherches ont été effectués dans des cadres différents. En particulier, les méthodes dites *batch* ont comme objectif d'identifier une politique optimale à partir d'une séquence d'état-action-récompense observée au préalable [Antos et al., 2007; Ernst et al., 2006; Kalyanakrishnan and Stone, 2007]. On différencie deux grandes familles de méthodes d'apprentissage par renforcement *« on-line »*. Les méthodes dites *« model-free »* et *« model-based »* diffèrent selon que l'on maintient ou non un modèle explicite de la dynamique du MDP. Les méthodes « model-free » tentent de trouver une politique optimale en estimant la table état-action Q^*. Dans les méthodes dites « model-based », l'agent estime le modèle tout au long de l'interaction avec l'environnement et utilise les algorithmes de planification présentés dans le paragraphe précédent pour déterminer la politique à suivre.

Différentes méthodes ont été proposées pour garantir le bon fonctionnement et mesurer la performance d'un algorithme d'apprentissage par renforcement. Une première garantie est de prouver que la fonction de valeur de la politique obtenue à chaque étape de l'algorithme converge vers la fonction de valeur optimale. Ceci prouve, qu'asymptotiquement, la politique obtenue en suivant l'algorithme est optimale. Cette première garantie est essentielle mais semble loin d'être suffisante. En effet, si la vitesse de convergence est extrêmement faible, cette convergence n'a que très peu d'intérêt en pratique. Une deuxième mesure est donc la vitesse de convergence de la fonction de valeur vers la fonction de valeur optimale. Les algorithmes peuvent alors être comparés en utilisant ce critère de performance. Cependant, cette mesure ne prend aucunement en compte la perte, en terme de récompense, durant la phase de convergence. Pour deux algorithmes ayant une même vitesse de convergence, il serait intéressant de choisir l'algorithme qui permet à l'agent de prendre des décisions de manière à accumuler le plus de récompenses, même pendant la phase de convergence. Une troisième mesure de performance, appelée regret, permet de faire cette distinction. Il s'agit de la somme des récompenses reçues lorsque l'agent suit l'algorithme comparée à la somme des récompenses qui seraient accumulées si l'agent connaissait le modèle d'interaction et jouait de manière optimale.

La suite de cette section est divisée en deux parties. La première concerne les méthodes « model-free » et la seconde celles « model-based ». Pour ces deux familles de méthodes, nous

exposons les principes fondamentaux des différentes approches et les quelques algorithmes de référence. Cependant, les travaux présentés dans la suite de cette thèse étant des méthodes « model-based », nous décrivons plus en détail les approches proposées dans la littérature pour cette famille de méthodes.

Comme mentionné précédemment, nous nous intéressons dans cette thèse à des processus de décisions markoviens à espaces d'états et d'actions finis. Nous suggérons une liste non-exhaustive de travaux considérant des modèles à espaces d'états ou d'actions infinis pour le lecteur intéressé [Baird, 1995; Bradtke and Barto, 1996; Sutton et al., 2000; Boyan, 2002; Szepesvári and Smart, 2004; Munos, 2003; Ernst et al., 2006; Van Roy, 2006; Munos and Szepesvári, 2008].

1.3.1 Méthodes « model-free »

Les méthodes « model-free » [1] consistent à estimer la table état-action tout au long de l'interaction entre l'agent et l'environnement en agissant de manière à ce que la table estimée converge vers la table d'état-action optimale Q^* sans se préoccuper du modèle sous-jacent. Pour chaque couple état-action (x, a), $Q^*(x, a)$, définie paragraphe 1.1.5, désigne l'espérance de la somme (potentiellement pondérée par un facteur γ) des récompenses reçues si l'environnement est dans l'état x et que l'agent joue l'action a puis suit la politique optimale. Comme expliqué dans la section précédente, les équations de la programmation dynamique permettent de calculer la fonction de valeur optimale V^*, ou la fonction de valeur V^π associée à une politique π, en fonction des probabilités de transitions $P(\cdot, \cdot; \cdot)$ et de l'espérance des récompense $r(\cdot, \cdot)$. Si P et r ne sont pas connues de l'agent, celui-ci peut estimer les fonctions de valeur à partir de l'interaction avec l'environnement en observant quelles récompenses il accumule après être passé par chaque état. Nous présentons, dans ce paragraphe, des méthodes pour estimer la fonction de valeur V^π d'une politique stationnaire π le long de cette interaction puis exposons des algorithmes classiques d'apprentissage par renforcement « model-free » basés sur ces méthodes. Nous nous limitons ici au cas du critère actualisé qui est le plus souvent utilisé pour ce type d'approches. Cependant, des formules similaires peuvent être énoncées pour les autres critères.

Méthodes de différences temporelles pour évaluer la performance d'une politique

La fonction de valeur V_γ^π associée à une politique stationnaire déterministe π fixée peut être estimée à l'aide de simulations. Une méthode naturelle pour ce faire est d'utiliser des trajectoires indépendantes durant lesquelles la politique π est suivie. Ces trajectoires doivent alors être générées à l'aide d'un simulateur. On peut ainsi construire K trajectoires indépendantes $(X_0^k, X_1^k, \ldots, X_{N_k}^k)_{k \leq K}$ commençant toutes à l'état x. La fonction de valeur $V_\gamma^\pi(x)$ est alors estimée en utilisant un algorithme itératif d'estimation d'une moyenne. Cela consiste à définir la suite des vecteurs $(\hat{V}_k)_k$ telle que pour tout $k \leq 1$

$$\hat{V}_0(x) = 0 \ ,$$
$$\hat{V}_{k+1}(x) = (1 - \alpha_k)\hat{V}_k(x) + \alpha_k \sum_{t=0}^{N_k} \gamma^t r(X_t^k, \pi(X_t^k)) \quad \forall 0 \leq k \leq K \ ,$$

où les α_k sont des pas d'apprentissage. La longueur N_k de ces trajectoires peut être fixée à l'avance ou être aléatoire (s'il s'agit du temps d'atteinte d'un état particulier, par exemple). Elle peut donc varier d'une expérience à l'autre. Cette suite converge vers la fonction de valeur

1. dans la littérature francophone, ces méthodes sont souvent appelées *méthodes directes* [Sigaud and Buffet, 2008].

V^π. Cette méthode est appelé *méthode de Monte-Carlo*. Comme mentionné précédemment, nous nous intéressons à des méthodes d'apprentissage par renforcement « on-line ». L'agent doit donc estimer la fonction de valeur durant l'interaction sans avoir recours à un simulateur. Pour cela, il mémorise les états qu'il visite et peut utiliser une méthode similaire en mettant à jour $\hat{V}(x)$ à chaque fois que l'état x est visité mais ceci introduit inévitablement un biais.

Les *méthodes de différences temporelles* (TD) introduites par [Sutton, 1988] permettent de résoudre cette difficulté en utilisant l'incrémentalité de la programmation dynamique. L'algorithme initial TD(0) est un algorithme stochastique dont le but est de résoudre l'équation du point fixe vérifiée par la fonction de valeur V_γ^π. L'idée est de remarquer qu'un estimateur non biaisé de $\mathcal{L}_\gamma^\pi V(x)$ est $r(x,\pi(x)) + \gamma V(x')$ où $x' \sim P(x,\pi(x);.)$. Ainsi, la fonction de valeur de la politique π peut être calculée de manière itérative selon l'algorithme d'approximation stochastique du point fixe (voir [Bertsekas and Tsitsiklis, 1996]). On note $\hat{V}_t$ la t-ième itérée de l'estimation de la fonction de valeur. Pour tout t,

$$\begin{aligned} \hat{V}_{t+1}(X_t) &= (1-\alpha_t)\hat{V}_t(X_t) + \alpha_t\left(R_t + \gamma\hat{V}_t(X_{t+1})\right) \\ &= \hat{V}_t(X_t) + \alpha_t\delta_t \end{aligned} \tag{1.19}$$

où $\delta_t = R_t + \gamma\hat{V}_t(X_{t+1}) - \hat{V}_t(X_t)$ et α_t est un pas d'apprentissage. Pour tout $x \neq X_{t+1}$, $\hat{V}_{t+1}(x) = \hat{V}_t(x)$. Le nom de *différence temporelle* vient du fait que δ_t est défini comme étant la différence entre entre les valeurs d'états successifs. Cet algorithme converge dès que les pas d'apprentissage vérifient $\sum_{t\geq 0}\alpha_t = \infty,\ \sum_{t\geq 0}\alpha_t^2 < \infty$.

Trace d'éligibilité

Les deux méthodes Monte-Carlo et TD(0) peuvent être regroupées sous une même approche appelée $TD(\lambda)$ [Sutton, 1988], le paramètre $\lambda \in [0,1]$ permettant d'interpoller entre ces deux méthodes. Cette approche peut être expliquée et motivée de différentes façons [Sutton and Barto, 1998; Bertsekas, 1995; Szepesvári, 2010], nous décrivons ici brièvement l'algorithme TD(λ) comme un algorithme stochastique du point fixe pour l'opérateur

$$\mathcal{H}_\lambda = (1-\lambda)\sum_{k>0}\lambda^k(\mathcal{L}_\gamma^\pi)^{k+1}\ .$$

On observe que la fonction de valeur V_γ^π est solution de l'équation du point fixe de tous les opérateurs $(\mathcal{L}_\gamma^\pi)^k$ pour $k \geq 0$. Ainsi, pour tout $0 \leq \lambda \leq 1$, V_γ^π est solution de l'équation du point fixe de $\mathcal{H}_\lambda$. La règle de mise à jour de la fonction de valeur est alors la suivante : pour tout $x \in \mathcal{X}$,

$$\hat{V}_{t+1}(x) = \hat{V}_t(x) + \alpha_t\delta_t z_t(x)\ ,$$

où z_t est la *trace d'éligibilité* définie par

$$z_t(x) = \sum_{k=0}^{t}(\gamma\lambda)^{t-k}\mathbb{1}_{\{X_k=x\}}$$

et $z_0(x) = 0$. L'impact des différences temporelles des transitions futures sur l'estimation de la valeur de l'état courant est pondérée par λ. En supposant que tous les états sont visités par une infinité de trajectoires et que les pas $(\alpha_t)_{t\geq 0}$ satisfont $\sum_{t\geq 0}\alpha_t = \infty,\ \sum_{t\geq 0}\alpha_t^2 < \infty$ alors la suite des $(\hat{V}_k)_k$ converge presque sûrement vers V_γ^π.

Algorithmes

Les méthodes $TD(\lambda)$ permettent d'estimer la valeur d'une politique. En apprentissage par renforcement, il est certes intéressant de pouvoir mesurer la valeur d'une politique donnée mais le but est de trouver la politique optimale. Pour cela, il existe principalement deux méthodes dans la littérature : l'une, appelée *SARSA*, se rapproche de l'algorithme d'itération sur les politiques (voir l'algorithme 1.2) tandis que l'autre, *Q-learning*, est une méthode plus proche de l'itération sur les valeurs (voir l'algorithme 1.1).

L'*algorithme SARSA* appartient à la famille dite « Actor-Critic » : il consiste à alterner entre jouer une politique (acteur) puis évaluer cette politique (critique) afin d'en déduire une meilleure par la suite. Cet algorithme, détaillé ci-dessous, est semblable à l'algorithme TD(0) et utilise les tables état-action Q à la place des fonctions de valeur V. Il s'agit donc d'un algorithme d'itération sur les politiques dans lequel l'étape d'évaluation de la politique est faite selon l'itération suivante :

$$Q_{t+1}(X_t, A_t) = (1 - \alpha_t)Q_t(X_t, A_t) + \alpha_t \left(R_t + \gamma Q_t(X_{t+1}, A_{t+1})\right) \ .$$

Algorithme 1.6 Algorithme SARSA

Entrées: $(\alpha_t)_t$, une politique π_0, une table Q_0
Observer X_0,
Choisir $A_0 \sim \pi_{Q_0}$
Pour $t \geq 0$ **faire**
Réception de R_t et observer X_{t+1},
Choisir $A_{t+1} \sim \pi_{Q_t}$
Évaluation de la politique :
$Q_{t+1}(X_t, A_t) = (1 - \alpha_t)Q_t(X_t, A_t) + \alpha_t(R_t + \gamma Q_t(X_{t+1}, A_{t+1}))$
Amélioration de la politique :
Calculer π_{t+1} à partir de Q_{t+1} : $\forall x \in \mathcal{X}$, $\pi_{t+1}(x) = \text{argmax}_{a \in \mathcal{A}} Q_{t+1}(x, a)$
fin Pour

Tout le long de l'interaction avec l'environnement, l'agent choisit les actions à effectuer en fonction de l'estimée courante Q_t de la table état action, en suivant la politique notée π_{Q_t}. Cette politique est, par exemple, être la politique gloutonne par rapport à la table état-action (voir section 1.1.5). Dans ce cas, on dit que l'agent suit *une politique d'exploitation*. On remarque que, en suivant une telle approche, l'agent peut ne pas visiter certains états et donc ne jamais trouver la politique optimale. Une autre méthode consisterait à essayer de visiter tous les couples état-action un grand nombre de fois en suivant à chaque instant une politique aléatoire sans tenir compte de la table état-action calculée. Il s'agirait alors d'*une politique d'exploration* uniquement. Tout bon algorithme d'apprentissage par renforcement doit trouver un équilibre entre exploration et exploitation. Citons deux stratégies couramment utilisées pour gérer cet équilibre :

- la politique ϵ-gloutonne propose de sélectionner l'action gloutonne par rapport à Q_t avec une probabilité ϵ et une action aléatoire avec probabilité $1 - \epsilon$,
- la politique *softmax*, ou de *Boltzmann*, exécute l'action a selon une probabilité :

$$\pi_t(x, a) \stackrel{\text{def}}{=} \frac{\mathrm{e}^{\frac{1}{T}Q_t(x,a)}}{\sum_{a' \in \mathcal{A}} \mathrm{e}^{\frac{1}{T}Q_t(x,a')}} \ ,$$

la constante T étant un paramètre que l'on peut éventuellement faire varier au cours du temps et qui contrôle le taux d'exploration.

Les propriétés de convergence de l'algorithme SARSA sont liées à la nature de la politique π_{Q_t} utilisée, qui dépend de la table état-action Q_t calculée à chaque instant. Ceci complique de façon notable les preuves de convergence. Néanmoins, la fonction Q_t ainsi construite converge vers Q^*_γ lorsque la politique d'apprentissage est telle que chaque action est exécutée une infinité de fois dans chacun des états et, à la limite, la politique est gloutonne par rapport à la table état-action [Singh et al., 2000].

Le deuxième algorithme classique « model-free » est l'*algorithme Q-learning*, proposé par Watkins en 1989. Il consiste à estimer directement la table état-action optimale en s'appuyant sur l'équation du point fixe de l'opérateur de Bellman. La table état-action Q est mise à jour à chaque instant t en fonction de l'état du système X_t, l'action $A_t \sim \pi_t(X_t, .)$ sélectionnée en utilisant une politique quelconque π fixée au préalable et l'état du système à l'instant suivant X_{t+1} :

$$Q_t(X_t, A_t) = Q_{t-1}(X_t, A_t) + \alpha_t \left(r(X_t, A_t) + \gamma \max_{b \in \mathcal{A}} Q_{t-1}(X_{t+1}, b) - Q_{t-1}(X_t, A_t) \right) .$$

Pour tous les couples état-action (x, a) tels que $X_t \neq x$ ou $A_t \neq a$, $Q_t(x, a) = Q_{t-1}(x, a)$. L'algorithme est décrit ci-dessous.

Algorithme 1.7 Algorithme Q-learning

Entrées: $(\alpha_t)_{t\geq 0}$, une politique π, une table Q_0
Observer X_0,
Pour $t \geq 0$ **faire**
Choisir A_t en suivant π
Réception de R_t et observer X_{t+1}.
$Q_{t+1}(X_t, A_t) = (1 - \alpha_t)Q_t(X_t, A_t) + \alpha_t(R_t + \gamma \max_{a\in\mathcal{A}} Q_t(X_{t+1}, a))$
fin Pour

En supposant que tous les couples état-action sont visités une infinité de fois et sous l'hypothèse habituelle sur les pas d'apprentissage, la suite des $(Q_t)_{t\geq 0}$ converge presque sûrement vers Q^*_γ [Watkins and Dayan, 1992]. Il est important de noter que, contrairement à l'algorithme SARSA, les propriétés de convergence de l'algorithme Q-learning sont indépendantes de la politique utilisée lors de l'interaction. C'est ce que l'on appelle un algorithme « off-policy », à opposer aux algorithmes tels que SARSA qui sont dits « on-policy ». L'algorithme Q-learning converge quelle que soit la manière dont les couples état-action sont sélectionnés tant qu'ils le sont tous régulièrement. Ainsi, lors de l'interaction, n'importe quelle politique permettant une exploration suffisante peut être utilisée.

Remarque 1. Comme pour les méthodes de différence temporelle, on peut déduire des algorithmes précédents les algorithmes SARSA(λ) et $Q(\lambda)$ en introduisant une trace d'éligibilité (voir [Sutton and Barto, 1998; Bertsekas and Tsitsiklis, 1996]).

1.3.2 Méthodes « model-based »

Contrairement aux méthodes « model-free », les algorithmes dits « model-based » [2] consistent à apprendre le modèle tout au long de l'interaction avec l'environnement et à jouer une politique basée sur cette estimation du modèle. Nous commencerons à présenter les méthodes « model-based » dans les modèles de bandits (voir exemple dans paragraphe 1.1.1) et expliquerons dans la suite comment ils ont été étendus aux processus de décision markovien quelconques.

2. dans la littérature francophone, ces méthodes sont souvent appelées *méthodes indirectes* [Sigaud and Buffet, 2008].

Modèle de bandit

Le modèle de bandit est un cas particulier de processus de décision markovien avec un seul état qui a suscité un grand intérêt depuis les années 50 [Robbins, 1952; Gittins, 1979]. A chaque instant t, un agent choisit une action A_t parmi un ensemble fini d'action $\{1, \dots, |\mathcal{A}|\}$ et reçoit une récompense aléatoire R_t suivant une distribution qui dépend de l'action choisie. Pour chaque action a, la distribution de probabilité de la récompense est inconnue de l'agent. On note $r(a) = \mathbb{E}[R_t \,|\, A_t = a]$ la moyenne de la récompense reçue lorsque le bras a est joué. L'agent cherche à sélectionner, à chaque instant, un bras qui lui permette de maximiser la récompense reçue à long terme. Il est donc confronté au fameux compromis entre exploration et exploitation. En effet, d'un côté il doit jouer régulièrement tous les bras pour accumuler de l'information sur la loi des récompenses en jouant les différents bras et d'un autre, il doit jouer le plus souvent possible le bras qui lui semble le meilleur pour maximiser ses gains [Auer et al., 2002; Audibert, 2010].

Soit $a^* \in \operatorname{argmax}_{a \in \mathcal{A}} \mathbb{E}[R_t \,|\, A_t = a]$ un *bras optimal*. Dans le modèle de bandit, la politique optimale est très simple : elle consiste à toujours jouer le bras optimal. On définit la récompense estimée à l'instant t par

$$\forall a \in \mathcal{A},\ \hat{r}_t(a) = \frac{\sum_{k=0}^{t-1} R_t \mathbb{1}_{\{A_k = a\}}}{\sum_{k=0}^{t-1} \mathbb{1}_{\{A_k = a\}}} \ . \tag{1.20}$$

De manière intuitive, jouer la politique optimale par rapport à la récompense estimée (aussi appelée politique gloutonne pour le modèle estimé) peut mener à de très mauvaises performances. En effet, supposons par exemple que la distribution des récompenses est une loi de Bernoulli. Imaginons que chacun des bras ait été joué une fois et que la récompense reçue en jouant le bras optimal a^* est de 0 alors que celle reçue en jouant un bras sous-optimal a vaut 1. Dans ce cas, les récompenses estimées de ces deux bras sont $\hat{r}_{|\mathcal{A}|}(a^*) = 0$ et $\hat{r}_{|\mathcal{A}|}(a) = 1$. La politique optimale par rapport au modèle estimé est alors de jouer le bras a. En jouant ce bras, l'estimateur $\hat{r}_t(a)$ décroît vers l'espérance de la récompense moyenne $r(a)$ mais reste strictement supérieur à $\hat{r}_t(a^*)$ qui vaut 0. Ainsi, le bras optimal a^* ne sera jamais joué.

Une politique simple pour équilibrer exploration et exploitation est la politique ϵ-gloutonne qui consiste à jouer avec probabilité $1-\epsilon$ le bras optimal par rapport à la récompense estimée et de jouer avec probabilité ϵ un autre bras au hasard. Cet ajout d'exploration permet de garantir que le bras optimal est détecté au bout d'un certain temps. En effet, l'agent ne s'arrête jamais d'explorer et ses estimations des récompenses moyennes finissent donc par être suffisamment précises pour dissocier le bras optimal des autres bras. Toute la difficulté dans cette approche est de déterminer la valeur du paramètre ϵ afin de garantir des performances satisfaisantes.

Une autre approche très connue d'apprentissage par renforcement dans le modèle de bandit est basée sur le principe d'*optimisme face à l'incertain* proposé par [Lai and Robbins, 1985]. Leur algorithme consiste à construire, à chaque instant, un intervalle de confiance pour la récompense reçue en jouant chacun des bras et de jouer le bras ayant la borne de confiance supérieure la plus grande. [Agrawal, 1995] propose d'utiliser une borne de confiance supérieure de la forme $\hat{r}_t(a) + \beta_t^a$ où $\hat{r}_t(a)$ est la moyenne empiriques de la récompense reçue en jouant le bras a et β_t^a est choisi de manière à ce que l'espérance de la récompense $r(a)$ appartienne à l'intervalle $[\hat{r}_t(a) - \beta_t^a; \hat{r}_t(a) + \beta_t^a]$ avec grande probabilité. A chaque instant t, l'agent choisit donc l'action ayant la plus grande borne de confiance supérieure

$$A_t = \operatorname*{argmax}_{a \in \mathcal{A}} \{\hat{r}_t(a) + \beta_t^a\} \ .$$

On note que, dans ce type d'approches, exploration et exploitation ne sont pas explicitement dissociés. En effet, contrairement à un algorithme de type ϵ-glouton où, à chaque instant,

l'agent décide de sélectionner une action au hasard (exploration) ou l'action gloutonne (exploitation), l'approche optimiste permet de gérer le compromis entre exploration et exploitation en comparant les bornes supérieures des intervalles de confiance pour chaque bras. Ce sont donc les termes $(\beta_t^a)_{a\in\mathcal{A}}$ qui contrôlent l'exploration à chaque instant. La fonction β_t^a étant décroissante en t, au bout d'un certain temps l'exploration fait place à l'exploitation.

Ces algorithmes sont appelés *optimistes* car l'agent joue en supposant que le vrai modèle d'interaction est le meilleur des modèles compatibles avec les données dont il dispose. En effet, l'ensemble des intervalles de confiance décrivent les modèles de bandits qui pourraient expliquer les récompenses reçues et jouer le bras dont la borne supérieure de ces intervalles est la plus grande revient à espérer que le vrai modèle est celui qui mène à la plus grande récompense possible.

Pour mesurer la performance d'un algorithme de bandit, on le compare à un *oracle* qui connaîtrait le bras optimal et qui jouerait en permanence celui-ci. Ainsi, on calcule le *regret* défini comme étant la différence entre la somme des récompenses espérées si on connaissait le meilleur bras et la récompense espérée en suivant l'algorithme sur un horizon fini n. On a

$$\text{Regret}_n = \sum_{t=0}^{n}(r(a^*) - r(A_t)) = nr(a^*) - \sum_{t=0}^{n} r(A_t) \ .$$

On s'intéressera également au regret espéré qui peut se réécrire selon le nombre de fois où chaque action a été choisi :

$$\mathbb{E}\left[\text{Regret}_n\right] = \mathbb{E}\left[\sum_{t=0}^{n}(r(a^*) - r(A_t))\right] = \sum_{a\in\mathcal{A}} \mathbb{E}\left[N_{t+1}(a)\right](r(a^*) - r(a)) \ .$$

où $N_{t+1}(a) = \sum_{t=0}^{n} \mathbb{1}_{\{A_t=a\}}$.

Un algorithme est dit *consistant* [Robbins, 1952] s'il vérifie

$$\lim_{n\to\infty} \sum_{t=0}^{n} \frac{\sum_{a\in\mathcal{A}} r(a)\mathbb{E}\left[\mathbb{1}_{\{A_t=a\}}\right]}{n+1} = r(a^*) \ .$$

Dans un cadre paramétrique, [Lai and Robbins, 1985] ont montré que tout algorithme consistant joue asymptotiquement chacun des bras sous-optimaux au moins $\log(n)$ fois sur un horizon de n instants. Plus précisément :

$$\liminf_{n\to\infty} \frac{\mathbb{E}\left[N_n(a)\right]}{\log(n)} \geq C(a^*, a) \tag{1.21}$$

La constante $C(a^*, a)$ est inversement proportionnelle à la divergence de Kullback-Leibler entre la distribution de probabilité des récompenses en jouant le bras a et celle en jouant le bras a^*. [Burnetas and Katehakis, 1996] ont étendu ce résultat à des modèles non paramétriques. Les algorithmes proposés par [Lai and Robbins, 1985; Agrawal, 1995; Honda and Takemura, 2010] sont asymptotiquement efficaces, au sens où, lorsque l'horizon n temps vers l'infini, le regret est proportionnel à $\log(n)$. [Honda and Takemura, 2010] ont prouvé que le regret asymptotique de leur algorithme est égal à la borne inférieure de celui-ci montrée par [Burnetas and Katehakis, 1996].

[Auer et al., 2002] ont proposé une étude de l'algorithme optimiste appelé UCB fournissant des bornes de regret non asymptotiques. Ils s'intéressent, quant à eux, à un modèle de récompense non paramétrique où les récompenses sont bornées. Le fameux algorithme UCB (pour *upper confidence bound*), décrit ci-dessous, est une variante de l'algorithme de [Agrawal, 1995]

où le bonus d'exploration est défini par $\beta_t^a = \sqrt{\frac{\alpha \log(t)}{N_t(a)}}$. Les auteurs prouvent que le regret au bout d'un horizon n est de la forme

$$\mathbb{E}\left[\text{Regret}_n\right] \leq C\frac{\log(n)}{\Delta}$$

où

$$\Delta = \min_{a\in\mathcal{A}, r(a)<r(a^*)} r(a^*) - r(a)$$

mesure la différence entre le meilleur bras et le meilleur bras sous-optimal et C est une constante indépendante du modèle. La constante α intervenant dans la définition de β_t^a doit être strictement supérieure à 1/2 [Garivier and Moulines, 2008; Bubeck, 2010].

Algorithme 1.8 Algorithme UCB

Entrées: Paramètre $\alpha > 1/2$

Jouer une fois chacun des bras et réception de $R_0, \dots R_{|\mathcal{A}|-1}$.

Pour $t \geq |\mathcal{A}|$ **faire**

Calculer $\hat{r}_t(a)$ pour tout $a \in \mathcal{A}$ d'après (1.20)

Jouer

$$A_t \in \underset{a\in\mathcal{A}}{\text{argmax}} \left\{ \hat{r}_t(a) + \sqrt{\frac{\alpha \log(t)}{N_t(a)}} \right\}$$

Réception de la récompense R_t

fin Pour

[Audibert et al., 2007] ont proposé d'utiliser une estimation de la variance dans le calcul de β_t^a permettant ainsi d'améliorer la performance de l'algorithme lorsque la variance de certains bras sous-optimaux est nettement plus petite que $R_{\max}^2$. L'algorithme proposé, appelé UCB-V, est présenté ci-dessous.

Algorithme 1.9 Algorithme UCB-V

Entrées: Paramètre $\alpha > 1$

Jouer une fois chacun des bras et réception de $R_0, \dots R_{|\mathcal{A}|-1}$.

Pour $t \geq |\mathcal{A}|$ **faire**

Calculer $\hat{r}_t(a)$ défini équation 1.20 et $W_t(a)$ défini, pour tout $a \in \mathcal{A}$, par

$$W_t(a) = \frac{1}{N_t(a)} \sum_{k=0}^{t-1} (R_t \mathbb{1}_{\{A_t=a\}} - \hat{r}_t(a))^2$$

Jouer

$$A_t \in \underset{a\in\mathcal{A}}{\text{argmax}} \left\{ \hat{r}_t(a) + \sqrt{\frac{2\alpha \log(t) W_t(a)}{N_t(a)}} + 3\alpha \frac{\log(t)}{N_t(a)} \right\}$$

Réception de R_t

fin Pour

Une autre problématique a suscité un certain intérêt dans les modèles de bandit. Il s'agit de déterminer l'action optimale en n pas de temps pour un horizon n fixé à l'avance. Cette fois, les récompenses reçues pendant l'interaction n'ont aucune importance, le seul but étant de trouver l'action optimale dans le temps imparti. Le principe des algorithmes répondant à cette problématique est d'éliminer au fur et à mesure toutes les actions qui semblent être moins bonnes que d'autres avec une grande probabilité. Un tel algorithme construit des intervalles

de confiance comme précédemment et élimine une action a si la borne supérieure de son intervalle de confiance est inférieure à une borne inférieure de l'intervalle de confiance associé à une autre action [Even-Dar et al., 2002; Mannor and Tsitsiklis, 2004; Mnih et al., 2008; Bubeck et al., 2009a]. On élimine a si il existe $b \in \mathcal{A}$ tel que

$$\hat{r}_t(a) + \beta_t^a \leq \hat{r}_t(b) - \beta_t^b \ .$$

Pour des MDP

La grande différence entre les modèles de bandit et les processus de décision markovien classiques est la présence d'états et de probabilités de transition entre ces états. Ces probabilités sont, elles aussi, inconnues et doivent donc être estimées par l'agent durant l'interaction avec l'environnement. Cette estimation se fait généralement en comptant les transitions observées. Notons, pour tout état $x \in \mathcal{X}$ et toute action $a \in \mathcal{A}$,

$$N_t(x,a) = \sum_{k=0}^{t-1} \mathbb{1}_{\{X_k=x, A_k=a\}}$$

le nombre de fois, avant l'instant t, où l'action a a été jouée alors que le système était dans l'état x. De manière similaire,

$$N_t(x,a,x') = \sum_{k=0}^{t-1} \mathbb{1}_{\{X_k=x, A_k=a, X_{k+1}=x'\}}$$

désigne le nombre de transitions vers l'état x' alors que l'état était x et que l'agent a joué l'action a. Ainsi, pour tous $x, x' \in \mathcal{X}$ et tout $a \in \mathcal{A}$, on définit l'estimateur de la probabilité de transition $P(x,a;x')$ par

$$\hat{P}_t(x,a;x') = \frac{N_t(x,a,x')}{\max\{N_t(x,a),1\}} \ .$$

On notera $\widehat{\mathbf{M}}_t$ le MDP défini par $(\mathcal{X}, \mathcal{A}, \hat{r}_t, \hat{P}_t)$ où $\hat{r}_t$ est défini par

$$\forall x \in \mathcal{X},\ a \in \mathcal{A},\ \hat{r}_t(x,a) = \frac{\sum_{k=0}^{t-1} R_t \mathbb{1}_{\{X_k=x,\ A_k=a\}}}{\sum_{k=0}^{t-1} \mathbb{1}_{\{X_k=x,\ A_k=a\}}} \ . \tag{1.22}$$

Comme expliqué dans le modèle de bandits, jouer de manière gloutonne par rapport au modèle estimé $\widehat{\mathbf{M}}_t$ peut engendrer de très grandes pertes par manque d'exploration. Illustrons cela sur l'exemple de la rivière (voir paragraphe 1.1.1). Imaginons que, après avoir nagé quelques temps dans les deux sens, l'agent est arrivé sur la rive gauche et reçoit une récompense. A ce moment là, s'il joue de façon gloutonne par rapport au modèle estimé, rien ne le motivera à retourner dans l'eau pour nager à contre-courant jusque l'autre rive puisque il ne sait pas qu'il peut recevoir une récompense beaucoup plus élevée. Il est important de noter qu'un algorithme générique d'apprentissage par renforcement ne suppose pas la structure du MDP connue. L'agent (le nageur) ne sait donc pas comment les états sont liés les uns aux autres et ne peut pas savoir, par exemple, qu'il est impossible de passer en un coup d'une rive à l'autre. Ainsi, même s'il savait que la rive de droite permettait de gagner une très grande récompense, il ne connaît pas la suite d'actions à sélectionner pour l'atteindre. Une manière d'introduire de l'exploration et donc de permettre à l'agent de découvrir les différentes transitions entre les états serait de suivre un algorithme ϵ-glouton qui jouerait avec probabilité ϵ une politique d'exploration et avec probabilité $1-\epsilon$ la politique optimale par rapport au modèle $\widehat{\mathbf{M}}_t$. Cependant, il est difficile de régler la valeur de ϵ pour garantir de bonnes performances

tant théoriques que pratiques.

Plusieurs algorithmes optimistes ont été proposés ces dernières années pour les processus de décision markovien. Ces algorithmes peuvent être classés en deux familles en fonction des garanties théoriques associées. On différenciera les analyses dites probablement approximativement correct (PAC) des approches garantissant des bornes de regret.

Analyse probablement approximativement corrects (PAC)

Les approches dites *probablement approximativement correctes*, ou encore PAC-MDP, fournissent des résultats garantissant qu'avec une grande probabilité l'algorithme joue de manière proche de l'optimale tout le temps sauf un « petit » nombre de fois. Appelons π_t la règle de décision proposée par l'algorithme à l'instant t. Un algorithme est dit PAC-MDP si la valeur de la politique suivie par l'agent dans tous les états visités est proche de la fonction de valeur optimale sauf pour un nombre d'instant polynomial en les caractéristiques du MDP. Plus précisément, on cherche à garantir qu'avec probabilité $1-\delta$

$$V^{\pi_t}(X_t) \geq V^*(X_t) - \epsilon$$

pour tout t sauf un nombre d'instants polynomial en $|\mathcal{X}|$, $|\mathcal{A}|$, l'horizon n, $1/\epsilon$ et $1/\delta$.

L'*algorithme* E^3 [Kearns and Singh, 2002] est un des premiers algorithmes de ce type. Son principe est de maintenir une liste $\mathcal{X}_v$ des états qui ont été visités plus de m fois, où m est fixé à l'avance. Tous les autres états sont regroupés dans ce qui est appelé un état absorbant. Lorsque l'état courant est dans l'état absorbant, l'agent suit une politique d'exploration. Dès que l'état courant appartient à l'ensemble $\mathcal{X}_v$, l'algorithme d'itération sur les valeurs à horizon fini est utilisé pour calculer deux politiques différentes. La première politique $\hat{\pi}$ est optimale par rapport au modèle estimé en considérant l'état absorbant comme un seul état, tandis que la politique d'exploration $\hat{\pi}_E$ est optimale pour un modèle où les récompenses sont mises à 0 partout sauf dans l'état absorbant pour lequel la récompense est maximale. Le choix entre ces deux alternatives dépend de la valeur $V_n^{\hat{\pi}}$ de la politique $\hat{\pi}$ sur un horizon fini n pré-défini. La politique choisie est alors suivie jusqu'à ce que l'état courant appartienne à l'état absorbant. L'*algorithme R-max* [Brafman and Tennenholtz, 2003] est un raffinement de celui-ci permettant de simplifier la gestion du compromis entre exploration et exploitation en utilisant une initialisation optimale des récompenses estimées. [Szita and Lőrincz, 2008] ont proposé un algorithme similaire aux deux précédents, dans lequel l'exploration est également gérée en introduisant un état absorbant x_E associé à une récompense maximale égale à $R_{\max}$. A chaque instant, deux tables états-actions Q^r et Q^e sont calculées de manière indépendantes à partir de $\hat{r}_t$ et $\hat{P}_t$:

$$Q^r_{t+1}(x,a) = \hat{r}_t(x,a) + \gamma \sum_{x'} \hat{P}_t(x,a;x') \max_{a'} Q^r_t(x',a')$$

et

$$Q^e_{t+1}(x,a) = \frac{\hat{P}_t(x,a;x_E) R_{\max}}{1-\gamma} + \gamma \sum_{x'} \hat{P}_t(x,a;x') \max_{a'} Q^e_t(x',a') \ .$$

La première table assigne une valeur à un couple état-action selon le modèle estimé tandis que la deuxième calcule la valeur d'un couple sous un modèle où une récompense maximale serait obtenue pour l'état absorbant et aucune récompense ailleurs. L'action choisie est alors gloutonne par rapport à $Q^r + Q^e$.

Une autre approche PAC-MDP est l'algorithme *Model Based Interval Estimation* proposé par [Strehl and Littman, 2008] qui consiste à construire des intervalles de confiance autour

des récompenses et probabilités de transitions estimées : pour tout état x et toute action a

$$CI_R(t,x,a) = \left[\hat{r}_t(x,a) \pm \frac{C_R}{\sqrt{N_t(x,a)}} \right]$$

et

$$CI_P(t,x,a) = \left\{ P \in \mathbb{S}^{|\mathcal{X}|},\ \left\| P - \hat{P}_t(x,a;.) \right\|_1 \leq \frac{C_P}{\sqrt{N_t(x,a)}} \right\} ,$$

où $\mathbb{S}^{|\mathcal{X}|}$ est l'ensemble des vecteurs de probabilité sur $\mathcal{X}$. Une table état-action $\tilde{Q}$ est alors définie par : pour tout x et tout a,

$$\tilde{Q}(x,a) = \max_{\tilde{r} \in CI_R(t,x,a)} \tilde{r} + \max_{\tilde{P} \in CI_P(t,x,a)} \gamma \sum_{x'} \tilde{P}(x') \max_{a'} \tilde{Q}(x',a') .$$

A chaque instant, l'action choisie est gloutonne par rapport à $\tilde{Q}$. Les auteurs proposent également un algorithme plus simple, ne nécessitant pas de maximisation sur un ensemble de modèles définis par des intervalles de confiance, mais en ajoutant un *bonus d'exploration* de la forme $\frac{\beta}{\sqrt{N_t(x,a)}}$. Dans ce cas, la table état-action $\tilde{Q}$ vérifie

$$\tilde{Q}(x,a) = \hat{r}_t(x,a) + \gamma \sum_{x'} \hat{P}_t(x,a;x') \max_{a'} \tilde{Q}(x',a') + \frac{\beta}{\sqrt{N_t(x,a)}} .$$

Apprentissage par renforcement et minimisation du regret

Certaines approches d'apprentissage par renforcement dans des MDP visent à minimiser le regret cumulé. Inspiré par les travaux de Lai et Robbins, [Burnetas and Katehakis, 1997] ont proposés un tel algorithme. Ils introduisent la définition du regret suivante. Pour toute politique π et tout état initial x_0,

$$\text{Regret}_n^{\pi}(x_0) = \sup_{\pi'} \left(\mathbb{E}^{\pi'} \left[\sum_{k=0}^{n} R_k \,\middle|\, X_0 = x_0 \right] \right) - \mathbb{E}^{\pi} \left[\sum_{k=0}^{n} R_k \,\middle|\, X_0 = x_0 \right] .$$

Leur algorithme mémorise l'ensemble des actions suffisamment jouées pour un état donné

$$D_t(x) = \{ a \in \mathcal{A}, N_t(x,a) \geq \log^2 N_t(x) \}$$

et calcule le gain η_t et un biais h_t optimaux pour le modèle estimé restreint à cet ensemble d'actions. Ils proposent ensuite de calculer un indice optimiste pour chaque couple état-action

$$U_t(x,a) = \max_{q, KL(\hat{P}_t(x,a;.),q) \leq \frac{C_P}{N_t(x,a)}} \sum_{x'} q(x') h_t(x') ,$$

où $KL(p;q)$ désigne la divergence de Kullback-Leibler entre p et q. L'action sélectionnée est alors soit une des actions qui doivent être jouées pour rester dans l'ensemble $D_t(x)$, soit l'action qui maximise l'index $U_t(x_t,a)$. Dans cet algorithme, l'exploration est gérée à la fois par l'optimisme de l'indice et par la sélection d'actions sous-échantillonnées. Les auteurs prouvent que le regret de leur algorithme est asymptotiquement logarithmique :

$$\forall x_0 \in \mathcal{X},\ \limsup_{n \to \infty} \frac{\text{Regret}_n^{\pi_{BT}}(x_0)}{\log n} \leq C(\mathbf{M}) ,$$

où π_{BT} est la politique suivie par l'algorithme. De plus, ils fournissent une borne inférieure du regret pour toute politique π telle que, pour tout $\alpha > 0$, $\text{Regret}_n^\pi(x_0) = o(n^\alpha)$ quand n tend vers l'infini :

$$\forall x_0 \in \mathcal{X},\ \liminf_{n\to\infty} \frac{\text{Regret}_n^\pi(x_0)}{\log n} \geq C(\mathbf{M}) \ .$$

La constante $C(\mathbf{M})$ dépend du modèle. L'algorithme *Optimistic Linear Programming* proposé [Tewari and Bartlett, 2008] est similaire à celui de [Burnetas and Katehakis, 1997] à ceci près que la divergence de Kullback-Leibler est remplacée par une norme L^1. Les auteurs reformulent et simplifient considérablement les preuves des bornes de regret grâce à ce changement de métrique. Notons que, contrairement à l'algorithme *Model Based Interval Estimation* (MBIE), le critère utilisé ici est le critère moyen (voir [Tewari and Bartlett, 2007] pour des explications théoriques du lien entre le critère moyen et le critère pondéré pour les méthodes d'estimation de paramètres avec des intervalles).

Les travaux de [Auer and Ortner, 2007; Jaksch et al., 2010; Bartlett and Tewari, 2009] introduisent des algorithmes qui garantissent des regrets logarithmiques non-asymptotiques pour une grande classe de MDP. Dans ces travaux, l'ensemble des modèles compatibles avec les observations sont définis comme pour l'algorithme MBIE par :

$$\mathcal{M}_t = \left\{ \mathbf{M} = (\mathcal{X}, \mathcal{A}, P, r) : \forall x \in \mathcal{X}, \forall a \in \mathcal{A},\ \hat{r}_t(x,a) - r(x,a)| \leq \frac{C_R}{\sqrt{N_t(x,a)}} \right.$$
$$\left. \text{et } \left\| \hat{P}_t(.;x,a) - P(.;x,a)) \right\|_1 \leq \frac{C_P}{\sqrt{N_t(x,a)}} \right\} \ .$$

La politique choisie est $\arg\max_\pi \{\eta^\pi(\mathbf{M}) : \mathbf{M} \in \mathcal{M}_t\}$ où $\eta^\pi(\mathbf{M})$ est la récompense moyenne reçue en suivant la politique π dans le MDP $\mathbf{M}$. Pour éviter d'alterner continûment entre différentes politiques sans pouvoir les évaluer, les auteurs proposent de n'en changer que lorsque le diamètre des intervalles de confiance a considérablement diminué [Auer and Ortner, 2007] ou lorsque le nombre de visites aux différents couples état-action en suivant chacune des politiques dépassent un seuil pré-fixé [Auer et al., 2009a]. La définition du regret est, dans ce cas, sensiblement différente. Les récompenses accumulées par l'algorithme sont comparées avec celles qui seraient reçues, en moyenne, par un agent jouant une politique optimale. Le *regret* de l'algorithme après n instants est défini par :

$$\text{Regret}_n = \sum_{t=0}^{n} \left(\eta^*(\mathbf{M}) - R_t\right) \ ,$$

où $\eta^*(\mathbf{M})$ est la récompense moyenne optimale dans le vrai MDP $\mathbf{M}$, inconnu de l'agent. Dans ces articles, il est prouvé que l'espérance du regret est bornée par un terme logarithmique en n et polynomial en $|\mathcal{A}|$, $|\mathcal{X}|$:

$$\mathbb{E}\left[\text{Regret}_n\right] \leq C(\mathbf{M}) \frac{|\mathcal{X}|^2 |\mathcal{A}| \log(n)}{\Delta}$$

où $\Delta = \eta^*(\mathbf{M}) - \max_{\pi:\eta^\pi(\mathbf{M})<\eta^*(\mathbf{M})} \eta^\pi(\mathbf{M})$ avec $C(\mathbf{M})$ une constante dépendant du modèle.

Élimination d'actions sous-optimales

Comme pour le modèle de bandit, [Even-Dar et al., 2006] ont proposé un algorithme basé sur l'élimination d'actions sous-optimales dans le cadre des MDP. L'idée est de maintenir des bornes supérieure et inférieure de la table Q estimée. A chaque instant $\overline{Q}(s,a)$ et $\underline{Q}(s,a)$ sont calculés en utilisant les estimées des vecteurs de récompenses et des probabilités de transition.

Lorsque $\overline{Q}(s,a) \leq \max_{a'} \underline{Q}(s,a')$, l'action a dans l'état s est éliminée. De proche en proche, les actions sous-optimales sont ainsi éliminées. Un critère d'arrêt de la phase d'exploration basé sur $\left\|\overline{Q}(s,a) - \underline{Q}(s,a)\right\|_\infty$ est proposé. Ensuite, la politique suivie est gloutonne par rapport à $\underline{Q}$. Il est prouvé que cette politique ainsi obtenue est ϵ-optimale avec grande probabilité.

1.4 POMDP

Une extension des processus de décision markoviens, très utile en pratique, est de considérer que l'agent n'observe pas (ou pas entièrement) l'état de l'environnement. Les processus de décision markoviens partiellement observés (POMDP) modélisent l'interaction dans ce cas. Dans cette section, nous présentons ce modèle et expliquons brièvement les différentes approches de la littérature pour sélectionner les actions dans un tel cadre.

1.4.1 Définitions

Un POMDP est un processus de décision markovien dans lequel l'agent n'observe pas l'état de l'environnement. A la place, à chaque instant t, il perçoit une observation Y_t qui dépend de l'état du système X_t, et, parfois, de l'action effectuée précédemment A_{t-1}. On note $(Y_t)_{t\geq 0}$ le processus des observations et $\mathcal{Y}$ l'ensemble des observations, que l'on supposera fini. On définit la probabilité d'observation $G : \mathcal{A} \times \mathcal{X} \times \mathcal{Y} \to [0;1]$ telle que pour tout $t \geq 0$, pour toute observation $y \in \mathcal{Y}$,

$$G(A_{t-1}, X_t; y) = \mathbb{P}(Y_t = y | A_{t-1}, X_t). \tag{1.23}$$

Un POMDP est défini par $(\mathcal{X}, \mathcal{A}, P, \mathcal{R}, \mathcal{Y}, G, B_0)$ où l'espace des états $\mathcal{X}$, l'espace des actions $\mathcal{A}$, le noyau de transition des états contrôlés par les actions P, et la loi des récompenses $\mathcal{R}$ sont déterminés comme dans un MDP (voir section 1.1). De plus,

- $\mathcal{Y}$ est l'espace des observations,
- la probabilité d'émission G est définie dans l'équation (1.23) ci-dessus,
- $B_0 : \mathcal{X} \to [0,1]$ est la distribution de probabilité initiale sur les états.

Reprenons les exemples du paragraphe 1.1.1 et modifions les sensiblement pour rendre l'état non observable.

Dans l'exemple de la *rivière*, supposons que le nageur ne sait pas exactement où il se trouve : il sait seulement si il est sur une rive ou dans l'eau. Il y a donc une confusion possible entre les états $\{1,6\}$ s'il est sur une rive et entre les états $\{2,3,4,5\}$ s'il est dans l'eau. Ce premier exemple illustre le phénomène appelé « perceptual aliasing » (voir [Chrisman, 1992; Shani, 2004]) qui a lieu lorsqu'une même observation peut être obtenue alors que l'environnement est dans des états différents. La séquence des observations et des actions passées permettent souvent de deviner l'état sous-jacent. Dans l'exemple considéré, on remarque que si le nageur se trouve sur une rive et qu'il vient de nager vers la gauche, alors il sait qu'il se trouve sur la rive de gauche. Notons que l'agent ne peut faire cette déduction que s'il connaît la structure du MDP. Si le nageur est dans la rivière et qu'il nage vers la droite depuis un certain temps, même s'il connaît cette structure, il ne peut pas déterminer sa position dans la rivière puisque la transition entre les états est aléatoire.

Dans l'exemple de *gestion de stock*, on peut imaginer un modèle où le vendeur n'observe pas l'état de son stock tous les jours. Par contre, il pourrait disposer, par exemple, d'une information, potentiellement bruitée, sur le nombre de fruits vendus la veille. Selon la fonction aléatoire qui relie l'observation perçue par l'agent et le vrai nombre de cageots présents dans le magasin, il est possible ou non de déterminer exactement l'état sous-jacent. Une approche qui peut être pertinente dans un tel cas est de ne pas essayer de deviner l'état sous-jacent

mais plutôt de décider le nombre de cageots à acheter directement en fonction de l'observation disponible. Une autre variante de cet exemple serait de considérer que le marchand doit payer une certaine somme d'argent afin de pouvoir consulter l'état de son stock. Dans ce cas, une action supplémentaire est introduite et le marchand a alors le choix entre payer pour obtenir des informations ou acheter des provisions de cageot en n'ayant qu'une indication bruitée de l'état de son stock.

Ces deux exemples donnent une idée de la complexité et de la diversité des problématiques modélisées par des POMDP. C'est ce qui rend très difficile la recherche d'une méthode générique pour déterminer une politique d'action optimale. Lorsque le modèle est entièrement connu de l'agent, il existe quelques méthodes de planification pour des POMDP quelconques. Mais, en apprentissage par renforcement, cette généralité n'est plus possible, les approches proposées dans la littérature sont alors assez variées et spécifiques à des POMDP ayant des structures particulières.

1.4.2 Etat interne

Dans un POMDP, l'agent ne connaissant pas l'état de l'environnement, il lui est nécessaire de rassembler l'ensemble des informations auquel il a accès afin de déterminer les actions à effectuer. L'information utile à chaque instant est généralement appelée *état interne* ou *état d'information*.

Un exemple simple d'état d'information est l'*état d'information complet*, noté I_t^C, constitué de toutes les données connues [Hauskrecht, 2000] : $I_t^C = \{Y_{0:t}, A_{0:t-1}, B_0\}$. On remarque que, pour tout t, l'état d'information complet est l'image par une fonction déterministe de l'état d'information précédent, de l'action et de l'observation courante :

$$I_{t+1}^C = \tau(I_t^C, Y_{t+1}, A_t) \ .$$

Bien que cet état résume à la perfection toute l'information passée, il a un grand inconvénient : sa dimension augmente de manière linéaire avec le temps. Dès que l'on s'intéresse à des problèmes à horizon très grand, voir infini, cette variable n'est donc plus manipulable.

On définit alors la notion d'*information exhaustive*. Un processus d'état d'information $(I_t)_t$ est dit être un processus d'information exhaustif si il vérifie les conditions suivantes :

- il existe une fonction déterministe τ telle que $I_{t+1} = \tau(I_t, Y_{t+1}, A_t)$,
- pour toute fonction $f \geq 0$,

$$\mathbb{E}\left[f(X_t) \,|\, I_0, Y_{0:t}, A_{0:t-1} \right] = \mathbb{E}\left[f(X_t) \,|\, I_t \right] \ . \tag{1.24}$$

Un état d'information exhaustif I_t est une fonction déterministe de l'état d'information précédent, de l'action précédente et de l'observation courante. Pour toute fonction $g \geq 0$, on a alors

$$\mathbb{E}\left[g(Y_{t+1}) \,|\, I_0, Y_{0:t}, A_{0:t} \right] = \mathbb{E}\left[g(Y_{t+1}) \,|\, I_t, A_t \right] \ .$$

Cet état interne préserve donc les informations nécessaires concernant la probabilité d'occupation, c'est-à-dire la probabilité de l'état (caché) de l'environnement (voir équation (1.24)), et la probabilité des observations connaissant les états.

Nous représentons sur la figure 1.4 le graphe de dépendance des différentes variables lorsque l'état interne est exhaustif et que la politique est une fonction aléatoire (pas forcément stationnaire) de l'état interne :

$$\pi_t : \mathcal{I} \times \mathcal{A} \to [0; 1] \ .$$

A chaque instant t, l'agent perçoit une observation Y_t qui dépend de l'action A_{t-1} effectuée à l'instant précédent et de l'état caché de l'environnement X_t. A partir de cette nouvelle

observation, de l'action A_{t-1} et de l'état interne à l'instant précédent I_{t-1}, l'agent calcule son état interne I_t ce qui lui permet de résumer l'information dont il dispose à propos de l'environnement ainsi que de choisir la nouvelle action A_t en suivant la politique π_t.

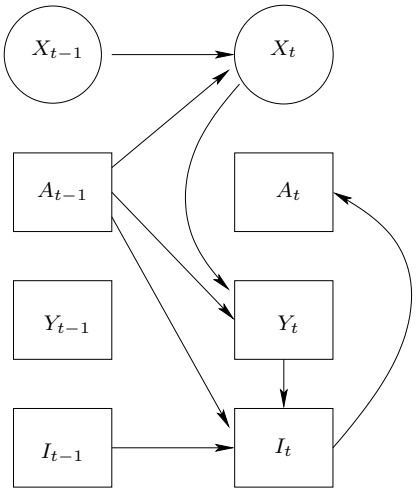

FIGURE 1.4 – Graphe de dépendance pour un état interne exhaustif..

Un résultat important dans la théorie des POMDP est que le processus d'état d'information exhaustif $(I_t)_{t\geq 0}$ est un processus de Markov contrôlé par le processus des actions $(A_t)_{t\geq 0}$. En effet, l'équation 1.24 ainsi que le fait que $I_{t+1} = \tau(I_t, Y_{t+1}, A_t)$ impliquent que, pour toute fonction f positive,

$$\mathbb{E}\left[f(I_{t+1}) \,|\, I_0, Y_{0:t}, A_{0:t}\right] = \mathbb{E}\left[f(I_{t+1}) \,|\, I_t, A_t\right] \ .$$

Le POMDP peut alors être transformé en un MDP où les états sont les états internes I_t et les actions sont les mêmes que pour le POMDP initial. On l'appelle *processus de décision markovien d'état d'information*. Soit Ψ le noyau de transition de la chaîne de Markov contrôlée $(I_t)_{t\geq 0}$ défini pour toute fonction f positive réelle et tout t par

$$\mathbb{E}\left[f(I_{t+1}) \,|\, A_{0:t}, I_{0:t}\right] = \Psi(I_t, A_t; f) \ .$$

Soit $\mathcal{I}$ l'ensemble des états d'information exhaustifs. Pour toute politique stationnaire déterministe π telle que $\pi : \mathcal{I} \to \mathcal{A}$, on définit une fonction de valeur pour les états d'information exhaustifs :

$$V_\gamma^\pi(I) = \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t r(X_t, A_t) \,\middle|\, I_0 = I\right] \ .$$

On peut écrire :

$$\begin{aligned} V_\gamma^\pi(I) &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}^\pi\left[r(X_t, A_t) \,|\, I_0 = I\right] \\ &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}^\pi\left[\mathbb{E}\left[r(X_t, A_t) \,|\, \mathcal{F}_t\right] \,|\, I_0 = I\right] \\ &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}^\pi\left[\sum_{x\in\mathcal{X}} \mathbb{P}\left[X_t = x \,|\, I_t\right] r(x, \pi(I_t)) \,\middle|\, I_0 = I\right] \end{aligned}$$

En posant $\rho(I_t, A_t) \stackrel{\text{def}}{=} \sum_{x\in\mathcal{X}} \mathbb{P}\left[X_t = x \,|\, I_t\right] r(x, A_t)$, l'espérance de la récompense reçue en jouant l'action A_t sachant que l'état interne vaut I_t, on a

$$V_\gamma^\pi(I) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t \rho(I_t, \pi(I_t)) \,\middle|\, I_0 = I \right] .$$

Soit Π^{ISD} l'ensemble des politiques stationnaires déterministes π telles que $\pi : \mathcal{I} \to \mathcal{A}$. On définit la fonction de valeur optimale :

$$V_\gamma^*(I) = \sup_{\pi\in\Pi^{ISD}} \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t \rho(I_t, \pi(I_t)) \,\middle|\, I_0 = I \right] \tag{1.25}$$

Celle-ci vérifie l'équation de Bellman :

$$V^*(I) = \max_{a\in\mathcal{A}} \left\{ \rho(I, a) + \gamma \Psi(I, a; V^*) \right\} .$$

En utilisant la dépendance déterministe de l'état interne par rapport à l'action A_t et à l'observation Y_{t+1}, l'équation de Bellman s'écrit de la manière suivante :

$$V^*(I) = \max_{a\in\mathcal{A}} \left\{ \rho(I, a) + \gamma \sum_{y'\in\mathcal{Y}} V^* \circ \tau(I, y', a) \mathbb{P}\left[Y_{t+1} = y' \,\middle|\, I_t = I, A_t = A \right] \right\} .$$

Une politique déterministe stationnaire optimale $\pi^* : \mathcal{I} \to \mathcal{A}$ peut alors être déduite de l'équation précédente : pour tout $I \in \mathcal{I}$

$$\pi^*(I) = \operatorname*{argmax}_{a\in\mathcal{A}} \left\{ \rho(I, a) + \gamma \Psi(I, a; V^*) \right\} \tag{1.26}$$

Un état d'information particulier est l'*état de croyance* (*belief state* en anglais) introduit par [Astrom, 1965]. Cet état de croyance est défini comme étant la probabilité d'être dans l'état x connaissant les observations et les actions passées [Astrom, 1965; Kaelbling et al., 1996] :

$$B_t(x) = \mathbb{P}(X_t = x | \mathcal{F}_t) ,$$

où $\mathcal{F}_t = \{Y_{0:t}, A_{0:t-1}, B_0\}$. L'ensemble des valeurs prises par B_t est l'espace des probabilités sur $\mathcal{X}$. L'espace d'état $\mathcal{X}$ étant supposé fini, il s'agit du simplexe des probabilités sur l'espace de cardinal $|\mathcal{X}|$, que l'on notera $\mathbb{S}^{|\mathcal{X}|}$. On peut facilement montrer que

$$B_{t+1}(X_{t+1}) = \tau(B_t, Y_{t+1}, A_t) \tag{1.27}$$

$$\stackrel{\text{def}}{=} \frac{\sum_{X_t\in\mathcal{X}} G(A_t, X_{t+1}; Y_{t+1}) P(X_t, A_t; X_{t+1}) B_t(X_t)}{\sum_{x'\in\mathcal{X}} \sum_{x\in\mathcal{X}} G(A_t, x'; Y_{t+1}) P(x, A_t; x') B_t(x)} , \tag{1.28}$$

que l'on reconnaît comme étant la formule de mise à jour récursive de filtre dans un modèle de Markov caché [Cappé et al., 2005]. Ainsi, l'état de croyance est un état interne exhaustif ; c'est l'état interne le plus utilisé dans la littérature.

1.4.3 Planification

Les deux paragraphes suivants sont dédiés à l'énumération de travaux de recherche en planification ou en apprentissage par renforcement dans des POMDP. Ces tâches étant complexes, de nombreuses approches très différentes ont été proposées dans la littérature. Une étude détaillée de toutes ces approches dépasse le cadre de cette thèse et cette liste est donc loin d'être exhaustive.

Un POMDP pouvant être ramené à un MDP à l'aide de l'état de croyance, on peut penser utiliser des méthodes classiques similaires à celles présentées dans le paragraphe 1.2 pour déterminer la politique optimale. Cependant, l'espace des états de croyance est un espace continu et la recherche de la politique optimale dans ce cas est un problème difficile [Hauskrecht, 2000]. Des méthodes d'approximation de la fonction de valeur utilisant les propriétés particulières de cette fonction dans un MDP à état de croyance ont été proposées par [Smallwood and Sondik, 1973; Sondik, 1978; Kaelbling et al., 1996; Cassandra et al., 1997]. Ces méthodes dites de « résolution exactes » optimisent la fonction de valeur sur l'ensemble des états de croyance. D'autres méthodes d'approximation proposent de calculer ces fonctions de valeur seulement en certains points appelés les *points de croyance*. Ces points peuvent être fixés à l'avance ou évoluer au cours du temps [Aberdeen, 2003b; Bonet, 2002; Poupart and Boutilier, 2004b; Pineau et al., 2006; Hsu et al., 2007]. Contrairement à ces méthodes qui cherchent la politique optimale d'un POMDP à partir des fonctions de valeur, certains travaux proposent des approches qui recherchent directement dans l'espace des politiques. Ce sont des méthodes du type de l'algorithme d'itération sur les politiques [Sondik, 1978; Hansen, 1998; Meuleau et al., 1999; Aberdeen, 2003a; Poupart and Boutilier, 2004a].

1.4.4 Apprentissage par renforcement

Il est très difficile de proposer des méthodes génériques d'apprentissage par renforcement dans un POMDP. On pourrait penser à utiliser la réécriture du modèle sous la forme d'un MDP à état de croyance comme proposé pour les méthodes de planification. Cependant, il est important de noter que l'état interne de croyance dépend inéluctablement du modèle. En effet, le calcul de l'état de croyance à chaque instant en fonction de l'état à l'instant précédent nécessite la connaissance de P et de G (voir équation (1.28)). Lorsque ces probabilités ne sont pas connues, ce qui est le cas en apprentissage par renforcement, il n'est pas possible de se ramener à un MDP à état continu.

Certains algorithmes utilisent l'espace des observations à la place de l'espace des états. Il s'agit de considérer des politiques qui sont des fonctions de l'observation courante $\pi : \mathcal{Y} \rightarrow \mathcal{A}$. Les méthodes habituelles d'apprentissage par renforcement présentées dans le cadre des MDP (voir paragraphe 1.2) telles que le Q-learning ou SARSA ont alors été utilisées. La fonction de valeur observation-action $Q(y, a)$ est calculée au lieu de la fonction de valeur état-action $Q(x, a)$ [Singh et al., 1994; Jaakkola et al., 1995]. Bien évidemment, ces algorithmes ne convergent pas nécessairement vers une politique optimale puisque la suite des observations conditionnellement aux actions n'est en général pas une chaîne de Markov.

Le problème d'apprentissage par renforcement dans un POMDP général paraissant difficilement soluble, des travaux ont été effectués dans des cas particuliers de POMDP. Par exemple, le problème de « perceptual aliasing » peut être résolu en utilisant une mémoire interne contenant les actions et observations passées sur un horizon fini [Shani, 2004]. Ces méthodes partent de l'hypothèse que l'utilisation d'une suite d'observations, d'actions et de récompenses passées permettent à l'agent de déterminer l'état sous-jacent de l'environnement. Plusieurs méthodes utilisent un arbre représentant l'histoire des actions effectuées et des observations perçues ; une action est alors associée à chaque racine de l'arbre [McCallum, 1996; Meuleau et al., 1999; Dutech and Samuelides, 2003].

A l'instar des méthodes d'apprentissage par renforcement dans les MDP, on peut différencier des méthodes *« model-free »* et des méthodes *« model-based »*. Ces dernières consistent à apprendre le modèle du POMDP durant l'interaction et d'en déduire une politique optimale. Deux types d'approches ont principalement été utilisées dans la littérature. Certains se basent sur la théorie des chaînes de Markov cachées pour prédire l'observation étant donné la suite des observations et actions passées [Chrisman, 1992; McCallum, 1996; Aberdeen, 2003a] tandis que d'autres proposent de déterminer un modèle du POMDP à l'aide d'un arbre de suffixe

[McCallum, 1996; Shani et al., 2005; Brafman et al., 2005]. Des travaux ont également proposé d'utiliser des méthodes d'échantillonage préférentiel pour représenter l'état de croyance et des approximations Monte-Carlo pour évaluer sa propagation [Thrun, 2000].

Une autre technique est de considérer une classe de politique paramétrée $\{\pi_\theta, \theta \in \mathbb{R}^K\}$ et de chercher la politique optimale parmi les éléments de cette classe. Le paramètre θ peut être déterminé de différentes façon, notamment en utilisant des méthodes de descente de gradient [Baxter and Bartlett, 2001; Baxter et al., 2001; Aberdeen and Baxter, 2002; Greensmith et al., 2004].

1.5 Conclusion

Les modèles d'interaction utilisés en apprentissage par renforcement sont les MDP, POMDP et le modèle de bandit. Les équations d'optimalité déterminent les politiques optimales lorsque le modèle d'interaction est connu de l'agent. Les algorithmes d'itération sur les valeurs et d'itération sur les politiques permettent de résoudre cette tâche de planification dans des MDP à espaces d'état et d'action finis. La tâche de planification est plus complexe lorsque l'état n'est que partiellement observé par l'agent, la notion d'état interne est alors utilisée pour la recherche de politique dans ce cadre.

Deux familles d'approches d'apprentissage par renforcement peuvent être distinguées : les approches « model-free » et « model-based ». Dans la suite de cette thèse, nous nous intéresserons à des méthodes « model-based ». Le modèle d'interaction particulier que nous considérons dans la chapitre 2 dépendant d'un petit nombre de paramètres, il paraît pertinent d'estimer les paramètres pour déterminer les politiques optimales dans ce cadre. Nous fournissons donc un algorithme d'apprentissage par renforcement « model-based » et continuons alors à suivre ce type d'approche par la suite.

1.

Dans les algorithmes d'apprentissage par renforcement existant, l'équilibre entre l'exploration et l'exploitation est contrôlé de différentes manières. Les algorithmes que nous proposons dans la suite suivent deux approches distinctes. L'algorithme présenté dans le chapitre 2 consiste à diviser l'interaction en deux phases successives, la première étant dédiée à l'exploration et à l'estimation des paramètres tandis que, dans la deuxième, l'agent suit une politique d'exploitation. Dans ce cas, il est crucial de déterminer au mieux la longueur de la phase d'exploration. Dans les chapitres 3 et 4, nous proposons de suivre des approches optimistes. L'exploration est alors implicite et dépend de la largeur des intervalles de confiance.

Pour analyser la performance des algorithmes que nous fournissons, nous proposons de calculer le regret. Ce critère présente en effet les garanties de performance les plus fortes. Cependant, pour pouvoir faire une telle analyse, certaines hypothèses restrictives sur le modèle sont nécessaires. Cela nécessite par exemple de supposer que l'espace d'état est fini et que la récompense est bornée. Ces hypothèses n'étant pas gênantes dans les modèles qui nous intéressent ici, nous pouvons fournir des bornes du regret des algorithmes que nous exposons.

CHAPITRE 2

Apprentissage par renforcement dans un modèle d'écoute de canal

Dans ce chapitre, nous considérons un processus de décision markovien partiellement observé (POMDP) assez particulier dans lequel l'agent a la possibilité d'observer l'état de l'environnement, ou une partie de celui-ci de son choix, en sélectionnant une action spécifique. Plus précisément, nous nous intéressons à un modèle où l'état est un vecteur de dimension N et l'agent peut choisir quelles composantes de l'état il souhaite observer à chaque instant. Ce modèle permet de considérer une application d'intérêt dans le domaine de la radio cognitive. Il s'agit de l'accès opportuniste à un réseau de communication par un utilisateur secondaire. Dans ce chapitre, nous avons choisi de présenter notre recherche en se focalisant sur ce modèle applicatif. Néanmoins, l'algorithme original que nous proposons dans ce cadre pour gérer de manière adaptative le compromis entre exploration et exploitation pourrait être étendu à d'autres POMDP.

2.1 Introduction

L'accès opportuniste aux ressources spectrales pour la radio cognitive a été l'objet de nombreuses recherches ces dernières années [Akyildiz et al., 2008; Haykin, 2005; Mitola, 2000]. Dans la radio à bandes licenciées, les ressources spectrales sont divisées en bandes de fréquences, aussi appelées canaux, attribuées par licence à des utilisateurs. Chaque canal est donc réservé à un utilisateur fixe. Le nombre d'utilisateurs et le besoin en ressources spectrales étant en constante croissance ces dernières années, un besoin de nouveau système d'allocation des ressources se fait ressentir. Des études ont montré que l'utilisation de ces ressources varie très fortement d'un instant à l'autre et d'un canal à l'autre [Force, 2002; Zhao and Sadler, 2007].

L'idée originale de la radio cognitive est celle d'un système de communication intelligent qui détecterait les besoins des utilisateurs et fournirait les ressources radio et les services sans fil les plus appropriés en fonction des ressources disponibles [Mitola III and Maguire Jr, 1999]. La radio cognitive propose d'optimiser l'utilisation du spectre en exploitant de manière ingénieuse la grande portion de bandes de fréquences inutilisée à chaque instant. Le but est de partager les bandes de fréquences attribuées (par licence) à des *utilisateurs primaires* avec d'autres utilisateurs, qui ne possèdent pas de licence. Ces seconds utilisateurs sont appelés *utilisateurs secondaires* ou *utilisateurs cognitifs*. En radio cognitive, ces derniers

identifient prudemment les ressources spectrales disponibles afin de communiquer en évitant de perturber le réseau primaire. Cet accès opportuniste aux ressources spectrales permet donc potentiellement d'améliorer de manière très significative l'efficacité du réseau.

La communication opportuniste sur bandes licenciées peut être modélisée par un processus de décision markovien partiellement observé (POMDP). Pour cause de limitations techniques et étant donné le coût énergétique de la surveillance du spectre, on admet que l'utilisateur secondaire ne peut pas observer l'état de toutes les bandes de fréquences simultanément [Lai et al., 2008; Liu and Zhao, 2008; Zhao et al., 2008]. Il doit alors sélectionner de manière adéquate un ensemble de canaux, dont il observera la disponibilité. Nous nous intéressons dans ce travail à la *politique d'écoute* des canaux que l'utilisateur secondaire suit pour déterminer quels canaux observer à chaque instant. Dans un premier temps, nous nous limiterons au cas où l'utilisateur secondaire connaît les informations statistiques concernant le trafic des utilisateurs primaires. Nous suivrons une approche, similaire à celle proposée par [Bonet, 2002] pour les POMDP généraux, adaptée aux propriétés particulières du modèle de communication opportuniste. L'algorithme proposé permet de trouver une stratégie proche de la politique optimale dont la performance théorique peut être étudiée. Néanmoins, la complexité de cet algorithme croît de manière quadratique avec le nombre de canaux. Nous considérerons donc également des simplifications du problème initial qui peuvent être introduites afin de déterminer des politiques d'écoute proches de l'optimale même lorsque le nombre de canaux considérés est grand.

En pratique, les informations statistiques concernant le trafic dans le réseau primaire ne sont pas connues à l'avance par l'utilisateur secondaire. Celui-ci doit alors les estimer avant de rechercher la politique d'écoute optimale. Cette approche plus réaliste s'apparente à un problème d'apprentissage par renforcement dans un POMDP. Nous proposons un algorithme composé d'une première phase d'estimation des paramètres et d'une deuxième phase durant laquelle l'utilisateur secondaire suit la politique optimale pour les paramètres estimés. Cet algorithme original permet de gérer de manière optimale le compromis entre exploration et exploitation qui, dans ce cadre, se traduit par la détermination de la durée de la phase d'estimation. En plus de résultats numériques encourageants, nous fournissons des garanties de performance théorique de l'algorithme en termes de borne de l'espérance du regret. Celles-ci sont similaires à celles connues jusqu'à présent dans le cas des modèles de bandits ou de MDP à espaces d'états et d'actions finis (voir section 1.3.2).

Ce chapitre est organisé de la manière suivante. Le modèle d'allocation de ressources spectrales, appelé dans la suite *modèle d'allocation de canal*, est décrit dans la section 2.2. Dans la section 2.3, plusieurs approches pour résoudre le problème de planification sont présentées. L'algorithme original d'apprentissage par renforcement que nous proposons est explicité dans la section 2.4. Cet algorithme est ensuite appliqué à un modèle d'écoute à un seul canal puis à un modèle d'allocation de canal lorsque les statistiques d'utilisation des utilisateurs primaires sont identiques.

2.2 Modèle d'allocation de canal

Le modèle de communication opportuniste que nous considérons consiste en N canaux indépendants dont l'occupation varie dans le temps. Ces N canaux sont attribués par licence à un réseau primaire composé d'utilisateurs qui communiquent selon une structure à trames temporelles. On notera $X_t(i)$ l'état du canal i au début de chaque intervalle de temps $t \in \mathbb{N}$. $X_t(i)$ est égal à 0 quand le canal i est occupé et à 1 quand le canal est libre (voir figure 2.1). On suppose que la disponibilité du canal évolue de manière markovienne : l'état du canal i au début de chaque trame t dépend uniquement de l'état du canal au début de la trame $t-1$.

Considérons maintenant un utilisateur secondaire recherchant des opportunités pour trans-

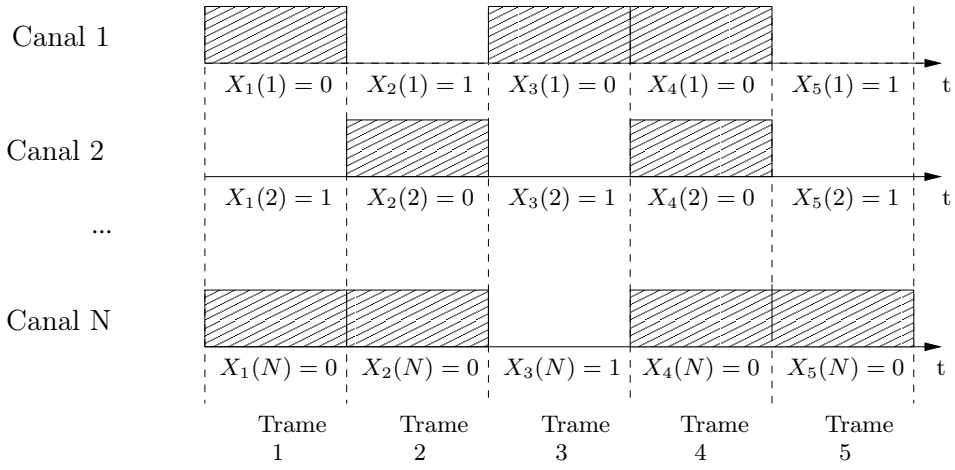

FIGURE 2.1 – Le réseau primaire

mettre ses données dans les canaux temporairement inutilisés par les utilisateurs primaires. Nous nous intéressons à un modèle avec un seul utilisateur secondaire. Des cadres similaires avec plusieurs utilisateurs secondaires ont également été considérés : en particulier, un algorithme pour un modèle décentralisé dans le cas où les canaux sont sans mémoire a été introduit récemment par [Liu and Zhao, 2010a]. De plus, certaines recherches ont considéré un cas où l'observation de l'état n'est pas parfaite [Chen et al., 2008]. L'utilisateur secondaire ne pouvant observer à chaque instant l'ensemble des canaux, une de ses tâches principales consiste à choisir quels canaux observer à chaque intervalle de temps afin de transmettre le maximum de données. Dans ce modèle, l'accès au canal peut être interprété comme une tâche de planification dans une classe particulière de processus de décision markoviens partiellement observés appelés « restless bandits » [Liu and Zhao, 2008; Zhao et al., 2007b].

2.2.1 Modélisation par un POMDP

Le *modèle d'allocation de canal* peut être interprété comme un processus de décision markovien partiellement observé (POMDP)(voir paragraphe 1.4 pour une définition des POMDP). Au début de l'intervalle de temps t, l'état du réseau est $X_t = [X_t(1), \ldots, X_t(N)]' \in \mathcal{X} \stackrel{\text{def}}{=} \{0,1\}^N$. Les états de différents canaux sont supposés indépendants : pour $i \neq j$, les variables aléatoires $X_t(i)$ et $X_t(j)$ sont indépendantes. L'utilisateur secondaire sélectionne un ensemble de L canaux à observer. Ce choix correspond à une action $A_t = [A_t(1), \ldots, A_t(N)]'$, où $A_t(i) = 1$ si l'agent choisit d'observer le i-ème canal et $A_t(i) = 0$ sinon. Puisque L canaux sont observés à chaque instant, $\sum_{i=1}^{N} A_t(i) = L$. On note $\mathcal{A}$ l'ensemble des actions possibles. L'observation est un vecteur $Y_t = [Y_t(1), \ldots, Y_t(N)]'$ tel que, pour tout $i \in \{1, \ldots, N\}$

$$Y_t(i) = \begin{cases} X_t(i) & \text{si } A_t(i) = 1 \\ \emptyset & \text{sinon} \end{cases} \tag{2.1}$$

où $\emptyset$ est l'ensemble vide. L'utilisateur secondaire communique alors dans les canaux libres parmi les L canaux observés et reçoit une récompense R_t à la fin de l'intervalle de temps t. Cette récompense dépend de l'action A_t, de l'observation Y_t et de la disponibilité des canaux observés.

Au début de l'intervalle de temps suivant, l'état du canal est X_{t+1}. Ce nouvel état dépend uniquement de l'état X_t des canaux à l'instant précédent. Les actions prises par l'utilisateur secondaire n'ont aucun impact sur le réseau primaire. Si l'utilisateur secondaire a choisi de

transmettre des informations dans le canal i durant l'intervalle de temps t, ce choix n'aura aucune incidence sur l'état du réseau primaire au début de l'intervalle de temps $t+1$. Par abus de langage, on parlera dans la suite d'instant t pour désigner l'intervalle de temps t.

Supposons que les probabilités de transition des états de tous les canaux sont connues. En particulier, notons $\alpha(i)$ la probabilité que le canal i passe de l'état 0 (occupé) à l'état 1 (libre) et $\beta(i)$ la probabilité que le canal i reste dans l'état 1 (voir figure 2.2).

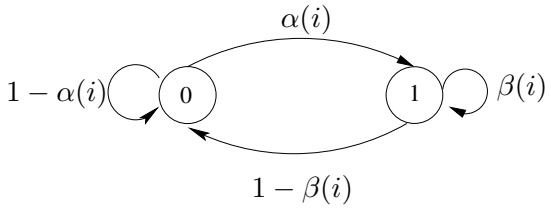

FIGURE 2.2 – Probabilités de transitions du i-ème canal.

La récompense gagnée à chaque instant est égale au nombre de transmissions que l'utilisateur secondaire a pu effectuer :

$$R_t = r(X_t, A_t) \stackrel{\text{def}}{=} \sum_{i=1}^{N} \mathbb{1}_{\{A_t(i)=1, X_t(i)=1\}} \cdot$$

Une des principales préoccupations de l'utilisateur secondaire est de trouver une *politique d'observation* π , que l'on appellera aussi *politique d'écoute*, des canaux de manière à maximiser la somme des gains à long terme.

2.2.2 Modélisation par un « restless bandit »

Le *modèle d'allocation de canal* défini dans la section précédente peut également être vu comme un « restless bandit » [Whittle, 1988].

Différents modèles de bandits ont été proposés dans la littérature. Dans le modèle standard présenté dans le paragraphe 1.1.1, un agent choisit un parmi N bras indépendants et reçoit une récompense aléatoire dont la moyenne dépend du bras choisi [Robbins, 1952]. Une extension du modèle de bandit à bras multiples consiste à introduire des états. A chaque instant t, chacun des N bras indépendants est dans un état noté $X_t(i)$, $i = 1, \ldots, N$. Dès qu'un bras i est sélectionné, l'agent reçoit une récompense qui dépend de l'état $X_t(i)$ puis cet état évolue de manière markovienne et devient $X_{t+1}(i)$ [Gittins, 1979; Frostig and Weiss, 1999]. Dans ces modèles de bandit, l'état d'un bras varie uniquement lorsque le bras est joué. Le joueur connaît donc l'état de chaque bandit. Ce modèle est un cas particulier de MDP où l'évolution de l'état du système $X_t = (X_t(1), \ldots, X_t(N))'$ se décompose en N évolutions markoviennes indépendantes. Gittins a proposé une stratégie, appelée *politique d'indice de Gittins*, permettant de jouer de manière optimale dans ce modèle (voir [Gittins, 1979; Frostig and Weiss, 1999]).

Dans le modèle « restless bandit » l'état de tous les bras varie à chaque instant, qu'ils soient joués ou non. Le *modèle d'allocation de canal* peut être vu comme un « restless bandit », puisque l'état de chaque canal évolue selon la dynamique imposée par le réseau primaire, que le canal soit observé ou non. La recherche de la politique optimale est alors nettement plus compliquée que dans un modèle de bandit à bras multiples puisque l'état de chaque canal a potentiellement nettement évolué depuis le dernier instant où il a été observé. L'utilisateur secondaire ne peut donc pas deviner si le canal est libre ou occupé. Le modèle de « restless bandit » et la recherche de politiques optimales dans ce contexte a été principalement étudié

par [Whittle, 1988; Weber and Weiss, 1990; Bertsimas and Niño-Mora, 2000; Niño-Mora, 2001; Glazebrook et al., 2002; Guha et al., 2008].

2.3 Planification

Avant de proposer un algorithme d'apprentissage par renforcement pour une version simplifiée de ce modèle d'allocation de canal, nous commençons par résoudre le problème de planification. La recherche de politique optimale dans un processus de décision markoviens partiellement observé est une tâche complexe pour laquelle il n'existe pas à ce jour de méthode générique. Nous proposons dans la section 2.3.1 une méthode approchée basée sur une grille de points. Cet algorithme est accompagné de résultats théoriques permettant de quantifier la distance entre la valeur de la politique obtenue et la valeur de la politique optimale. Cependant, une des limites de cette méthode est que la taille de la grille construite augmente de manière linéaire avec le nombre N de canaux. La complexité de l'algorithme est alors quadratique en N. Des *méthodes* dites *d'indice* pallient ce problème de complexité en découplant le problème d'optimisation en N sous-problèmes spécifiques à chaque canal. Dans le paragraphe 2.3.2, nous présentons ces méthodes d'indices permettant de trouver une politique quasi optimale. Nous considérerons au paragraphe 2.3.3 le cas particulier de canaux stochastiquement identiques.

2.3.1 Modèle général

Définition de l'état interne

A chaque instant, l'état n'est que partiellement observable puisque seuls L canaux peuvent être écoutés à la fois. Au début d'un intervalle de temps t, l'utilisateur secondaire ne connaît l'état d'aucun des N canaux, cependant, certains d'entre eux ont été récemment observés et il connaît donc la probabilité d'occupation des canaux. Pour choisir quel canal écouter, l'utilisateur secondaire garde en mémoire un état interne qui résume toutes les décisions et observations passées. Une approche standard est d'utiliser l'*état de croyance* (voir paragraphe 1.4.2) qui est un vecteur B_t de dimension 2^N tel que

$$B_t(x) = \mathbb{P}\left[X_t = x \,|\, A_{1:t-1}, Y_{1:t-1}, B_0\right] \ , \quad x \in \mathcal{X} \ .$$

On définit B_0 comme étant la probabilité initiale d'occupation des canaux qui est supposée connue. Il a été prouvé que l'état de croyance est un état interne exhaustif [Astrom, 1965]. Dans le *modèle d'allocation de canal*, l'indépendance entre les canaux peut être exploitée de manière à construire un état interne exhaustif de dimension N, beaucoup plus petite que l'état de croyance. Soit $p_t = [p_t(1), \ldots, p_t(N)]'$ tel que $p_t(i)$ est la probabilité que le canal i soit libre au début de l'intervalle de temps t conditionnellement aux observations et décisions passées :

$$p_t(i) = \mathbb{P}\left[X_t(i) = 1 \,|\, A_{1:t-1}, Y_{1:t-1}, p_0\right] \ , \quad i \in \{1, \ldots, N\} \ , \tag{2.2}$$

où $p_0(i) \stackrel{\text{def}}{=} \mathbb{P}(X_0(i) = 1)$. Le vecteur de probabilité p_t sera appelé dans la suite *probabilité d'occupation*. La proposition suivante montre que l'état interne défini par l'équation (2.2) est un état interne exhaustif.

Proposition 2.1. *Si la probabilité initiale B_0 peut être écrite comme un produit des probabilités marginales, c.a.d $\forall x \in \mathcal{X}$, $B_0(x) = \prod_{i=1}^{N} p_0(i)^{x(i)}(1 - p_0(i))^{1-x(i)}$, alors, pour tout état $x \in \mathcal{X}$,*

$$B_t(x) = \prod_{i=1}^{N} p_t(i)^{x(i)}(1 - p_t(i))^{1-x(i)} \ , \tag{2.3}$$

et il existe une fonction $\tau : [0,1]^N \times \mathcal{A} \times \mathcal{Y} \to [0,1]^N$ *telle que, pour chaque composante* i,

$$p_{t+1}(i) = \tau(p_t, A_t, Y_t)(i) \stackrel{\text{def}}{=} \begin{cases} \alpha(i) & si\ A_t(i) = 1, Y_t(i) = 0 \\ \beta(i) & si\ A_t(i) = 1, Y_t(i) = 1 \\ p_t(i)\beta(i) + (1 - p_t(i))\alpha(i) & sinon \end{cases} . \tag{2.4}$$

Fonction de valeur et équation de Bellman

Nous cherchons à déterminer une politique optimale pour le modèle présenté ci-dessus. Comme on l'a vu dans le premier chapitre, une politique optimale est définie en fonction d'un critère. Nous utiliserons dans cette section le critère γ-pondéré, les équations d'optimalité s'écrivant de manière plus simple dans ce cadre. La politique optimale recherchée ici est donc la politique maximisant

$$\mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} R(X_t, A_t) \right] , \quad 0 < \gamma < 1 .$$

Soit V_γ^π la fonction de valeur de la politique π :

$$V_\gamma^\pi(p) = \mathbb{E}^\pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} R_t \,\middle|\, p_0 = p \right] .$$

La politique optimale de tout POMDP pouvant être déduite de la politique optimale dans le MDP continu équivalent, où l'état interne du POMDP joue le rôle de la variable d'état, il existe une politique déterministe stationnaire $\pi^* : [0,1]^N \to \mathcal{A}$ optimale (voir résultats du chapitre 1) :

$$V_\gamma^{\pi^*} = V_\gamma^* \stackrel{\text{def}}{=} \max_{\pi \in \Pi^{SD}} V_\gamma^\pi ,$$

où Π^{SD} est l'ensemble des politiques stationnaires déterministes. La fonction de valeur optimale V_γ^* satisfait l'équation de Bellman

$$V_\gamma^*(p) = \max_{a \in \mathcal{A}} \left\{ \rho(p,a) + \gamma \sum_{y \in \mathcal{Y}} \phi(p,a;y) V_\gamma^*(\tau(p,a,y)) \right\} ,$$

où $\rho(p,a)$ est l'espérance de la récompense gagnée à un pas de temps étant donné l'état interne p et l'action a :

$$\rho(p,a) = \sum_{i, a(i)=1} \{ p(i)\beta(i) + (1 - p(i))\alpha(i) \} ,$$

et $\phi(p,a;y)$ est la probabilité d'observer y conditionnellement à l'état interne p et à l'action a : $\phi(p,a;y) = \prod_{i=1}^N \phi_i(p(i), a(i); y(i))$ où

$$\phi_i(p(i), a(i); y(i)) = \begin{cases} p(i)\beta(i) + (1 - p(i))\alpha(i) & \text{si } a(i) = 1 \text{ et } y(i) = 1 \\ p(i)(1 - \beta(i)) + (1 - p(i))(1 - \alpha(i)) & \text{si } a(i) = 1 \text{ et } y(i) = 0 \\ 1 & \text{si } a(i) = 0 \text{ et } y(i) = \emptyset \\ 0 & \text{sinon} \end{cases} .$$

Notons que $\phi_i(p(i), a(i); y(i))$ ne dépend que de la i-ème composante des vecteurs p, a, y, α et β. Une politique optimale peut être calculée à partir de la fonction de valeur optimale V_γ^* ; il s'agit de la politique gloutonne par rapport à V_γ^* (voir section 1.1.5) :

$$\forall p \in [0,1]^N , \ \pi^*(p) = \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ \rho(p,a) + \gamma \sum_{y \in \mathcal{Y}} \phi(p,a;y) V_\gamma^*(\tau(p,a,y)) \right\} .$$

Pour des processus de décision markoviens à espace d'états fini, l'algorithme d'itération sur les valeurs peut être utilisé pour calculer V_γ^* : il s'agit de chercher de manière itérative le point fixe de l'équation de Bellman (voir algorithme 1.1). Cependant, l'algorithme est d'un intérêt purement théorique dans ce cas puisque l'ensemble des états internes est infini.

Discrétisation de l'espace d'état interne

Une solution pratique est de discrétiser l'espace des états internes. Cette discrétisation induit inévitablement une perte d'efficacité mais, comme on le verra ci-dessous, cela permet néanmoins de trouver une politique proche de l'optimale. Il existe différentes méthodes pour discrétiser l'espace des états internes. Nous nous concentrons sur une solution qui prend en compte la structure particulière de l'état interne.

Notons $p^{k,0}(i)$ (respectivement $p^{k,1}(i)$) la probabilité que le canal i soit libre étant donné qu'il n'a pas été observé depuis k intervalles de temps et que sa dernière observation était 0 (respectivement 1). Les probabilités $p^{k,0}(i)$ et $p^{k,1}(i)$ satisfont les équations récursives suivantes :

$$p^{k,0}(i) \stackrel{\text{def}}{=} \mathbb{P}\left[X_t(i) = 1 \,|\, X_{t-k}(i) = 0, A_{t-k}(i) = 1, A_{t-k+1}(i) = \cdots = A_{t-1}(i) = 0\right] \tag{2.5}$$

$$= \begin{cases} \alpha(i) & \text{si } k = 1 \\ p^{k-1,0}(i)\beta(i) + (1 - p^{k-1,0}(i))\alpha(i) & \text{sinon} \end{cases} \tag{2.6}$$

$$p^{k,1}(i) \stackrel{\text{def}}{=} \mathbb{P}\left[X_t(i) = 1 \,|\, X_{t-k}(i) = 1, A_{t-k}(i) = 1, A_{t-k+1}(i) = \cdots = A_{t-1}(i) = 0\right] \tag{2.7}$$

$$= \begin{cases} \beta(i) & \text{si } k = 1 \\ p^{k-1,1}(i)\beta(i) + (1 - p^{k-1,1}(i))\alpha(i) & \text{sinon} \end{cases} \tag{2.8}$$

On déduit de l'équation (2.4) que, pour chaque canal i observé à l'instant $t-1$ (c'est-à-dire tel que $A_{t-1}(i) = 1$), la i-ème composante de l'état interne $p_t(i)$ est soit égale à $\alpha(i)$ soit à $\beta(i)$ selon que le canal a été vu occupé ou libre. Puisque L canaux sont observés à chaque instant, L composantes de l'état interne sont égales à $\alpha(i)$ ou $\beta(i)$. Si le canal i n'a pas été observé à l'instant $t-1$, la i-ème composante de l'état interne p_t dépend de la dernière observation faite de ce canal. Plus précisément, pour chaque canal i, $p_t(i) \in \{\nu(i), p_0(i), p^{k,0}(i), p^{k,1}(i), k > 0\}$, où $\nu(i)$ est la probabilité stationnaire que le canal i soit libre (dans la suite, on supposera que $p_0 = \nu$). Le vecteur $(1-\nu(i), \nu(i))'$ est la loi stationnaire de la chaîne de Markov $(X_t(i))_t$ telle que

$$\begin{pmatrix} 1-\nu(i) & \nu(i) \end{pmatrix} = \begin{pmatrix} 1-\nu(i) & \nu(i) \end{pmatrix} \begin{pmatrix} 1-\alpha(i) & \alpha(i) \\ \beta(i) & 1-\beta(i) \end{pmatrix},$$

d'où

$$\nu(i) = \frac{\alpha(i)}{1 - \beta(i) + \alpha(i)}.$$

L'ensemble des états internes $\tilde{\mathcal{P}}_{p_0,K}$ que nous construisons est composé de tous les états internes atteignables en au plus K intervalles de temps à partir de l'état initial $p_0 = \nu$. Ces point internes sont des vecteurs de dimension N tels que $p_t(i) \in \{\nu(i), p^{k,0}(i), p^{k,1}(i), 1 \leq k \leq K\}$ sous la contrainte que p_t a exactement L composantes égales à $\alpha(i)$ ou $\beta(i)$. La constante K est choisie de manière à limiter la complexité de l'algorithme. L'ensemble $\tilde{\mathcal{P}}_{p_0,K}$ peut être vu comme une grille adaptée.

L'ensemble des points atteignables à partir de $p_0 = \nu$ est représenté sur la figure 2.3 pour un modèle à deux canaux avec $L = 1$. Puisque $N = 2$, ces vecteurs de dimension 2 ont tous une composante i égale soit à $\alpha(i)$ soit à $\beta(i)$. L'autre coordonnée (qui correspond à un canal qui n'a pas été observé à l'instant précédent) tend vers la probabilité stationnaire. On observe sur la figure que les points se resserrent vers le point représentant la probabilité stationnaire.

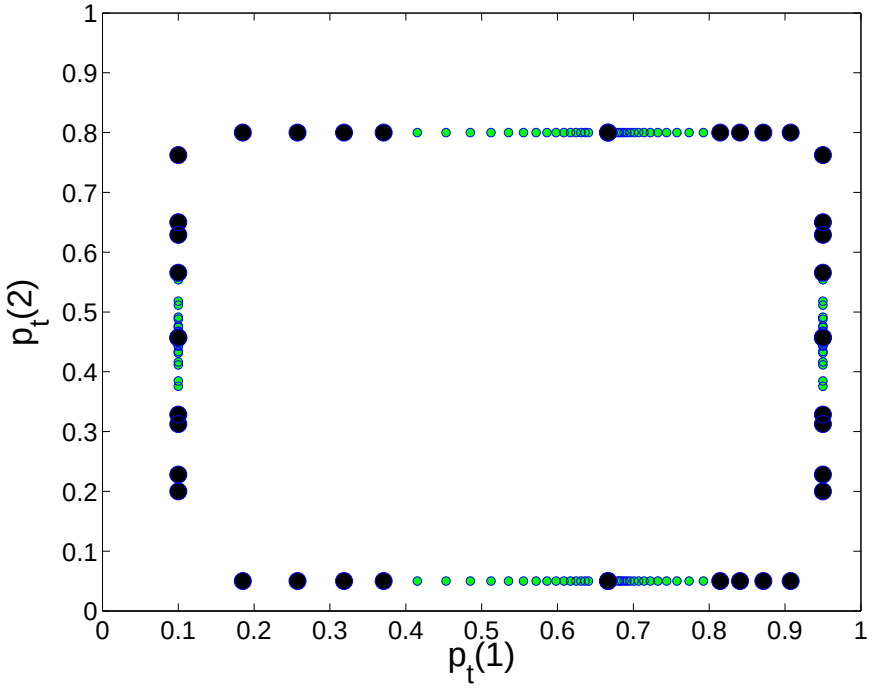

FIGURE 2.3 – Ensemble des états internes atteignables à partir de l'état interne initial p_0 (tous les points) et l'ensemble $\tilde{\mathcal{P}}_{p_0,K}$ (les gros points noirs) pour un modèle à deux canaux où un canal est observé à la fois ($L = 1$) pour $K = 4$. Les paramètres sont fixés à $\alpha = (0.1, 0.8)'$, $\beta = (0.95, 0.05)'$, et $K = 4$.

Les gros points noirs constituent l'ensemble $\tilde{\mathcal{P}}_{p_0,K}$ pour $K = 4$. Il s'agit donc des points atteignables à partir de p_0 en au plus 4 pas de temps.

Algorithme

Nous pouvons maintenant calculer la fonction de valeur optimale basée sur cet ensemble de points internes $\tilde{\mathcal{P}}_{p_0,K}$ en utilisant l'algorithme d'itération sur les valeurs (voir algorithme 1.1). Etant donné une fonction de valeur initiale $\tilde{V}_0$, on calcule de manière itérative $\tilde{V}_n$ comme suit

$$\forall p \in \tilde{\mathcal{P}}_{p_0,K}\ ,\ \tilde{V}_{n+1}(p) = \max_{a\in\mathcal{A}} \left\{ \rho(p,a) + \gamma \sum_{y\in\mathcal{Y}} \phi(p,a;y)\tilde{V}_n(\tilde{\tau}(p,a,y)) \right\}\ ,$$

où $\tilde{\tau} : \tilde{\mathcal{P}}_{p_0,K} \times \mathcal{A} \times \mathcal{Y} \to \tilde{\mathcal{P}}_{p_0,K}$ est tel que

$$\tilde{\tau}(p,a,y) = \begin{cases} \tau(p,a,y) & \text{si } p \text{ et } p' \text{ ont le même support} \\ \text{argmin}_{p'\in\tilde{\mathcal{P}}_{p_0,K}} d(p', \tau(p,a,y)) & \text{sinon} \end{cases}\ ,$$

où d est la distance définie par

$$d(p,p') = \begin{cases} \sum_{j=1}^{N} |p(j) - p'(j)| & \text{si } \forall i \in \{1,\dots,n\},\quad p(i) \in \{\alpha(i), \beta(i)\} \Rightarrow p(i) = p'(i) \\ \infty & \text{sinon} \end{cases}\ .$$

Deux états internes p et p' sont dits avoir le même support si, pour chacune des L coordonnées $i \in \{1, \cdots, N\}$ du vecteur p égales à $\alpha(i)$ ou à $\beta(i)$, la i-ème coordonnée du vecteur p' est égale à $p(i)$. La politique que nous proposons est la politique gloutonne par rapport à la fonction de valeur $\tilde{V}^*_\gamma = \lim_{n\to\infty} \tilde{V}_n$. La méthode est résumée sous la forme de l'algorithme 2.1.

Algorithme 2.1 Algorithme NOPT

Entrées: N, L, α, β, ϵ, γ

1: Pour tout $p \in \tilde{\mathcal{P}}_{p_0,K}$, $\tilde{V}_0(p) = 0$

2: **Pour tout** $n \geq 1$

3: **Pour** tout $p \in \tilde{\mathcal{P}}_{p_0,K}$ **faire**

$$\tilde{V}_n(p) = \max_{a \in \mathcal{A}} \left\{ \rho(p,a) + \gamma \sum_{y \in \mathcal{Y}} \phi(p,a;y) \tilde{V}_{n-1}(\tilde{\tau}(p,a,y)) \right\}$$

4: **fin Pour**

5: **jusqu'à ce que** $\|\tilde{V}_n - \tilde{V}_{n+1}\|_2 < \epsilon(1-\gamma)/(2\gamma)$.

6: **Pour** chaque $p \in \tilde{\mathcal{P}}_{p_0,K}$ **faire**

$$\pi_0^*(p) = \operatorname*{argmax}_{a \in \mathcal{A}} \left\{ \rho(p_{t-1},a) + \gamma \sum_{y \in \mathcal{Y}} \phi(p_{t-1},a;Y_t) \tilde{V}_n(\tilde{\tau}(p_{t-1},a,Y_t)) \right\}$$

7: **fin Pour**

Illustration sur des cas simples

Dans ce paragraphe, nous présentons le comportement de la politique obtenue à la fin de l'algorithme, appelée NOPT, et la comparons à la politique sous-optimale *myope* (utilisée par [Zhao et al., 2007a] dans ce modèle). La *politique myope*, aussi appelée stratégie à un pas, consiste à choisir l'action qui maximise la récompense espérée à un pas de temps :

$$\operatorname*{argmax}_a \rho(p,a) = \operatorname*{argmax}_a \mathbb{E}\left[R_t \,|\, A_t = a, p_{t-1} = p \right] .$$

Afin d'illustrer plus aisément le comportement de la politique, nous nous restreignons à un modèle à deux canaux. On utilise différentes valeurs pour les probabilités de transition α et β. Nous nous intéressons en particulier à des réseaux où certains canaux sont soit persistants (c'est-à-dire où l'occupation est la même pendant de longues phases temporelles) soit très fluctuants (c'est-à-dire où l'occupation du canal varie très fortement à chaque instant). Dans chacun des trois scénarios suivants, le canal 2 a une probabilité d'occupation indépendante de l'instant et est donc égale à la probabilité stationnaire d'occupation $\nu(2)$. L'occupation du canal 1, par contre, varie de manière markovienne. Sauf lorsque nous le précisons explicitement, le facteur de pondération est pris égal à $\gamma = 0.9$ et le paramètre de discrétisation égal à $K = 10$.

Dans le premier scénario, l'état du premier canal reste inchangé avec grande probabilité $1 - \alpha(1) = \beta(1) = 0.9$, et la probabilité que le deuxième canal soit libre est la même qu'il ait été vu libre ou occupé précédemment. On a $\alpha(2) = \beta(2) = \nu(2) = 0.51$. Nous représentons sur la figure 2.4, l'évolution de la probabilité d'occupation $p_t(1)$ du premier canal comparé à la probabilité stationnaire d'être libre du deuxième canal $p_t(2) = \nu(2) = 0.51$ sous les politiques myope et NOPT. Nous représentons en dessous le canal observé à chaque instant. Dans ce premier scénario, la probabilité stationnaire du premier canal ($\nu(1) = 0.5$) est plus petite que celle du deuxième canal. La politique myope sélectionne donc toujours le deuxième canal. En exploitant le fait que le premier canal est persistant (il reste dans le même état pour de longues périodes), la politique NOPT propose un choix différent d'actions consistant à observer le premier canal tant qu'il est vu libre, arrêter de l'observer dès qu'il est vu occupé et l'observer de nouveau lorsque la probabilité d'occupation dépasse un seuil (qui est autour de 0.3 dans cet exemple).

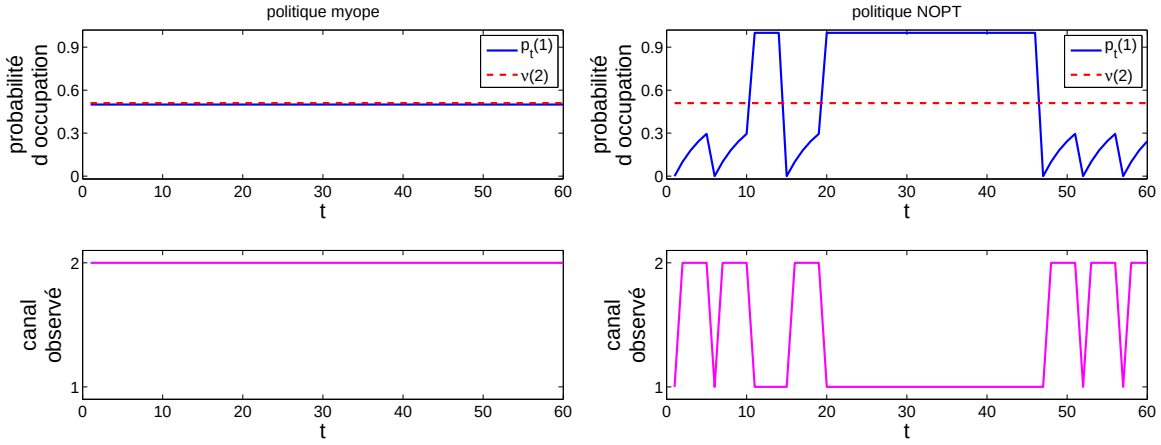

FIGURE 2.4 – Haut : Evolution de la première composante de l'état interne $p_t(1)$ (ligne continue) comparée à la probabilité stationnaire du deuxième canal $p_t(2) = \nu(2)$ (ligne pointillé) ; bas : évolution du canal observé ; pour la politique myope (à gauche) et la politique NOPT (à droite) dans le modèle à deux canaux avec $\alpha = (0.1, 0.51)'$ et $\beta = (0.9, 0.51)'$.

Dans le second scénario, le modèle est similaire sauf que les transitions entre les états dans le canal 1 sont négativement corrélées : $\alpha(1) = 1 - \beta(1) = 0.9$. On observe sur la figure 2.5 que la politique myope est la même que pour le scénario 1 alors que la politique NOPT utilise avantageusement les fluctuations du canal en l'observant juste après qu'il ait été vu libre.

Le troisième scénario permet d'étudier une situation où la probabilité stationnaire du canal 2 est plus petite que celle du canal 1. On a $\alpha = (0.1, 0.49)'$ et $\beta = (0.9, 0.49)'$ d'où $\nu(1) = 0.5$ et $\nu(2) = 0.49$. Dans ce cas, la politique myope a un comportement plus proche de la politique optimale déterminée par l'algorithme NOPT (voir figure 2.6) : quand le canal 1 est libre avec probabilité $\beta(1)$, il est observé, et, dès qu'il a été vu occupé, le canal 2 est observé tant que $p_t(1)$ est plus petite que $\nu(2) = 0.49$.

Majoration de l'erreur d'approximation

Le théorème suivant permet de majorer la différence entre la fonction de valeur optimale et la fonction de valeur obtenue en suivant l'algorithme NOPT.

Théorème 2.1. *Soit $\tilde{\mathcal{P}}_{p_0,\infty}$ l'ensemble des points atteignables à partir de la probabilité d'occupation p_0. Pour tout $p \in \tilde{\mathcal{P}}_{p_0,\infty}$, l'erreur entre la fonction de valeur optimale V^*_γ et la n-ième approximation de la fonction de valeur $\tilde{V}_n$ est bornée par*

$$\left|V^*_\gamma(p) - \tilde{V}_n(\upsilon(p))\right| \leq C_1 \sum_{i, p(i) \notin \{\alpha(i), \beta(i)\}} \frac{|\beta(i) - \alpha(i)|^K}{1 - \beta(i) + \alpha(i)} \max\{\alpha(i), 1 - \beta(i)\} + \frac{\gamma^n}{1 - \gamma} ,$$

où C_1 est la constante suivante

$$C_1 = \frac{(1 + \gamma) \max_i |\beta(i) - \alpha(i)|}{(1 - \gamma)^2 (1 - \gamma \max_i |\beta(i) - \alpha(i)|)} .$$

et υ désigne la projection d'un point de $[0,1]^N$ sur $\tilde{\mathcal{P}}_{p_0,K}$: pour tout $p \in [0,1]^N$

$$\upsilon(p) = \begin{cases} p & \text{si } p \in \tilde{\mathcal{P}}_{p_0,K} \\ \operatorname{argmin}_{p' \in \tilde{\mathcal{P}}_{p_0,K}} d(p', p) & \text{sinon} \end{cases} .$$

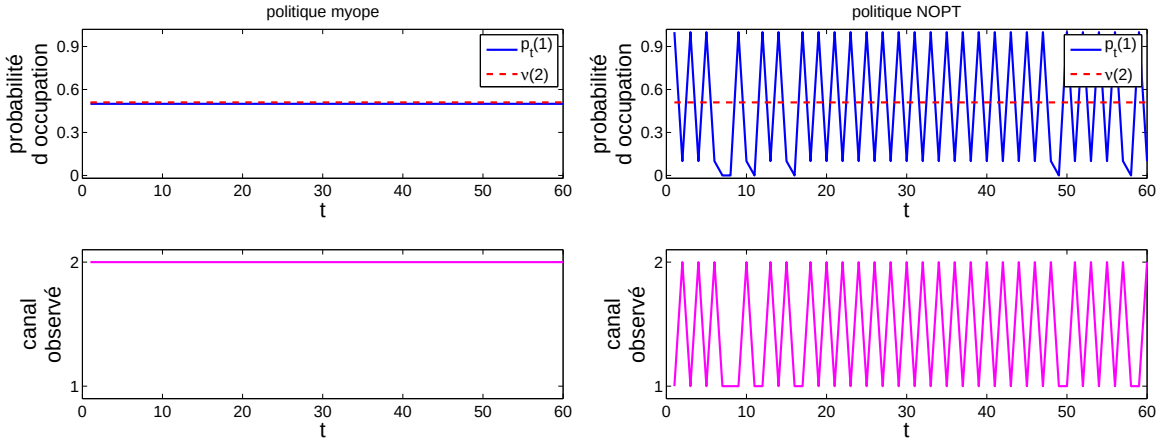

FIGURE 2.5 – Haut : Evolution de la première composante de l'état interne $p_t(1)$ (ligne continue) comparée à la probabilité stationnaire du deuxième canal $p_t(2) = \nu(2)$ (ligne pointillé) ; bas : évolution du canal observé ; pour la politique myope (à gauche) et la politique NOPT (à droite) dans le modèle à deux canaux avec $\alpha = (0.9, 0.51)'$ et $\beta = (0.1, 0.51)'$.

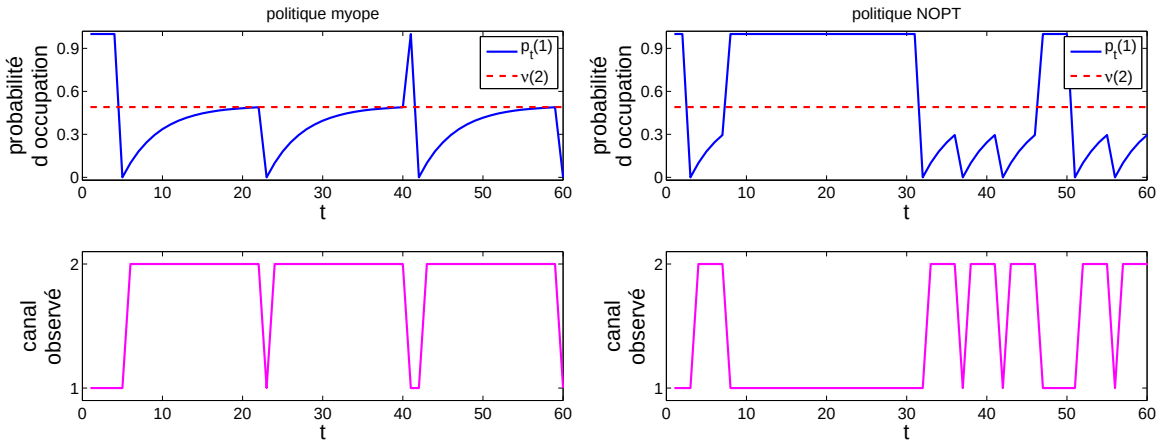

FIGURE 2.6 – Haut : Evolution de la première composante de l'état interne $p_t(1)$ (ligne continue) comparée à la probabilité stationnaire du deuxième canal $p_t(2) = \nu(2)$ (ligne pointillé) ; bas : évolution du canal observé ; pour la politique myope (à gauche) et la politique NOPT (à droite) dans le modèle à deux canaux avec $\alpha = (0.1, 0.49)'$ et $\beta = (0.9, 0.49)'$.

La borne décroît quand le paramètre de discrétisation K augmente. Le taux de décroissance dépend de la différence $|\beta(i)-\alpha(i)|$ pour $i = 1, \ldots, N$. Pour discuter de l'impact du paramètre de discrétisation K en pratique, nous avons simulé 200 trajectoires de longueur 10000 et calculé la récompense moyenne obtenue le long de chaque réalisation dans le troisième scénario. La récompense moyenne obtenue en suivant la politique NOPT avec différentes valeurs de K est représentée sur la figure 2.7. On observe que, même avec peu de points ($K \leq 5$), la politique NOPT obtient de bons résultats. Discrétiser l'espace des états internes en utilisant K plus grand que 5 n'a pas d'impact sur la récompense moyenne. Pour d'autres valeurs de probabilité de transition α et β (en particulier si $|\beta(i)-\alpha(i)|$ est très proche de 1), il peut être néanmoins nécessaire d'utiliser plus de points.

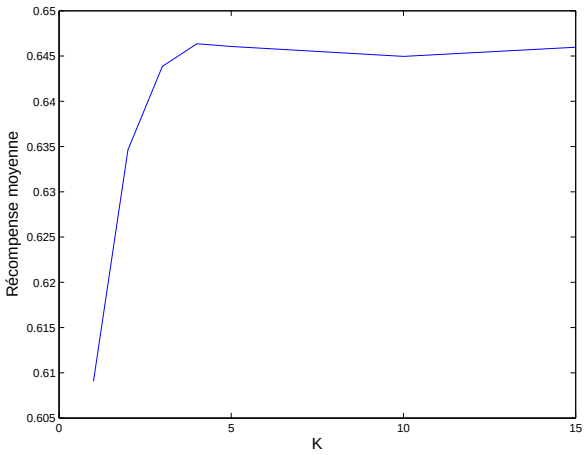

FIGURE 2.7 – Récompense moyenne en utilisant la politique NOPT pour différentes valeurs de K dans le modèle à 2 canaux avec $\alpha = (0.1, 0.51)'$ et $\beta = (0.9, 0.51)'$.

Démonstration. Posons $\mathcal{Q}_\gamma$ la fonction telle que, pour toute fonction $W : [0,1]^N \to \mathbb{R}$, tout $p \in [0,1]^N$ et tout $a \in \mathcal{A}$,

$$\mathcal{Q}_\gamma(W, p, a) = \rho(p, a) + \gamma \sum_{y \in \mathcal{Y}} \Phi(p, a; y) W(\tau(p, a, y))$$

et rappelons que $\mathcal{L}_\gamma$ est l'opérateur tel que

$$\mathcal{L}_\gamma W(p) = \max_{a \in \mathcal{A}} \mathcal{Q}_\gamma(W, p, a) \; .$$

Pour tout $p \in \tilde{\mathcal{P}}_{p_0,\infty}$,

$$\left| V_\gamma^*(p) - \tilde{V}_n(\upsilon(p)) \right| \leq |V_\gamma^*(p) - V_\gamma^*(\upsilon(p))| + |V_\gamma^*(\upsilon(p)) - \tilde{V}_n(\upsilon(p))|$$

Majorons séparément chacun de ces deux termes.

1. On sait que pour tout $p \in [0,1]$, $V_\gamma^*(p) = \lim_{k \to \infty} \mathcal{L}_\gamma^k V_0(p)$. Ainsi, pour majorer $|V_\gamma^*(p) - V_\gamma^*(\upsilon(p))|$, montrons par récurrence que, pour tout $p, p' \in \tilde{\mathcal{P}}_{p_0,\infty}$, et tout $k \in \mathbb{N}$,

$$|(\mathcal{L}_\gamma^k V_0)(p) - (\mathcal{L}_\gamma^k V_0)(p')| \leq C \; d(p, p') \tag{2.9}$$

avec

$$C = \frac{(1+\gamma) \max_{i=\{1\ldots N\}} |\beta(i) - \alpha(i)|}{(1-\gamma)(1 - \gamma \max_{i=\{1\ldots N\}} |\beta(i) - \alpha(i)|)} \; .$$

– Soit $a = \text{argmax}_{b\in\mathcal{A}} \max\{\mathcal{Q}_\gamma(V_0, p, b), \mathcal{Q}_\gamma(V_0, p', b)\}$, alors

$$|(\mathcal{L}_\gamma V_0)(p) - (\mathcal{L}_\gamma V_0)(p')| \leq |\rho(p, a) - \rho(p', a)| .$$

Or pour tout $p, p' \in [0, 1]$ et tout $a \in \mathcal{A}$,

$$|\rho(p, a) - \rho(p', a)| \leq \max_{i=\{1\ldots N\}} |\beta(i) - \alpha(i)| d(p, p') \leq C\ d(p, p') \tag{2.10}$$

Donc (2.9) est vraie au rang 1.

– Supposons que (2.9) soit vraie au rang k. Soit

$$a = \underset{b\in\mathcal{A}}{\text{argmax}}\ \max\left\{\mathcal{Q}_\gamma(V_0, p, b), \mathcal{Q}_\gamma(V_0, p', b)\right\} .$$

On a

$$\begin{aligned}
&|(\mathcal{L}_\gamma^{k+1} V_0)(p) - (\mathcal{L}_\gamma^{k+1} V_0)(p')| \\
&\qquad \leq |\rho(p, a) - \rho(p', a) + \gamma \sum_{y\in\mathcal{Y}} \Phi(p, a; y)(\mathcal{L}_\gamma^k V_0)(\tau(p, a, y)) \\
&\qquad\quad - \gamma \sum_{y\in\mathcal{Y}} \Phi(p', a; y)(\mathcal{L}_\gamma^k V_0)(\tau(p', a, y))| \\
&\qquad \leq |\rho(p, a) - \rho(p', a)| + \gamma \sum_{y\in\mathcal{Y}} \Phi(p, a; y)|(\mathcal{L}_\gamma^k V_0)(\tau(p, a, y)) - (\mathcal{L}_\gamma^k V_0)(\tau(p', a, y))| \\
&\qquad\quad + \gamma \sum_{y\in\mathcal{Y}} |\Phi(p, a; y) - \Phi(p', a; y)||(\mathcal{L}_\gamma^k V_0)(\tau(p', a, y))| \\
&\qquad \leq |\rho(p, a) - \rho(p', a)| + \gamma C \sum_{y\in\mathcal{Y}} \Phi(p, a; y)\ d(\tau(p, a, y), \tau(p', a, y)) \\
&\qquad\quad + \gamma \sum_{y\in\mathcal{Y}} |\Phi(p, a; y) - \Phi(p', a; y)||(\mathcal{L}_\gamma^k V_0)(\tau(p', a, y))| .
\end{aligned}$$

Or

$$d(\tau(p, a, y), \tau(p', a, y)) \leq \max_{i=\{1\ldots N\}} |\beta(i) - \alpha(i)| d(p, p') .$$

D'après l'inégalité (2.10) et celle ci-dessus, on a

$$\begin{aligned}
|(\mathcal{L}_\gamma^{k+1} V_0)(p) - (\mathcal{L}_\gamma^{k+1} V_0)(p')| \leq (1 + \gamma C) \max_{i=\{1\ldots N\}} |\beta(i) - \alpha(i)|\ d(p, p') \\
+ \gamma \left\| \mathcal{L}_\gamma^k V_0 \right\|_\infty \sum_{y\in\mathcal{Y}} |\Phi(p, a; y) - \Phi(p', a; y)| .
\end{aligned}$$

De plus, $\left\|\mathcal{L}_\gamma^k V_0\right\|_\infty \leq \sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$ (puisque $\rho(x, a)$ est majorée par 1 pour tout x et tout a) et

$$\sum_{y\in\mathcal{Y}} |\Phi(p, a; y) - \Phi(p', a; y)| \leq 2 \max_{i=\{1\ldots N\}} |\beta(i) - \alpha(i)| d(p, p') .$$

D'où

$$\begin{aligned}
|(\mathcal{L}_\gamma^{k+1} V_0)(p) - (\mathcal{L}_\gamma^{k+1} V_0)(p')| &\leq (1 + C\gamma + \frac{2\gamma}{1-\gamma}) \max_{i=\{1\ldots N\}} |\beta(i) - \alpha(i)| d(p, p') \\
&= C \max_{i=\{1\ldots N\}} |\beta(i) - \alpha(i)| d(p, p') .
\end{aligned}$$

Ainsi

$$|V_\gamma^*(p) - V_\gamma^*(\upsilon(p))| \leq C\, d(p, \upsilon(p)) \,. \tag{2.11}$$

2. Montrons maintenant que, pour tout $p \in \tilde{\mathcal{P}}_{p_0,\infty}$ et $k \geq 1$,

$$|V_\gamma^*(p) - \tilde{V}_k(p)| \leq C\, D \sum_{j=1}^{k-1} \gamma^{j-1} + \gamma^k \left\|V_\gamma^*\right\|_\infty \;, \tag{2.12}$$

où $D = \sup_{p \in \tilde{\mathcal{P}}_{p_0,\infty}} d(p, \upsilon(p))$. La preuve se fait par récurrence.

– Pour tout $p \in \tilde{\mathcal{P}}_{p_0,\infty}$, il existe $a \in \mathcal{A}$ tel que

$$|V_\gamma^*(p) - \tilde{V}_1(p)| \leq \gamma \sum_{y \in \mathcal{Y}} \Phi(p, a; y) |V_\gamma^*(\tau(p, a, y)) - \tilde{V}_0(\upsilon(\tau(p, a, y)))| \leq \gamma \left\|V_\gamma^*\right\|_\infty$$

donc (2.12) est vraie au rang 1.

– Supposons que (2.12) soit vraie au rang k. Pour tout $p \in \tilde{\mathcal{P}}_{p_0,\infty}$, il existe $a \in \mathcal{A}$ tel que

$$\begin{aligned} |V_\gamma^*(p) - \tilde{V}_{k+1}(p)| &\leq \gamma \sum_{y \in \mathcal{Y}} \Phi(p, a; y) |V_\gamma^*(\tau(p, a, y)) - \tilde{V}_k(\upsilon(\tau(p, a, y)))| \\ &\leq \gamma \sum_{y \in \mathcal{Y}} \Phi(p, a; y) \, |V_\gamma^*(\tau(p, a, y)) - V_\gamma^*(\upsilon(\tau(p, a, y))) \\ &\qquad + V_\gamma^*(\upsilon(\tau(p, a, y))) - \tilde{V}_k(\upsilon(\tau(p, a, y)))| \end{aligned}$$

D'après l'équation (2.11), on a $|V_\gamma^*(\tau(p, a, y)) - V_\gamma^*(\upsilon(\tau(p, a, y)))| \leq C\, D$. De plus, d'après (2.12),

$$|V_\gamma^*(\upsilon(\tau(p, a, y))) - \tilde{V}_k(\upsilon(\tau(p, a, y)))| \leq C\, D \sum_{j=1}^{k-1} \gamma^{j-1} + \gamma^k \left\|V_\gamma^*\right\|_\infty \,.$$

D'où

$$|V_\gamma^*(p) - \tilde{V}_{k+1}(p)| \leq \gamma C\, D + \gamma C\, D \sum_{j=1}^{k-1} \gamma^{j-1} + \gamma \gamma^k \left\|V_\gamma^*\right\|_\infty \,.$$

On a donc démontré (2.12) au rang $k + 1$.

En combinant les résultats précédents, on a

$$\left|V_\gamma^*(p) - \tilde{V}_n(\upsilon(p))\right| \leq C\, d(p, \upsilon(p)) + C\, D \sum_{j=1}^{n-1} \gamma^{j-1} + \gamma^n \left\|V_\gamma^*\right\|_\infty \,.$$

De plus, pour tout $p \in \tilde{\mathcal{P}}_{p_0,\infty}$

$$d(p, \upsilon(p)) \leq \sum_{i=1}^{N} |p(i) - \upsilon(p)(i)| = \sum_{i, p(i) \notin \{\alpha(i), \beta(i)\}} |p(i) - \upsilon(p)(i)| \,.$$

Pour une composante i fixée, la distance entre $p(i)$ et $\upsilon(p)(i)$ est plus petite que la distance entre $p^{K,0}(i)$ (ou $p^{K,1}(i)$) et $\nu(i)$, d'où

$$\begin{aligned} |p(i) - \upsilon(p)(i)| &\leq \max\{|\nu(i) - p^{K,0}(i)|, |\nu(i) - p^{K,1}(i)|\} \\ &\leq \frac{|\beta(i) - \alpha(i)|^K}{1 - \beta(i) + \alpha(i)} \max\{\alpha(i), 1 - \beta(i)\} \,. \end{aligned}$$

Donc, pour tout $p \in \tilde{\mathcal{P}}_{p_0,\infty}$

$$d(p, v(p)) \leq \sum_{i, p(i) \notin \{\alpha(i), \beta(i)\}} \frac{|\beta(i) - \alpha(i)|^K}{1 - \beta(i) + \alpha(i)} \max\{\alpha(i), 1 - \beta(i)\} \ . \tag{2.13}$$

En utilisant ce dernier résultat ainsi que le fait que $\left\|V_\gamma^*\right\|_\infty \leq \frac{1}{1-\gamma}$, on obtient le résultat du théorème. □

2.3.2 Politiques d'indice

[Papadimitriou and Tsitsiklis, 1994] ont établi que la tâche de planification dans le modèle « restless bandit » est PSPACE-dure [1]. La politique optimale ne peut donc pas être trouvée en pratique lorsque le nombre N de canaux est grand avec un algorithme tel que le précédent. Cependant, des travaux récents ont proposé des algorithmes permettant de déterminer des politiques proches de l'optimale appelées *politiques d'indice* [Guha and Munagala, 2007; Le Ny et al., 2008; Liu and Zhao, 2008], qui ont une complexité moindre. Une politique d'indice consiste à scinder la tâche d'optimisation en N sous-problèmes spécifiques à chaque canal. Cette idée a été tout d'abord proposée par [Whittle, 1988] qui s'était inspiré des travaux de [Gittins, 1979] menés dans le cadre des bandits à bras multiples.

Indice de Whittle

La politique d'indice de Whittle découle d'une heuristique basée sur une relaxation de la contrainte du problème d'optimisation initial : au lieu de chercher une politique qui maximise la somme des récompenses sur le long terme sous la contrainte que L canaux *exactement* sont observés à chaque instant, on suppose que *en moyenne* L canaux sont observés à chaque instant. Le problème d'optimisation est alors le suivant : trouver la politique π qui maximise

$$\lim_{n\to\infty} \mathbb{E}^\pi \left[\frac{1}{n} \sum_{t=0}^{n-1} \sum_{i=1}^{N} \mathbb{1}_{\{A_t(i)=1, X_t(i)=1\}} \right]$$

sous la contrainte

$$\lim_{n\to\infty} \mathbb{E}^\pi \left[\frac{1}{n} \sum_{t=0}^{n-1} \sum_{i=1}^{N} \mathbb{1}_{\{A_t(i)=0\}} \right] = N - L \ .$$

Cette optimisation sous contrainte revient alors à calculer

$$\sup_\pi \inf_{\lambda \geq 0} \lim_{n\to\infty} \frac{1}{n} \mathbb{E}^\pi \left[\sum_{t=0}^{n-1} \sum_{i=1}^{N} \mathbb{1}_{\{A_t(i)=1, X_t(i)=1\}} + \lambda \mathbb{1}_{\{A_t(i)=0\}} + \lambda(N - L) \right]$$

où λ est un multiplicateur de Lagrange. La maximisation précédente est équivalente à [Altman, 1999] :

$$\inf_\lambda \sum_{i=1}^{N} \sup_{\pi_i} \lim_{n\to\infty} \frac{1}{n} \mathbb{E}^{\pi_i} \left[\sum_{t=0}^{n-1} \mathbb{1}_{\{A_t(i)=1, X_t(i)=1\}} + \lambda \mathbb{1}_{\{A_t(i)=0\}} \right] + \lambda(N - L) \ .$$

1. La définition précise de PSPACE requiert une introduction à la théorie de la complexité qui sort du cadre de ce chapitre, et nous invitons le lecteur intéressé à consulter [Papadimitriou, 2003]. L'encyclopédie en ligne *Wikipedia* explique que le terme PSPACE désigne l'ensemble de tous les problèmes de décision qui peuvent être résolus par une machine Turing en utilisant une espace mémoire polynomiale. Un problème est dit PSPACE-dur si il est au moins aussi difficile que tous les problèmes de l'espace PSPACE.

Ainsi, pour tout λ fixé, il s'agit de trouver la politique π_i qui maximise la récompense moyenne dans chaque canal i à laquelle on rajoute une récompense égale à λ reçue lorsque l'agent n'observe pas le canal. Notons

$$\eta^{*,i,\lambda} = \max_{\pi} \lim_{n\to\infty} \frac{1}{n}\mathbb{E}^{\pi}\left[\sum_{t=1}^{n} \mathbb{1}_{\{A_t(i)=1, X_t(i)=1\}} + \lambda \mathbb{1}_{\{A_t(i)=0\}}\right]$$

la récompense moyenne optimale dans le canal i dans le modèle dans lequel on ajoute une récompense λ pour l'action correspondant à ne pas écouter le canal. Notons $h^{i,\lambda} : [0,1] \to \mathbb{R}$ une fonction de biais associée à la politique optimale dans ce modèle (voir paragraphe 1.2.2). A chaque instant, l'agent a le choix entre observer le canal i ou ne pas l'observer. Si le canal est observé, la récompense espérée à un coup est p et la récompense espérée pour les pas de temps suivants est $h^{i,\lambda}(\alpha)$ si le canal est libre (ce qui se produit avec probabilité p) et $h^{i,\lambda}(\beta)$ s'il est occupé. Si l'agent n'observe pas le canal, la récompense reçue est λ et l'espérance de la moyenne des récompenses futures est $h^{i,\lambda}(p')$ où $p' = p\beta + (1-p)\alpha$ est la probabilité que le canal soit libre à l'instant suivant. L'équation de Bellman s'écrit alors :

$$\eta^{*,i,\lambda} + h^{i,\lambda}(p) = \max\left\{p + ph^{i,\lambda}(\alpha(i)) + (1-p)h^{i,\lambda}(\beta(i)) \, , \lambda + h^{i,\lambda}(p\beta(i) + (1-p)\alpha(i)\right\} \, .$$

L'indice de Whittle $W_t(i)$ pour le canal i à l'instant t est défini comme étant la valeur de λ telle que, à l'instant t, les décisions d'observer ou de ne pas observer le canal soient aussi attractives l'une que l'autre. $W_t(i)$ est donc caractérisé dans notre cas par l'équation

$$p_t(i)+p_t(i)h^{i,W_t(i)}(\alpha(i))+(1-p_t(i))h^{i,W_t(i)}(\beta(i)) = W_t(i)+h^{i,W_t(i)}(p_t(i)\beta(i)+(1-p_t(i))\alpha(i)) \, .$$

Cet indice de Whittle existe sous certaines conditions dite d'*indexabilité du modèle*. Ces conditions sont aisément vérifiées dans le « restless bandit » considéré dans ce chapitre (voir [Le Ny et al., 2008; Liu and Zhao, 2010b]).

La politique d'indice proposée par Whittle consiste à calculer, à chaque instant t et pour chaque canal i, l'indice $W_t(i)$ et à observer les L canaux correspondants aux L plus grands indices. Whittle conjecture que lorsque le rapport L/N tend vers une constante quand N tend vers l'infini, sa politique d'indice est optimale pour le critère moyen [Weber and Weiss, 1990; Frostig and Weiss, 1999].

Modèle d'écoute de canal

Pour déterminer l'indice de Whittle associé à chaque canal, il faut pouvoir calculer $\eta^{*,i,\lambda}$ et $h^{i,\lambda}$ pour tout λ. Cela nécessite de résoudre le problème de planification dans un modèle à un seul canal dans lequel un gain λ est associé au fait de ne pas observer le canal [Le Ny et al., 2008; Liu and Zhao, 2008]. Ce modèle sera appelé par la suite *modèle d'écoute de canal*.

Une autre relaxation possible du *modèle d'allocation de canal* consiste à ne pas déterminer à l'avance le nombre de canaux à observer à chaque instant mais à introduire un coût pour écouter (observer) chaque canal. On remarque que la politique optimale dans le modèle à un canal avec une pénalité pour observer le canal est la même que celle dans les N *modèles d'écoute de canal* indépendants. Nous exposons dans la suite de ce paragraphe ce modèle d'écoute de canal ainsi que les politiques optimales selon les valeurs des paramètres du modèle.

Considérons un unique canal. Soit X_t l'état du canal qui est égal à 0 quand le canal est occupé et à 1 quand il est libre. Soit α (resp. β) la probabilité de transition de l'état 0 (resp. 1) à l'état 1.

A chaque instant, l'utilisateur secondaire peut choisir d'observer l'état du canal ($A_t = 1$) ou de ne pas l'observer ($A_t = 0$). L'observation Y_t ainsi obtenue est égale à X_t si le canal a été observé, et à l'ensemble vide sinon.

La récompense gagnée à chaque instant est définie par

$$r(X_t, A_t) = \begin{cases} 1 & \text{si } A_t = 1,\ X_t = Y_t = 1 \\ 0 & \text{si } A_t = 1,\ X_t = Y_t = 0 \\ \lambda & \text{sinon} \end{cases} .$$

Elle ne dépend de l'état X_t qu'à travers l'observation Y_t. La récompense $0 \leq \lambda \leq 1$ associée à l'action de non-observation (appelée subside par Whittle [Whittle, 1988]) peut également être interprétée comme un coût fixe à payer pour observer un canal : en soustrayant λ à toutes les récompenses, le modèle est équivalent au modèle suivant. Si l'agent observe le canal, il paye un coût λ et reçoit une récompense égale à 1 si ce dernier est libre et de 0 s'il est occupé. Si l'agent n'observe pas le canal, il ne reçoit ni ne paye rien.

Le *modèle d'écoute de canal* est entièrement déterminé par la connaissance des deux probabilités de transition α et β et celle du coût λ. Il peut être reformulé comme un MDP à état continu en utilisant l'état interne p_t qui résume toutes les décisions et les observations passées (voir section 1.4) : $p_t = \mathbb{P}\left[X_t = 1 \,|\, A_{0:t-1}, Y_{0:t-1}\right]$. On rappelle que cet état interne satisfait l'équation récursive suivante :

$$p_{t+1} = \begin{cases} \alpha & \text{si } A_t = 1, Y_t = 0 \\ \beta & \text{si } A_t = 1, Y_t = 1 \\ p_t\beta + (1 - p_t)\alpha & \text{sinon} \end{cases} . \tag{2.14}$$

L'état de croyance p_t est ainsi entièrement déterminé par le couple (K_t, U_t), où K_t est la durée depuis la dernière observation et U_t est l'état du canal lors de cette dernière observation : le canal a été observé pour la dernière fois à l'instant $t - K_t$ et il était dans l'état $U_t = X_{t-K_t} \in \{0, 1\}$. Ce couple (K_t, U_t) est un autre état interne du système. Cet état interne appartient à l'ensemble infini dénombrable $\mathbb{N}^* \times \{0, 1\}$. Notons que $p^{k,u}_{\alpha,\beta}$, défini par les équations (2.6) et (2.8), désigne la probabilité conditionnelle que le canal soit libre étant donné que $(K_t, U_t) = (k, u)$: pour $k > 1$,

$$p^{k,u}_{\alpha,\beta} = \mathbb{P}[X_t = 1 | A_{t-k+1:t-1} = 0, A_{t-k} = 1, Y_{t-k} = u]$$

et

$$p^{1,u}_{\alpha,\beta} = \mathbb{P}\left[X_t = 1 \,|\, A_{t-1} = 1, Y_{t-1} = u\right] .$$

Nous explicitons volontairement, à partir de maintenant, la dépendance en α et β de $p^{k,u}$. L'équation (2.4) implique que, pour $k > 1$:

$$p^{k,0}_{\alpha,\beta} = \frac{\alpha(1 - (\beta - \alpha)^k)}{1 - \beta + \alpha} , \tag{2.15}$$

$$p^{k,1}_{\alpha,\beta} = \frac{(\beta - \alpha)^k(1 - \beta) + \alpha}{1 - \beta + \alpha} . \tag{2.16}$$

Il est facile de prouver que l'état interne (K_t, U_t) est une *statistique exhaustive*. Il existe donc une politique optimale dépendant seulement de cet état interne.

Soit $\pi : \mathbb{N}^* \times \{0, 1\} \to \mathcal{A}$ une politique qui associe une action à chaque état interne (K_t, U_t), et Π l'ensemble des politiques possibles de ce type. Toute politique dans Π est caractérisée par le couple (m_0, m_1) qui définit combien de temps l'utilisateur secondaire attend avant d'observer de nouveau le canal en fonction de la dernière observation. La politique $\pi_{(m_0,m_1)}$ consiste à attendre $m_0 - 1$ (resp. $m_1 - 1$) instants avant d'observer le canal de nouveau si, la dernière fois que le canal a été observé, il était occupé (resp. libre). Soit π_∞ la politique qui consiste à ne jamais observer le canal.

Calculons maintenant la récompense moyenne $\eta^{\pi}_{\alpha,\beta}$ associée à tout politique $\pi \in \Pi$:

$$\eta^{\pi}_{\alpha,\beta} = \lim_{n\to\infty} \frac{1}{n} \mathbb{E}^{\pi}_{\alpha,\beta} \left[\sum_{t=1}^{n} r(X_t, A_t) \right] . \tag{2.17}$$

La récompense moyenne reçue en suivant une telle politique peut être calculée de manière exacte en fonction du couple des probabilités (α, β). Une fois la récompense moyenne $\eta^{\pi}_{\alpha,\beta}$ calculée pour toute politique π, on peut définir la politique optimale pour une valeur de (α, β) fixée par :

$$\pi^{*}_{\alpha,\beta} = \operatorname*{argmax}_{\pi\in\Pi} \eta^{\pi}_{\alpha,\beta} . \tag{2.18}$$

La récompense moyenne $\eta^{\pi}_{\alpha,\beta}$ de la politique π, définie par l'équation (2.17), peut être écrite comme une fonction de la probabilité stationnaire μ^{π} de la chaîne de Markov des états internes $\{(K_t, U_t)\}_t$ [Puterman, 1994] :

$$\eta^{\pi}_{\alpha,\beta} = \sum_{k\in\mathbb{N}^*} \sum_{u\in\{0,1\}} \mu^{\pi}(k,u) \left[p^{k,u}_{\alpha,\beta} \mathbb{1}_{\{\pi(k,u)=1\}} + \lambda \mathbb{1}_{\{\pi(k,u)=0\}} \right] . \tag{2.19}$$

Soient Ψ^{π} la probabilité de transition de la chaîne de Markov des états internes $\{(K_t, U_t)\}_t$, lorsque la politique π est suivie :

$$\Psi^{\pi}((k,u);(k',u')) = \mathbb{P}\left[(K_t, U_t) = (k',u') \,\middle|\, (K_{t-1}, U_{t-1}) = (k,u) \right] .$$

Rappelons que μ^{π} vérifie $\mu^{\pi} = \mu^{\pi}\Psi^{\pi}$.

Ainsi, les récompenses moyennes des politiques $\pi_{(m_0,m_1)}$ et π_{∞} peuvent être calculées en déterminant les probabilités stationnaires de la chaîne de Markov des états internes sous ces politiques.

Sous la politique $\pi_{(m_0,m_1)}$, où $m_0, m_1 \in \mathbb{N}^*$, l'état interne (K_t, U_t) peut seulement prendre une des valeurs suivantes : $(k, 0)$ avec $1 \le k \le m_0$ ou $(k', 1)$ avec $1 \le k' \le m_1$. En effet, si l'état était occupé (resp. libre) la dernière fois qu'il a été vu, on n'attend pas plus de m_0 (resp. m_1) instants avant de l'observer de nouveau. La probabilité de transition $\Psi^{\pi_{(m_0,m_1)}}$ entre ces états est, pour tout $k \in \mathbb{N}$ et $y \in \{0,1\}$,

$$\Psi^{\pi_{(m_0,m_1)}}((k,y),(k',y')) = \begin{cases} 1 & \text{si } 2 \le k' = k+1 \le m_y \text{ et } y = y' \\ p^{m_y,y}_{\alpha,\beta} & \text{si } k = m_y,\ k' = 1 \text{ et } y' = 1 \\ 1 - p^{m_y,y}_{\alpha,\beta} & \text{si } k = m_y,\ k' = 1 \text{ et } y' = 0 \\ 0 & \text{sinon} \end{cases} .$$

En résolvant l'équation $\mu^{\pi_{(m_0,m_1)}}\Psi^{\pi_{(m_0,m_1)}} = \mu^{\pi_{(m_0,m_1)}}$, on détermine la probabilité stationnaire

$$\mu^{\pi_{(m_0,m_1)}}(k,0) = \frac{1 - p^{m_1,1}_{\alpha,\beta}}{m_1 p^{m_0,0}_{\alpha,\beta} + m_0(1 - p^{m_1,1}_{\alpha,\beta})} \quad \text{pour tout } 1 \le k \le m_0 ,$$

$$\mu^{\pi_{(m_0,m_1)}}(k,1) = \frac{p^{m_0,0}_{\alpha,\beta}}{m_1 p^{m_0,0}_{\alpha,\beta} + m_0(1 - p^{m_1,1}_{\alpha,\beta})} \quad \text{pour tout } 1 \le k \le m_1 .$$

En utilisant l'équation (2.19), on obtient la récompense moyenne suivante

$$\eta^{\pi_{(m_0,m_1)}}_{\alpha,\beta} = \frac{p^{m_0,0}_{\alpha,\beta} + \lambda[(m_1-1)p^{m_0,0}_{\alpha,\beta} + (m_0-1)p^{m_1,1}_{\alpha,\beta}]}{m_0(1 - p^{m_1,1}_{\alpha,\beta}) + m_1 p^{m_0,0}_{\alpha,\beta}} . \tag{2.20}$$

Si m_0 ou m_1 est égal à l'infini, après un nombre fini d'instants l'utilisateur secondaire n'observe plus le canal. La récompense moyenne est alors

$$\eta_{\alpha,\beta}^{\pi_\infty} = \lambda \, . \tag{2.21}$$

Pour un paramètre (α, β) fixé, la politique optimale $\pi^*_{\alpha,\beta}$ est donnée par l'équation (2.18). Pour la déterminer, il suffit de comparer la récompense moyenne associée à chaque politique $\pi \in \Pi$ et de déterminer laquelle de ces politiques est supérieure aux autres pour la valeur du paramètre (α, β). Le calcul de l'indice de Whittle nécessite également la connaissance du biais associé à la politique optimale qui peut être trouvé en résolvant l'équation de Bellman. Les auteurs de [Le Ny et al., 2008] et de [Liu and Zhao, 2008] ont calculés les indices de Whittle pour ce modèle en fonction du paramètre (α, β).

2.3.3 Modèle de canaux stochastiquement identiques

Le *modèle d'allocation de canal* exposé schéma 2.1 peut être traité dans un cas particulier nettement plus simple où tous les canaux sont *stochastiquement identiques*, c'est-à-dire dont les probabilités de transition sont égales :

$$\forall i \in \{1, \dots, N\} \, , \; \alpha(i) = \alpha \, , \; \beta(i) = \beta \, .$$

Ce modèle a été considéré dans plusieurs travaux ces dernières années [Liu and Zhao, 2008; Zhao et al., 2008]. En particulier, il a été prouvé que, dans ce cas, la politique d'indice de Whittle est équivalente à la *politique myope* (voir [Liu and Zhao, 2008]) qui consiste à sélectionner les canaux à observer selon la récompense espérée à un pas de temps :

$$A_t = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \sum_{i=1}^{N} a(i) p_{\alpha,\beta}^{K_t(i),U_t(i)}$$

En suivant cette politique, l'utilisateur secondaire observe les L canaux ayant la plus grande probabilité d'être libre $p_{\alpha,\beta}^{k,y}$. La politique dépend alors uniquement du fait que le système est positivement $(\alpha \leq \beta)$ ou négativement $(\beta \leq \alpha)$ corrélé (voir [Liu and Zhao, 2008] pour plus de détails). Pour comprendre la différence notable entre ces deux cas, la figure 2.8 représente la probabilité $p_{\alpha,\beta}^{k,y}$ pour $y = 1$ et $y = 0$ en fonction de k. On distingue le cas corrélé du cas non corrélé. On observe que, pour tout $k \geq 1$, pour tout $y \in \{0, 1\}$,

$$\begin{cases} p_{\alpha,\beta}^{1,0} = \alpha \leq p_{\alpha,\beta}^{k,y} \leq \beta = p_{\alpha,\beta}^{1,1} & \text{si } \alpha \leq \beta \, , \\ p_{\alpha,\beta}^{1,1} = \beta \leq p_{\alpha,\beta}^{k,y} \leq \alpha = p_{\alpha,\beta}^{1,0} & \text{si } \beta \leq \alpha \, . \end{cases} \tag{2.22}$$

Le comportement de la politique myope se déduit facilement de ces équations. Dans le cas positivement corrélé $(\alpha \leq \beta)$,

- si le canal i a été observé l'instant précédent et était libre, c.a.d. $K_t(i) = 1$, $U_t(i) = 1$, on sait que la probabilité $p_{\alpha,\beta}^{K_t(j),U_t(j)}$ que tout autre canal j soit libre est inférieure ou égale à $p_{\alpha,\beta}^{1,1}$, donc la politique myope consiste à observer de nouveau le canal i ;
- si le canal i a été observé l'instant précédent et était occupé, c.a.d. $K_t(i) = 1$, $U_t(i) = 0$, on sait que la probabilité $p_{\alpha,\beta}^{K_t(j),U_t(j)}$ que tout autre canal j soit libre est supérieure ou égale à $p_{\alpha,\beta}^{1,0}$, donc la politique myope consiste à ne pas réobserver le canal i.

Au contraire, dans le cas négativement corrélé $(\beta \leq \alpha)$,

- si le canal i a été observé l'instant précédent et était occupée, c.a.d. $K_t(i) = 1$, $U_t(i) = 0$, la probabilité $p_{\alpha,\beta}^{K_t(j),U_t(j)}$ est inférieure à $p_{\alpha,\beta}^{1,0}$, donc la politique myope consiste à observer de nouveau le canal i ;

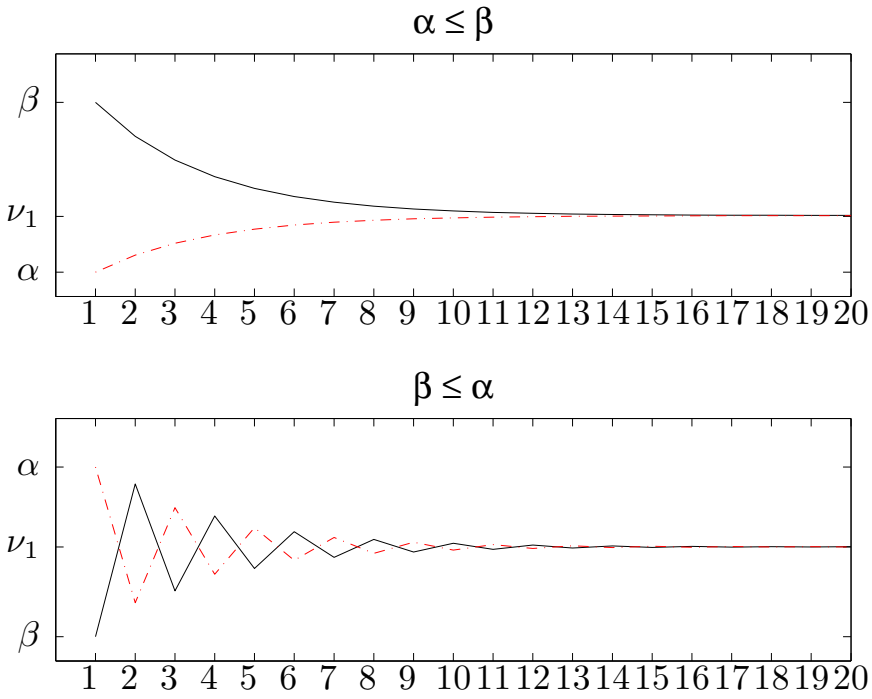

FIGURE 2.8 – Probabilités $p_{\alpha,\beta}^{k,y}$ pour $y = 1$ (ligne continue) et $y = 0$ (ligne pointillée) en fonction de k, dans le cas positivement corrélé (haut) et le cas négativement corrélé (bas).

– si le canal i a été observé l'instant précédent et était libre, c.a.d. $K_t(i) = 1$, $U_t(i) = 0$, la probabilité $p_{\alpha,\beta}^{K_t(j),U_t(j)}$ est supérieure ou égale à $p_{\alpha,\beta}^{1,0}$, donc la politique myope consiste à ne pas réobserver le canal i.

Notons π_+ la politique dans le cas positivement corrélé et π_- celle dans le cas négativement corrélé. La récompense à long terme reçue en suivant les politiques π_+ et π_- ne peuvent pas être calculées de manière exacte. Cependant, l'approche de [Zhao et al., 2008] peut être utilisée pour calculer une approximation de $\eta_{\alpha,\beta}^{\pi_+}$ et $\eta_{\alpha,\beta}^{\pi_-}$.

Soit $\bar{M}_{obs}$ la longueur moyenne d'une phase d'observation d'un canal, c'est-à-dire le nombre d'instants, en moyenne, entre le moment où l'utilisateur secondaire décide d'écouter un canal et le moment où il arrête. En suivant la politique π_+ (resp. π_-), l'utilisateur secondaire arrête d'observer un canal dès qu'il a été vu occupé (resp. libre). Donc, sur une phase d'observation de longueur $\bar{M}_{obs}$, sous la politique π_+ (resp. π_-), le canal est libre (resp. occupé) les $\bar{M}_{obs} - 1$ premiers instants et occupé (resp. libre) le dernier. Puisque L canaux peuvent être observés simultanément, la récompense moyenne reçue en suivant la politique π_+ (resp. π_-) est égale à L multipliée par la proportion du temps où le canal est libre sur les $\bar{M}_{obs}$ instants :

$$\begin{cases} \eta_{\alpha,\beta}^{\pi_+} = L\,\dfrac{\bar{M}_{obs} - 1}{\bar{M}_{obs}}\,, \\ \eta_{\alpha,\beta}^{\pi_-} = L\,\dfrac{1}{\bar{M}_{obs}}\,. \end{cases}$$

La longueur moyenne $\bar{M}_{obs}$ d'une phase d'observation d'un canal dépend de la probabilité p que le canal soit libre au début de la phase. On peut alors écrire $\bar{M}_{obs}$ en fonction de p (voir [Zhao et al., 2008] pour plus de détails) :

$$\begin{cases} \bar{M}_{obs}(p) = \dfrac{1-\beta+p}{1-\beta} & \text{si la politique est } \pi_+, \\ \bar{M}_{obs}(p) = \dfrac{1-p+\alpha}{\alpha} & \text{si la politique est } \pi_-\,. \end{cases}$$

En supposant que N est beaucoup plus grand que L, la récompense moyenne peut être approchée en remplaçant p par la probabilité stationnaire d'occupation ν :

$$\begin{cases} \eta_{\alpha,\beta}^{\pi_+} \approx L \, \dfrac{\nu}{1-\beta+\nu} \, , \\ \eta_{\alpha,\beta}^{\pi_-} \approx L \, \frac{\alpha}{1-\nu+\alpha} \, . \end{cases} \tag{2.23}$$

2.4 Apprentissage par renforcement

Dans tout le début de ce chapitre, tout comme dans la plupart des articles de la littérature sur ce sujet, les informations statistiques concernant le trafic des utilisateurs primaires sont supposées entièrement connues par l'utilisateur secondaire [Liu and Zhao, 2008; Zhao et al., 2008, 2007b]. Cependant, en pratique, celles-ci doivent être estimées d'une certaine manière par l'utilisateur secondaire. Le but du travail que nous présentons dans cette section est de déterminer une politique pour le *modèle d'allocation de canal* présenté dans la section 2.3.2 lorsque l'utilisateur secondaire n'a aucune information préalable sur les paramètres des canaux. [Long et al., 2008] proposent une règle heuristique basée sur le comportement asymptotique des paramètres estimés. [Lai et al., 2008] ont également considéré une tâche d'apprentissage dans le modèle d'accès de canal opportuniste dans le cas où l'utilisation de chaque canal est sans mémoire. De plus, un algorithme efficace a été proposé par [Anantharam et al., 1987a,b] à la fois pour le modèle sans mémoire et le modèle Markovien.

Même lorsque les paramètres des canaux sont connus à l'avance, la recherche d'une politique optimale est complexe. Nous avons présenté dans la section précédente la politique d'indice de Whittle qui est proche de l'optimale et qui permet de limiter la complexité en scindant la tâche d'optimisation en N sous-problèmes. L'indice de Whittle est calculé indépendamment pour chaque canal à travers la résolution du problème de planification dans le *modèle d'écoute de canal*. L'indice obtenu est une fonction non triviale des paramètres α et β. Lorsque ces paramètres sont inconnus et que l'agent dispose seulement d'une estimation de ceux-ci, il semble très difficile de s'appuyer sur la politique d'indice de Whittle pour chercher une politique optimale. Par ailleurs, une autre relaxation du *modèle d'allocation de canal* a été présentée dans la section précédente. Celle-ci consiste à permettre à l'utilisateur secondaire d'écouter tous les canaux à la fois et à payer un prix proportionnel au nombre de canaux qu'il écoute. A la fois la méthode d'indice de Whittle et cette deuxième relaxation sont basés sur le *modèle d'écoute de canal*. Dans cette section, nous présentons donc une stratégie permettant d'agir de manière optimale dans ce modèle lorsque les paramètre du modèle (α, β) sont inconnus.

Nous nous intéressons à un scénario où l'utilisateur secondaire accomplit une première phase d'exploration afin d'estimer les paramètres du canal. Dans une seconde phase, l'utilisateur se contente d'exploiter la politique optimale calculée à partir de ces estimations. Comme dans tout problème d'apprentissage par renforcement, la difficulté principale est d'atteindre un bon équilibre entre exploration et exploitation en arrêtant la phase d'exploration dès que les estimées permettent de déterminer la politique optimale. L'algorithme proposé est évalué à l'aide de bornes du regret (voir paragraphe 1.3.2 ou dans la suite pour une définition du regret).

Dans le domaine de l'apprentissage par renforcement, différentes approches ont été proposées ces dernières années pour équilibrer de manière explicite l'exploration et l'exploitation. Ceci est en particulier le but d'algorithmes dits « model-based » (voir paragraphe 1.3.2). Des bornes de regret pour des MDP à espaces d'états et d'actions finis ont notamment été fournies par [Auer et al., 2009a; Tewari and Bartlett, 2008] (voir chapitre 4 dans lequel ces algorithmes sont décrits). Les bornes données par [Auer et al., 2009a] sont de la forme $C|\mathcal{X}|^2 \log(n)$, où n est l'horizon temporel, $|\mathcal{X}|$ est la taille de l'espace d'état et C est une constante qui dépend

du modèle (inconnu) du MDP. [Auer et al., 2009a] fournissent également une borne uniforme de la forme $C|\mathcal{X}|\sqrt{n \log n}$, où C réfère cette fois à une constante universelle. Comme expliqué dans la section précédente, le *modèle d'écoute de canal* est un POMDP et peut éventuellement être réécrit comme un MDP lorsque l'on connaît les paramètres du modèle (voir section 1.4). L'espace d'état de ce MDP est un ensemble infini. Aucune de ces approches « model-based » ne s'applique donc à ce modèle. Cependant, une stratégie d'apprentissage par renforcement peut être obtenue en exploitant certaines spécificités du modèle, en particulier le fait que (1) l'état est partiellement observable et peut être observé en utilisant l'action d'observation, (2) le modèle est paramétré par un vecteur de paramètres de petite dimension, et (3) les transitions entre les états ne dépendent pas des actions prises par l'agent.

L'*algorithme de pavage* que nous proposons atteint un regret maximal en $C \log(n)$ (où C dépend de la valeur du paramètre) et un regret uniforme en $C(\log n)^{1/3}\, n^{2/3}$ pour le *modèle d'écoute de canal*. A notre connaissance, c'est le premier algorithme qui obtient de si fortes garanties pour un tel modèle.

Le paragraphe 2.4.1 présente l'idée générale de notre algorithme. Dans la section 2.4.2, nous introduisons le cadre plus abstrait de MDP ou POMDP paramétriques dans lequel l'algorithme peut être étudié. Les sections 2.4.3 et 2.4.4 décrivent l'algorithme et ses performances en terme de bornes de regret. L'application à l'accès opportuniste de canal est détaillée dans les sections 2.4.5 et 2.4.6, à la fois pour le *modèle d'écoute de canal* et dans le cas de N canaux stochastiquement identiques.

2.4.1 Idée générale de l'algorithme

Dans la section 2.3.2, nous avons calculé la récompense moyenne reçue en suivant les politiques qui associent une action à chaque état interne. Cette récompense moyenne dépend des paramètres α et β qui sont maintenant supposés inconnus de l'utilisateur secondaire. Notons $\eta^{\pi}_{\alpha,\beta}$ la récompense moyenne reçue en suivant la politique π sous le modèle déterminé par (α, β). Pour chaque valeur du paramètre (α, β), la politique optimale $\pi^{*}_{\alpha,\beta} = \operatorname{argmax}_{\pi \in \Pi} \eta^{\pi}_{\alpha,\beta}$ peut être calculée. En étudiant la répartition des politiques optimales dans l'espace de paramètre $[0,1] \times [0,1]$, il est possible de déterminer des *zones de politique*. Ces zones sont des régions de l'espace des paramètres qui correspondent à une unique politique optimale. Elles sont représentées sur la figure 2.9 pour une valeur du coût d'écoute du canal $\lambda = 0.3$. Soit $Z_{(m_0,m_1)}$ (resp. Z_{∞}) la région de l'espace de paramètre telle que $\pi_{(m_0,m_1)}$ (resp. π_{∞}) est la politique optimale. Remarquons que pour $\alpha > \lambda$ et $\beta > \lambda$, la politique optimale $\pi_{(1,1)}$ consiste à toujours observer le canal quel que soit l'état de celui-ci ; en effet, la récompense espérée en observant le canal (égale à α ou β) est toujours plus grande que la récompense λ reçue si le canal n'est pas observé (voir figure 2.9). Au contraire, quand $\alpha < \lambda$ et $\beta < \lambda$, il est optimal de ne jamais observer le canal. Lorsque $\beta < \lambda < \alpha$, la corrélation de la chaîne est négative et la politique optimale consiste à observer le canal s'il a été vu occupé et attendre un pas de temps avant de l'observer de nouveau s'il a été vu libre. Pour $\alpha < \lambda < \beta$, il existe une infinité de zones de politique. Chacune d'entre elles consiste à observer le canal s'il était libre à la dernière observation et à attendre $m_0 - 1$ instants avant de l'observer de nouveau sinon. Les valeurs de m_0 sont comprises entre 2 et l'infini.

Supposons que l'utilisateur secondaire connaisse ce pavage de l'espace des paramètres. Alors, il lui suffit de savoir dans quelle zone de politique les paramètres du modèle se trouvent pour déterminer la politique optimale. L'*algorithme de pavage* que nous proposons découle de cette idée. On présente cet algorithme dans un cadre plus général qui ne dépend pas de la forme exacte des zones de politique. Ce cadre abstrait met l'accent sur les caractéristiques du modèle d'écoute optimale ; elles sont résumées dans les hypothèses 1 et 3 de la section 2.4.4 ci-dessous.

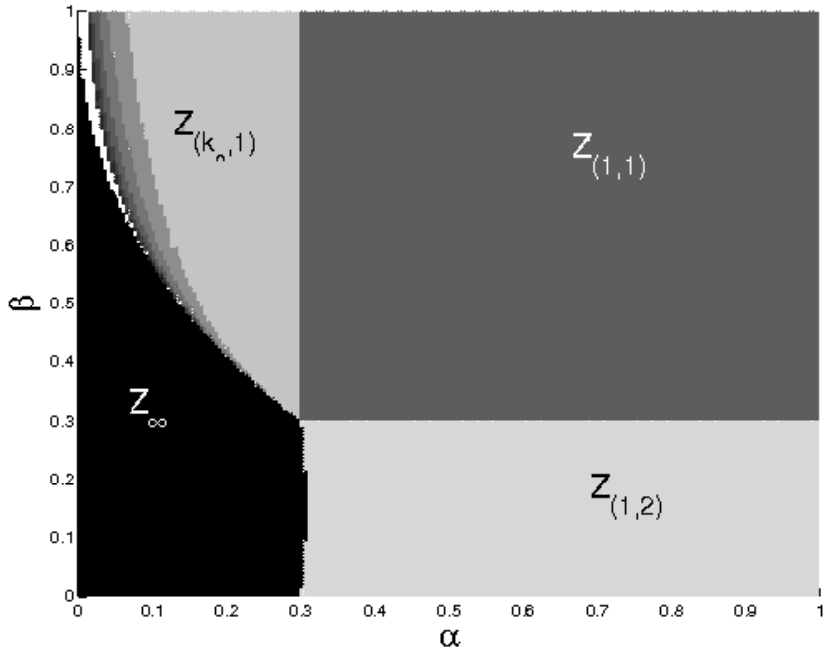

FIGURE 2.9 – Les régions de politique optimales dans le modèle à un canal avec $\lambda = 0.3$.

2.4.2 Modèle

Considérons un POMDP défini par $(\mathcal{X}, \mathcal{A}, \mathcal{Y}, P_\theta, f, r)$, où $\mathcal{X}$ est l'espace d'état discret, $\mathcal{Y}$ est l'espace d'observation, $\mathcal{A}$ est l'espace d'action fini, $P_\theta : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to [0, 1]$ est la probabilité de transition entre les états, $f : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ est la fonction d'observation, $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ est la fonction de récompense bornée et $\theta \in \Theta$ désigne un paramètre inconnu. Étant donné l'état caché courant $x \in \mathcal{X}$ du système, et une action $a \in \mathcal{A}$, la probabilité de l'état suivant $x' \in \mathcal{X}$ est donnée par $P_\theta(x, a; x')$. A chaque instant t, une action A_t est choisie selon la *politique* π suivie qui dépend a priori de l'ensemble des actions $A_{0:t-1}$ et des observations $Y_{0:t-1}$ passées. Le choix de cette action engendre l'observation $Y_t = f(X_t, A_t)$ et la récompense $r(X_t, A_t)$. Notons que l'observation ainsi que la récompense sont supposées dépendre, ici, de manière déterministe de l'état et de l'action. Sans perte de généralité on suppose que pour tout $x \in \mathcal{X}$, pour tout $a \in \mathcal{A}$, $r(x, a) \leq 1$.

Puisque nous nous intéressons à des récompenses cumulées sur un horizon fini mais très grand, nous considérons le critère moyen défini par

$$\eta_\theta^\pi = \lim_{n \to \infty} \frac{1}{n} \mathbb{E}_\theta^\pi \left(\sum_{t=0}^{n-1} r(X_t, A_t) \right) , \tag{2.24}$$

où π désigne une politique fixée.

La notation η_θ^π permet de souligner le fait que la récompense moyenne dépend à la fois de la politique π et de la valeur du paramètre θ. Pour une valeur de paramètre donnée, la récompense optimale à long terme est définie par $\eta_\theta^* = \sup_\pi \eta_\theta^\pi$ et π_θ^* désigne la politique optimale associée. Nous supposons que la dépendance de η_θ^π et π_θ^* par rapport à θ est entièrement connue et qu'il existe une politique particulière π_0 sous laquelle θ peut être estimé de manière consistante. Dans le *modèle d'écoute de canal*, par exemple, cette politique π_0 consiste à observer continuellement le canal de manière à estimer les probabilités de transition en comptant directement les occurrences des états observés.

Étant donné ce qui précède, il est possible de partitionner l'espace des paramètres Θ en des sous-ensembles d'intersection nulle $(Z_i)_i$, appelés *zones de politique*, tels que chaque zone Z_i correspond à une unique politique optimale, notée π_i^* : pour tout $\theta \in Z_i$, $\eta_\theta^* = \eta_\theta^{\pi_i^*}$. Dans chaque zone de politique Z_i, la politique optimale correspondante π_i^* est supposée connue. Il en est de même pour la récompense moyenne optimale $\eta_\theta^{\pi_i^*}$ pour tout $\theta \in \Theta$.

2.4.3 L'algorithme de pavage (AP)

Notons $\hat{\theta}_t$ le paramètre estimé obtenu après t instants et Δ_t la région de confiance associée, dont la construction sera expliquée de manière plus précise ci-dessous. Le principe de l'algorithme de pavage est d'utiliser les zones de politique $(Z_i)_i$ pour déterminer la longueur de la phase d'exploration : essentiellement, la phase d'exploration dure jusqu'à ce que la région de confiance estimée Δ_t soit entièrement contenue dans une des zones de politique. Cependant, il apparaît que ce principe un peu naïf ne permet pas un contrôle suffisant de la durée espérée de la phase d'exploration, et donc, du regret de l'algorithme. En effet, lorsque la valeur du paramètre est très proche du bord d'une zone de politique, il faut attendre potentiellement très longtemps avant que la région de confiance appartienne à cette zone. Pour permettre de traiter les valeurs proches des frontières, nous proposons d'introduire des zones, appelées *zones frontières*, $(F_j(n))_j$ dont la taille dépend de l'horizon temporel fixé n. Si la région de confiance est entièrement contenue dans une de ces zones frontières, la phase d'exploration s'arrête. La figure 2.10 représente le pavage de l'espace des paramètres pour un exemple avec trois zones de politique optimales distinctes. Dans ce cas, il y a quatre zones frontières : une entre chaque couple de zones de politique ($F_1(n)$, $F_2(n)$ et $F_3(n)$) et une autre ($F_4(n)$) pour l'intersection de toutes les zones de politique.

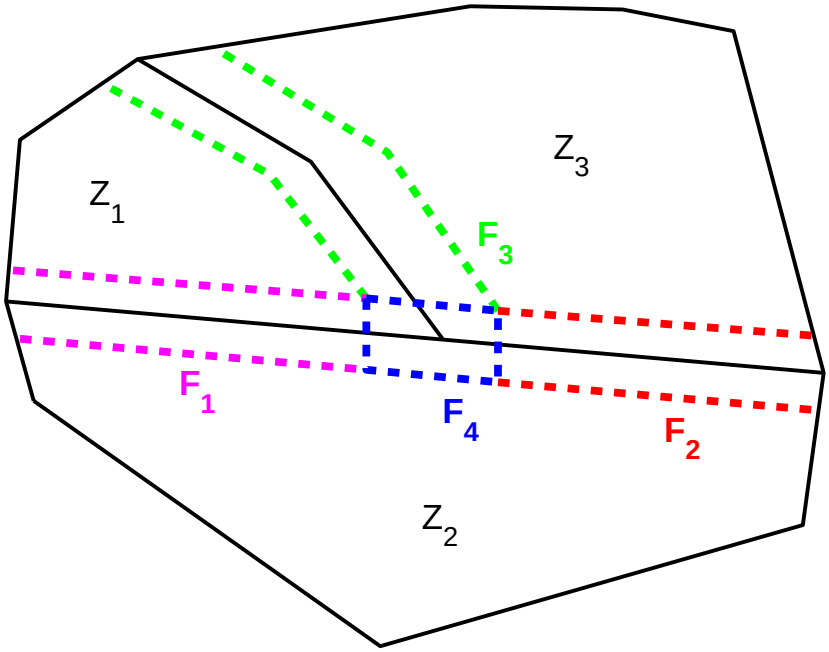

FIGURE 2.10 – Pavage de l'espace des paramètres pour un exemple avec trois zones de politique optimales distinctes

Soit

$$T_n = \inf\{t \geq 1 : \exists i,\ \Delta_t \subset Z_i \text{ ou } \exists j,\ \Delta_t \subset F_j(n)\} \tag{2.25}$$

l'instant aléatoire de la fin de la phase d'exploration. L'algorithme de pavage consiste à utiliser la politique d'exploration π_0 jusqu'à l'occurrence du temps d'arrêt T_n, selon (2.25). A partir de l'instant T_n, l'algorithme sélectionne une politique à utiliser jusqu'à la fin : si, à la fin de la phase d'exploration, la région de confiance est entièrement inclue dans une zone de politique Z_i, alors la politique sélectionnée est π_i^* ; sinon, la région de confiance est inclue dans une zone frontière $F_j(n)$ et la politique sélectionnée est une politique optimale π_k^* *compatible* avec la zone frontière $F_j(n)$. Une politique optimale π_k^* est dite *compatible* avec une zone frontière $F_j(n)$ si l'intersection entre la zone de politique Z_k et la zone frontière est non-vide. Dans l'exemple de la figure 2.10, π_1^* et π_2^* sont compatibles avec la zone frontière $F_1(n)$, alors que toutes les politiques optimales $(\pi_i^*)_{i=1,2,3}$ sont compatibles avec la zone frontière centrale $F_4(n)$. Si la phase d'exploration se termine dans une zone frontière, alors l'algorithme sélectionne simplement une des politiques optimales compatibles avec la zone frontière. Le but des zones frontières est donc de garantir que la phase d'exploration s'arrête même pour des valeurs de paramètre pour lesquelles la discrimination entre certaines politiques optimales est difficile. Bien sûr, en pratique, d'autres considérations pourraient suggérer de sélectionner une politique compatible plutôt qu'une autre mais la borne de regret générale donnée ci-dessous suppose seulement qu'une politique compatible quelconque est sélectionnée à la fin de la phase d'exploration.

2.4.4 Analyse de la performance de l'algorithme

Pour évaluer la performance de l'algorithme proposé, nous majorons l'espérance du regret accumulé jusqu'à un horizon fini n fixé. On rappelle que l'espérance du regret est défini comme la différence entre l'espérance des récompenses cumulées sous la politique optimale et celles obtenues en suivant l'algorithme :

$$\mathbb{E}_{\theta^*}\left[\text{Regret}_n\right] = \mathbb{E}_{\theta^*}^{\pi_{\theta^*}^*}\left[\sum_{t=0}^{n-1} r(X_t, A_t)\right] - \mathbb{E}_{\theta^*}^{\text{AP}}\left[\sum_{t=0}^{n-1} r(X_t, A_t)\right] , \tag{2.26}$$

où θ^* est la valeur inconnue du paramètre. Pour tout sous-ensemble λ de Θ, notons

$$\delta(\lambda) = \sup\left\{\left\|\theta - \theta'\right\|_\infty,\ \theta, \theta' \in \lambda\right\}$$

le diamètre de λ. Pour obtenir des bornes de $\mathbb{E}_{\theta^*}\left[\text{Regret}_n\right]$ qui ne dépendent pas de θ^*, il est nécessaire de faire les hypothèses suivantes :

Hypothèse 2.1. *La région de confiance Δ_t, qui est en général aléatoire, est construite de manière à ce qu'il existe $c_1, c_1', c_3 \in \mathbb{R}_+$ tels que, pour tout $\theta \in \Theta$, et pour tout $c_3 \log(n) \leq t \leq n$,*

$$\mathbb{P}_\theta\left(\theta \in \Delta_t,\ \delta(\Delta_t) \leq c_1 \frac{\sqrt{\log n}}{\sqrt{t}}\right) \geq 1 - c_1' n^{-1/3} .$$

Hypothèse 2.2. *Etant donné un réel positif $\epsilon(n)$, il est possible de construire des zones frontières $(F_j(n))_j$ telles qu'il existe des constantes $c_2, c_2' \in \mathbb{R}_+$ pour lesquelles*

- *$\delta(\Delta_t) \leq c_2\epsilon(n)$ implique qu'il existe soit i tel que $\Delta_t \subset Z_i$ soit j tel que $\Delta_t \subset F_j(n)$,*
- *si $\theta \in F_j(n)$, il existe $\theta_i' \in Z_i$ tel que $\|\theta - \theta_i'\|_\infty \leq c_2'\epsilon(n)$, pour toutes les zones de politiques Z_i compatibles avec $F_j(n)$ (c.a.d, telles que $Z_i \bigcap F_j(n) \neq \emptyset$).*

Hypothèse 2.3. *Pour tout i, il existe $d_i \in \mathbb{R}_+$ tel que pour tout $\theta, \theta' \in \Theta$,*

$$|\eta_\theta^{\pi_i^*} - \eta_{\theta'}^{\pi_i^*}| \leq d_i\left\|\theta - \theta'\right\|_\infty .$$

L'hypothèse 2.1 concerne la construction des régions de confiance. Elle peut généralement être satisfaite en utilisant des inégalités de concentration. La constante 1/3 est arbitraire et a été choisie de manière à coïncider avec le regret espéré minimal dans le pire des cas présenté dans le théorème 2.2 ci-dessous. Une autre constante pourrait être choisie (voir par exemple l'hypothèse 2.4 ci-dessous).

L'hypothèse 2.2 implique que toute région de confiance de diamètre plus petit que $\epsilon(n)$ est entièrement contenue soit dans une zone de politique soit dans une zone frontière, tout en assurant que, localement, la taille de la zone frontière est de l'ordre de $\epsilon(n)$. Plus précisément, dès que le diamètre de la région de confiance Δ_t devient plus petit que la moitié de la largeur des zones frontières, alors Δ_t est forcément entièrement incluse soit dans une zone de politique, soit dans une zone frontière. L'algorithme de pavage dépend de manière cruciale de la construction de ces zones frontières. La figure 2.11 représente l'exemple de pavage précédent et une région de confiance sphérique de diamètre égale à la moitié de la largeur de la frontière F_4. Ainsi, où qu'elle soit centrée, cette boule est toujours contenue, soit dans une zone frontière, soit dans une zone de politique. Ceci ne serait pas vrai si le diamètre de la boule était ne serait-ce qu'un peu plus grand.

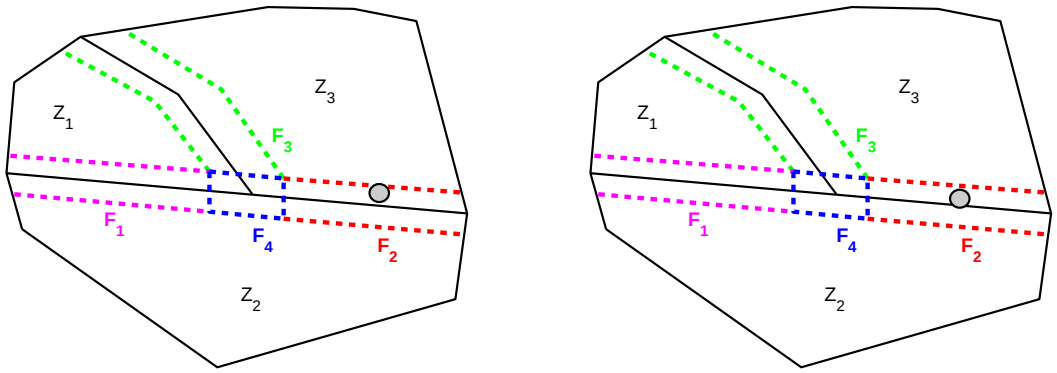

FIGURE 2.11 – Boule de confiance (grise) dans le pavage de l'espace des paramètres présenté précédemment. Le diamètre de la boule étant égale à la moitié de la largeur de la zone frontière, la boule est soit incluse dans une zone de politique (à gauche), soit dans la zone frontière (à droite).

Finalement, l'hypothèse 2.3 est une simple condition de régularité (continuité Lipschitzienne de la fonction de valeur). Le théorème suivant expose la performance de l'approche utilisant un pavage de l'espace des paramètres.

Théorème 2.2. *Sous les hypothèses 2.1, 2.2 et 2.3, et pour tout n, la durée de la phase d'exploration est majorée, en espérance, par*

$$\mathbb{E}_{\theta^*}(T_n) \leq c\,\frac{\log n}{\epsilon^2(n)}\,, \tag{2.27}$$

et l'espérance du regret par

$$\mathbb{E}_{\theta^*}\left[Regret_n\right] \leq \mathbb{E}_{\theta^*}(T_n) + c' n \epsilon(n) + c'' n^{2/3}\,, \tag{2.28}$$

où $c = (c_1/c_2)^2$, $c' = c'_2 \max_{i,k}(d_i + d_k)$ et $c'' = c'_1 + c_3$.

Le regret espéré minimal dans le pire des cas est obtenu en sélectionnant $\epsilon(n)$ de l'ordre de $(\log n/n)^{1/3}$, ce qui permet d'obtenir la borne

$$\mathbb{E}_{\theta^*}\left[Regret_n\right] \leq C(\log n)^{1/3}\, n^{2/3}$$

pour une constante C.

La borne de la durée de la phase d'exploration de l'équation (2.27) vient du fait que l'exploration termine seulement quand la région de confiance, définie dans l'hypothèse 2.1, atteint une taille qui est de l'ordre du diamètre de la frontière, c'est-à-dire $\epsilon(n)$. Le deuxième terme du membre de droite de l'équation (2.28) correspond au regret maximal si la phase d'exploration se termine dans une zone frontière. Le taux $(\log n)^{1/3} n^{2/3}$ est obtenu en équilibrant ces deux termes ($\mathbb{E}_{\theta^*}(T_n)$ et $c' n\epsilon(n)$).

Démonstration. La région de confiance est telle que, pour tout instant $c_3 \log(n) \leq t \leq n$,

$$\mathbb{P}_{\theta^*}\left(\theta^* \in \Delta_t \ , \ \delta(\Delta_t) \leq c_1 \sqrt{\log n}/\sqrt{t}\right) \geq 1 - c_1' n^{-1/3} \ .$$

A la fin de la phase d'exploration, si le vrai paramètre θ^* est dans la région de confiance, deux cas sont possibles : soit la région de confiance Δ_t est contenue dans une zone de politique Z_i, soit elle est incluse dans une zone frontière $F_j(n)$. Si la région de confiance est contenue dans une zone de politique, le regret est la somme de la perte accumulée pendant la phase d'exploration et de la perte correspondant au fait que la région de confiance ne contient pas le vrai paramètre θ^*. La perte accumulée durant la phase d'exploration peut être majorée par la durée de celle-ci. De plus, l'hypothèse 2.1 étant vraie pour $t \geq c_3 \log(n)$, on a donc

$$\mathbb{E}_{\theta^*}\left[\text{Regret}_n\right] \leq c_3 \log(n) + \mathbb{E}_{\theta^*}(T_n) + c_1' n \ n^{-1/3} \ .$$

Si la région de confiance est dans une zone frontière $F_j(n)$, un terme supplémentaire est ajouté au regret. Il s'agit de la perte due au fait que la politique sélectionnée à la fin de la phase d'exploration n'est pas nécessairement l'optimale pour le vrai paramètre θ^*. Soit π_i^* la politique optimale pour θ^* et π_k^* la politique sélectionnée. Notons que Z_i et Z_k sont compatibles avec $F_j(n)$. La perte est alors

$$\eta_{\theta^*}^{\pi_i^*} - \eta_{\theta^*}^{\pi_k^*} = (\eta_{\theta^*}^{\pi_i^*} - \eta_{\theta}^{\pi_i^*}) + (\eta_{\theta}^{\pi_k^*} - \eta_{\theta^*}^{\pi_k^*}) + (\eta_{\theta}^{\pi_i^*} - \eta_{\theta}^{\pi_k^*}) \ ,$$

pour tout $\theta \in Z_k \bigcap F_j(n)$. Le dernier terme est négatif puisque π_k^* est la politique optimale pour θ. Les deux autres termes peuvent être majorés en utilisant l'hypothèse 2.3. Alors,

$$|\eta_{\theta^*}^{\pi_i^*} - \eta_{\theta^*}^{\pi_k^*}| \leq (d_i + d_k) \|\theta^* - \theta\|_\infty \ .$$

La région de confiance étant entièrement incluse dans la zone frontière $F_j(n)$, les paramètres θ et θ^* appartiennent tous les deux à celle-ci. D'après l'hypothèse 2.2, on peut donc choisir θ tel que $\|\theta^* - \theta\|_\infty < c_2' \epsilon(n)$ alors

$$\mathbb{E}_{\theta^*}\left[\text{Regret}_n\right] \leq c_3 \log(n) + \mathbb{E}_{\theta^*}(T_n) + n c' \epsilon(n) + c_1' n \ n^{-1/3} \ ,$$

où $c' = c_2' \max_{i,k}(d_i + d_k)$.

Le regret maximal est obtenu quand la région de confiance est contenue dans une zone frontière. D'après les hypothèses 2.1 et 2.2, si t satisfait $c_1(\log n/t)^{1/2} < c_2 \epsilon(n)$ alors $t \geq T_n$ avec grande probabilité. Donc, $\mathbb{E}_{\theta^*}(T_n) \leq (c_1^2 \log n)/(c_2 \epsilon(n))^2$. L'espérance du regret est alors majorée par

$$\max_{\theta^*} \mathbb{E}_{\theta^*}\left[\text{Regret}_n\right] \leq c_3 \log(n) + \frac{c_1^2 \log n}{c_2^2 \epsilon^2(n)} + n c' \epsilon(n) + c_1' n^{2/3} \ ,$$

qui est minimisé pour $\epsilon(n) = \left(\frac{2 c_1^2 \ \log n}{c_2^2 c' \ n}\right)^{1/3}$.

□

Un examen plus précis de la preuve montre que si l'on peut assurer que l'exploration termine dans une des zones de politique Z_i, l'espérance du regret peut être bornée par une expression similaire à (2.28) mais sans le terme $c'n\epsilon(n)$. Dans ce cas, en modifiant légèrement l'hypothèse 2.1, on peut obtenir des bornes de regret logarithmiques.

Hypothèse 2.4. *L'intervalle de confiance Δ_t est construit de manière à ce qu'il existe $c_1, c_1', c_3 \in \mathbb{R}_+$ tels que, pour tout $\theta \in \Theta$, pour tout n, pour tout $c_3 \log(n) \leq t \leq n$, et tout $x > 1$,*

$$\mathbb{P}_\theta \left(\theta \in \Delta_t,\ \delta(\Delta_t) \leq c_1 \frac{\sqrt{x}}{\sqrt{t}} \right) \geq 1 - c_1' \exp\{-2x\} \ .$$

Il est cependant nécessaire d'introduire des contraintes supplémentaires pour garantir que l'exploration termine dans une zone de politique plutôt que dans une zone frontière. Ces contraintes prennent typiquement la forme d'une distance minimale entre la valeur du vrai paramètre θ^* et les bords de la zone de politique associée. Ce résultat est formalisé par le théorème suivant.

Théorème 2.3. *Soit θ^* un paramètre dans une zone de politique Z tel qu'il existe κ pour lequel*

$$\min_{\theta \notin Z} \|\theta^* - \theta\|_\infty > \kappa \ .$$

Sous les hypothèses 2.2, 2.3 et 2.4, l'espérance du regret est majorée par

$$\mathbb{E}_{\theta^*} \left[Regret_n \right] \leq C(\kappa) \log(n) + C'(\kappa)$$

pour tout n et pour des constantes $C(\kappa)$ et $C'(\kappa)$ qui décroissent strictement lorsque κ augmente.

Démonstration. La condition $\min_{\theta \notin Z} \|\theta^* - \theta\|_\infty > \kappa$ signifie que la distance entre θ^* et tout bord de la zone de politique Z est plus grande que κ. Donc, dès que $\delta(\Delta_t) \leq \kappa$, la région de confiance Δ_t est contenue dans la zone de politique Z. L'espérance du regret de l'algorithme de pavage est alors majorée par $c_3 \log(n) + \mathbb{E}_{\theta^*}(T_n) + c_1' n \exp\{-2x\}$. D'après l'hypothèse 2.4, si t satisfait $c_1 (x/t)^{1/2} < \kappa$ alors $t \geq T_n$ avec grande probabilité. Donc, $\mathbb{E}_{\theta^*}(T_n) \leq c_1 x / \kappa^2$ et l'espérance du regret est majorée par

$$c_3 \log(n) + \frac{c_1 x}{\kappa^2} + c_1' n \exp\{-2x\} \ ,$$

qui est minimisé pour

$$x = \frac{\log(2c_1' n \kappa^2 / c_1^2)}{2} \ .$$

Pour cette valeur de x, on a

$$\mathbb{E}_{\theta^*} \left[\text{Regret}_n \right] = \frac{c_1^2}{2\kappa^2} (\log(n) + \log(2c_1' \kappa^2 / c_1^2) + 1) + c_3 \log(n) \ .$$

□

2.4.5 Application pour le modèle d'écoute de canal

Construction des zones frontières

Dans la suite, nous appliquons l'algorithme de pavage au *modèle d'écoute de canal* présenté dans le paragraphe 2.3.2. Nous avons montré précédemment que, pour une valeur fixée de λ, l'espace des paramètres $[0,1] \times [0,1]$ peut être partitionné en des zones distinctes correspondant aux politiques optimales. Sur la figure 2.9, ce pavage est représenté pour $\lambda = 0.3$. Les

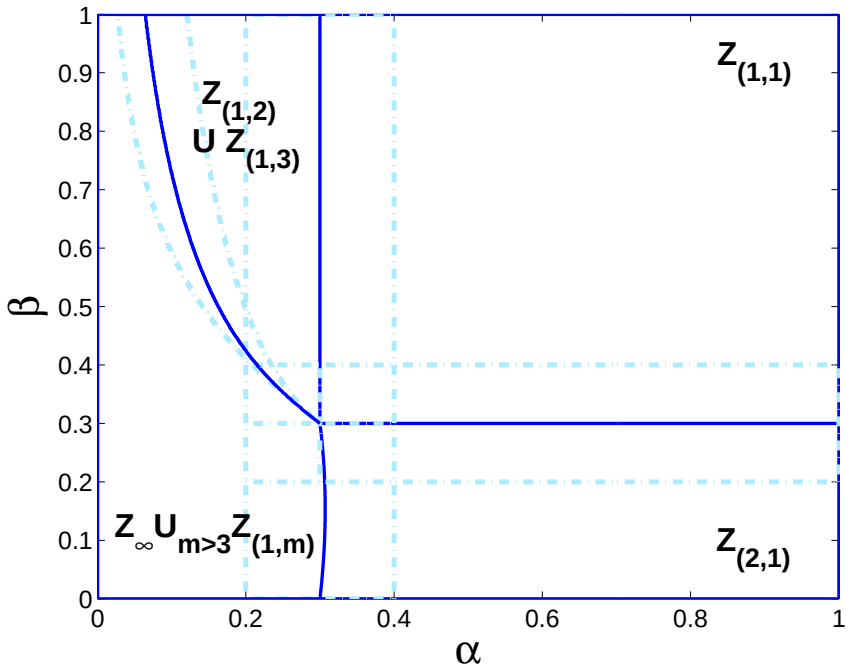

FIGURE 2.12 – Les zones frontières proposées pour $\lambda = 0.3$.

zones frontières peuvent être construites de différentes façon, la seule condition étant qu'elles satisfassent l'hypothèse 2.2.

Une construction possible des zones frontières est représentée sur la figure 2.12. On introduit

- une zone rectangulaire entre $Z_{(1,1)}$ et $Z_{(1,2)}$;
- une zone rectangulaire entre $Z_{(1,1)}$ et $Z_{(2,1)}$;
- une zone rectangulaire entre $Z_{(1,2)}$ et Z_{∞} ;
- une zone carrée centrée au point de jonction (λ, λ) des trois zones de frontière.

La largeur $\epsilon(n)$ des zones frontières rectangulaires dépend de l'horizon temporel n. Comme mentionné précédemment, il existe une accumulation de zones de politique dans le coin en haut à gauche (voir figure 2.9). Pour résoudre cette difficulté, nous fusionnons les zones $Z_{(2,1)}$ et $Z_{(3,1)}$. De plus, nous agrégeons la zone de non-observation Z_{∞} avec les zones $Z_{(m_0,1)}$ pour $m_0 \geq 4$. On introduit alors une zone frontière entre ces zones qui est l'union $Z_{(3,1)} \cup Z_{(4,1)} \cup Z_{(5,1)}$. L'équation des deux courbes délimitant cette zone frontière sont

$$\eta_{\alpha,\beta}^{\pi_{(2,1)}} = \eta_{\alpha,\beta}^{\pi_{(3,1)}}$$

et

$$\eta_{\alpha,\beta}^{\pi_{(5,1)}} = \eta_{\alpha,\beta}^{\pi_{(6,1)}}$$

(voir l'équation (2.20)). Plus de zones pourraient bien évidemment être construites mais, en pratique, l'algorithme proposé est déjà satisfaisant en faisant cette simple agrégation. En effet, la fonction de valeur a des variations très limitées dans les zones agrégées.

Il est important de noter que la construction du pavage n'a besoin d'être faite qu'une seule fois, avant l'estimation des paramètres. L'algorithme de pavage consiste ensuite à estimer le paramètre $\theta = (\alpha, \beta)$ jusqu'à ce que la région de confiance soit entièrement contenue dans une zone de politique ou dans une zone frontière. La politique d'observation, notée π_0, consiste à observer en permanence le canal. Ainsi, le paramètre peut être facilement estimé en comptant

directement les transitions entre les états : à l'instant t, le paramètre estimé est donné par

$$\hat{\alpha}_t = \frac{N_t^{0,1}}{N_t^0} \text{ et } \hat{\beta}_t = \frac{N_t^{1,1}}{N_t^1} , \tag{2.29}$$

où N_t^0 (resp. N_t^1) est le nombre de visites à l'état 0 (resp. 1) avant l'instant t et $N_t^{0,1}$ (resp. $N_t^{1,1}$) est le nombre de visites à l'état 0 (resp. 1) suivies d'une visite à l'état 1 avant l'instant t. On utilise la convention que $\hat{\alpha}_t = 0$ (resp. $\hat{\beta}_t = 0$) si $N_t^0 = 0$ (resp. $N_t^1 = 0$). Une fois la phase d'exploration terminée, l'utilisateur secondaire suit la politique optimale associée au paramètre estimé.

Vérification des hypothèses

Afin de vérifier que le modèle satisfait les conditions du théorème 2.2, il est nécessaire de faire une hypothèse d'irréductibilité sur la chaîne de Markov.

Hypothèse 2.5. *Il existe* $\zeta > 0$ *tel que* $(\alpha, \beta) \in \Theta \stackrel{\text{def}}{=} [\zeta, 1-\zeta]^2$.

Cette hypothèse permet de majorer le temps de retour $\tau_x = \inf\{t \geq 1, X_t = x\}$ à l'état $x \in \{0,1\}$ partant de $x' \in \{0,1\}$: $\mathbb{E}_{x'}[\tau_x] \leq 1/\zeta$. Notons que cette majoration est indépendante de la politique suivie. La quantité ζ est reliée à la notion de *diamètre* introduite par [Auer et al., 2009a] dans le cas des MDP.

On définit la région de confiance suivante :

$$\Delta_t = \left[\hat{\alpha}_t \pm \sqrt{\frac{\log n}{3N_t^0}}\right] \times \left[\hat{\beta}_t \pm \sqrt{\frac{\log n}{3N_t^1}}\right] . \tag{2.30}$$

Notons que Δ_t est une quantité aléatoire puisqu'elle dépend du nombre de fois où l'état a été vu libre et occupé.

Pour prouver que l'espérance du regret cumulé en suivant l'algorithme de pavage dans le *modèle d'écoute de canal* est borné, il est nécessaire de vérifier que chacune des trois hypothèses du théorème 2.2 sont satisfaites. Commençons par montrer que l'hypothèse 2.1 est vérifiée. Il s'agit de prouver qu'il existe c_1, c_1', c_3 deux réels positifs tels que, pour tout $c_3 \log(n) \leq t \leq n$

$$\mathbb{P}_{(\alpha,\beta)}\left((\alpha,\beta) \in \Delta_t,\ \delta(\Delta_t) \leq c_1 \frac{\sqrt{\log n}}{\sqrt{t}}\right) \geq 1 - c_1' n^{-1/3} .$$

On a

$$P_{(\alpha,\beta)}\left((\alpha,\beta) \in \Delta_t,\ \delta(\Delta_t) \leq c_1 \frac{\sqrt{\log n}}{\sqrt{t}}\right) \geq 1 - P_{(\alpha,\beta)}\left((\alpha,\beta) \notin \Delta_t\right) - P_{(\alpha,\beta)}\left(\delta(\Delta_t) > c_1 \frac{\sqrt{\log n}}{\sqrt{t}}\right)$$

Le premier terme peut être décomposé de la manière suivante

$$\begin{aligned}
&\mathbb{P}_{(\alpha,\beta)}\left((\alpha,\beta) \notin \Delta_t\right) \\
&\quad \leq \mathbb{P}_{(\alpha,\beta)}\left(|\hat{\alpha}_t - \alpha| > \sqrt{\frac{\log n}{3N_t^0}}\right) + \mathbb{P}_{(\alpha,\beta)}\left(|\hat{\beta}_t - \beta| > \sqrt{\frac{\log n}{3N_t^1}}\right) \\
&\quad \leq \mathbb{P}_{(\alpha,\beta)}\left(|N_t^{0,1} - \alpha N_t^0| > \sqrt{\frac{N_t^0 \log n}{3}}\right) + \mathbb{P}_{(\alpha,\beta)}\left(|N_t^{1,1} - \beta N_t^1| > \sqrt{\frac{N_t^1 \log n}{3}}\right) \\
&\quad \leq \mathbb{P}_{(\alpha,\beta)}\left(\frac{|N_t^{0,1} - \alpha N_t^0|}{\sqrt{N_t^0}} > \sqrt{\frac{\log n}{3}}\right) + \mathbb{P}_{(\alpha,\beta)}\left(\frac{|N_t^{1,1} - \beta N_t^1|}{\sqrt{N_t^1}} > \sqrt{\frac{\log n}{3}}\right) .
\end{aligned}$$

En utilisant l'inégalité de concentration pour des incréments de martingale démontrée par [Garivier and Moulines, 2008] et décrite dans le théorème A.6, on obtient

$$\mathbb{P}_{(\alpha,\beta)}\left(\frac{|N_t^{0,1}-\alpha N_t^0|}{\sqrt{N_t^0}} > \sqrt{\frac{\log n}{3}}\right) \leq 8\log(t)\exp\left\{-\frac{1.99\log(n)}{3}\right\} \leq 10n^{-1/3}\ .$$

En faisant un raisonnement similaire pour le terme concernant $|N_t^{1,1}-\beta N_t^1|$, on montre que

$$\mathbb{P}_{(\alpha,\beta)}\left((\alpha,\beta)\notin\Delta_t\right) \leq 10n^{-1/3}\ .$$

De plus, on remarque que

$$\left\{\delta(\Delta_t) \leq c_1\frac{\sqrt{\log n}}{\sqrt{t}}\right\} \subset \left\{N_t^0 \geq \frac{4t}{3c_1^2},\ N_t^1 \geq \frac{4t}{3c_1^2}\right\}\ . \tag{2.31}$$

Ainsi,

$$\begin{aligned}\mathbb{P}_{(\alpha,\beta)}\left(\delta(\Delta_t) > c_1\frac{\sqrt{\log(n)}}{\sqrt{t}}\right) &\leq \mathbb{P}_{(\alpha,\beta)}\left(N_t^1 < \frac{4t}{3c_1^2}\right) + \mathbb{P}_{(\alpha,\beta)}\left(N_t^0 < \frac{4t}{3c_1^2}\right)\\ &\leq \mathbb{P}_{(\alpha,\beta)}\left(N_t^1 < \frac{\zeta t}{2}\right) + \mathbb{P}_{(\alpha,\beta)}\left(N_t^0 < \frac{\zeta t}{2}\right)\end{aligned}$$

où on a posé $c_1 = 2\sqrt{2}/\sqrt{3\zeta}$. Le théorème 2 de [Glynn and Ormoneit, 2002] (voir le théorème A.7) permet de majorer $\mathbb{P}_{(\alpha,\beta)}\left(N_t^1 < \frac{\zeta t}{2}\right)$. Notons $\nu_1 = \alpha/(1-\alpha+\beta)$ la probabilité stationnaire que le canal soit libre. Remarquons que $\inf_{\alpha,\beta}\nu_1 = \zeta$ et que la constante de minoration $1-|\beta-\alpha|$ est majorée par 2ζ. On a donc

$$\mathbb{P}_{(\alpha,\beta)}\left(N_t^1 < \frac{\zeta t}{2}\right) \leq \mathbb{P}_{(\alpha,\beta)}\left(N_t^1 - \nu_1 t < -\frac{\zeta t}{2}\right) \leq \exp\left\{-\frac{4\zeta^2(\zeta t/2 - 1/\zeta)^2}{2t}\right\}\ .$$

Pour tout $c_3\log(n) \leq t \leq n$, où la dépendance de c en ζ est de $1/\zeta^4$,

$$2\zeta^2(t\zeta/2 - 1/\zeta)^2/t \geq \log(n)/3$$

et donc

$$\mathbb{P}_{(\alpha,\beta)}\left(N_t^1 < \frac{c\zeta t}{2}\right) \leq \mathbb{P}_{(\alpha,\beta)}\left(N_t^1 - \nu_1 t < -\frac{\zeta t}{2}\right) \leq n^{-1/3}\ .$$

En suivant un raisonnement similaire pour majorer la probabilité que N_t^0 soit plus petite que $\zeta t/2$, on obtient que

$$\mathbb{P}_{(\alpha,\beta)}\left(\delta(\Delta_t) > c_1\frac{\sqrt{\log n}}{\sqrt{t}}\right) \leq 2n^{-1/3}\ ,$$

avec $c_1 = 2\sqrt{2}/\sqrt{3\zeta}$. On a donc

$$\begin{aligned}&\mathbb{P}_{(\alpha,\beta)}\left((\alpha,\beta)\in\Delta_t,\ \delta(\Delta_t) \leq c_1\frac{\sqrt{\log n}}{\sqrt{t}}\right)\\ &\quad\geq 1 - \mathbb{P}_{(\alpha,\beta)}\left(\delta(\Delta_t) > c_1\frac{\sqrt{\log n}}{\sqrt{t}}\right) - \mathbb{P}_{(\alpha,\beta)}\left((\alpha,\beta)\notin\Delta_t,\ \delta(\Delta_t) \leq c_1\frac{\sqrt{\log n}}{\sqrt{t}}\right)\\ &\quad\geq 1 - 12n^{-1/3}\ .\end{aligned}$$

Montrons maintenant que le schéma de partition de l'espace des paramètres proposé paragraphe 2.4.5 satisfait l'hypothèse 2.2. La première partie de cette hypothèse requiert que toute

région de confiance de diamètre plus petit que $c_2\epsilon(n)$ soit entièrement contenue, soit dans une zone de politique, soit dans une zone frontière. Ceci est trivialement satisfait pour toutes les zones frontières rectangulaires en prenant $c_2 = 1$. La difficulté concernant la zone frontière non rectangulaire $Z_{(3,1)} \cup Z_{(4,1)} \cup Z_{(5,1)}$ est que la largeur de la zone frontière décroît quand α et β approchent λ. Cependant, la zone centrale permet d'éviter que cela soit problématique. En effet, les deux courbes définies par $\eta_{\alpha,\beta}^{\pi_{(3,1)}} = \eta_{\alpha,\beta}^{\pi_{(4,1)}}$ et $\eta_{\alpha,\beta}^{\pi_{(5,1)}} = \eta_{\alpha,\beta}^{\pi_{(6,1)}}$ interceptent toutes les deux la droite verticale, d'équation $\alpha = \lambda - \epsilon(n)$, délimitant la zone de frontière rectangulaire à gauche. Il suffit alors de choisir c_2 de manière à ce que $c_2\epsilon(n)$ soit égale à la distance entre les deux points d'intersection. De plus, la deuxième partie de l'hypothèse 2.2, qui requiert que la distance entre tout point θ dans la zone frontière et toutes les zones de politique compatibles soit majorée, est satisfaite de manière évidente. Pour finir, pour toute politique optimale, la récompense moyenne, définie par les équations (2.20) et (2.21), est une fonction continue lipschitzienne de (α, β) pour $\alpha, \beta \in [\zeta, 1-\zeta]$, et, donc, la troisième hypothèse est également satisfaite.

Résultats numériques

Comme suggéré par les théorèmes 2.2 et 2.3, la longueur de la phase d'exploration de l'algorithme de pavage dépend de la valeur du vrai paramètre (α^*, β^*). Cette longueur varie également d'une simulation à l'autre puisqu'elle est aléatoire. Pour illustrer ces effets, nous prenons $\lambda = 0.3$ et considérons deux valeurs différentes des paramètres : $(\alpha^*, \beta^*) = (0.8, 0.05)$ est inclus dans la zone de politique $Z_{(1,2)}$ et loin de toute zone frontière, et, $(\alpha^*, \beta^*) = (0.8, 0.2)$ se trouve dans la zone frontière entre $Z_{(1,1)}$ et $Z_{(1,2)}$ et est proche du bord de la zone frontière. La distribution empirique de la longueur de la phase d'exploration dans les deux cas est représentée sur la figure 2.13. Remarquons que les formes de ces deux distributions sont assez différentes et que la moyenne empirique de la longueur de la phase d'exploration est plus petite pour une valeur de paramètre qui est loin de toute zone frontière que pour une valeur au bord d'une zone frontière.

Sur la figure 2.14, nous comparons les regrets cumulés Regret_n^{AP} de l'algorithme de pavage avec les regrets $\text{Regret}_n^{LD}(l_{expl})$ d'un algorithme qui aurait une phase d'exploration de longueur déterministe l_{expl}. On utilise $(\alpha^*, \beta^*) = (0.8, 0.05)$ et deux valeurs différentes de l_{expl} : une plus petite ($l_{expl} = 20$) et l'autre plus grande ($l_{expl} = 300$) que la longueur moyenne de la phase d'exploration qui varie de 40 à 150 pour cette valeur du paramètre (voir figure 2.13). Les algorithmes sont exécutés 4 fois indépendamment et les regrets cumulés correspondant à chaque simulation sont représentés sur la figure 2.14. Remarquons que, (α^*, β^*) étant à l'intérieur d'une zone de politique, le regret de l'algorithme de pavage est nul pendant la phase d'exploitation si la politique optimale pour le vrai paramètre est utilisée. Quand la longueur de la phase d'exploration l_{expl} est suffisamment grande, l'estimation du paramètre est assez précise et donc le regret accumulé pendant la phase d'exploitation est nul. Cependant, une trop grande valeur de l_{expl} augmente le regret pendant la période d'exploration : on observe sur la figure 2.14 que le regret $\text{Regret}_n^{LD}(l_{expl})$ avec $l_{expl} = 300$ est plus grand que Regret_n^{AP}. Quand la longueur déterministe de la phase d'exploration est plus petite que la longueur moyenne de la phase d'exploration de l'algorithme de pavage, soit le paramètre est bien estimé à la fin de la phase d'exploration et le regret $\text{Regret}_n^{LD}(l_{expl})$ est plus petit que Regret_n^{AP}, soit, la valeur estimée est trop loin de la vraie valeur et la politique suivie pendant la phase d'exploitation n'est pas la politique optimale. Dans ce cas, le regret n'est pas nul pendant la phase d'exploitation et $\text{Regret}_n^{LD}(l_{expl})$ est considérablement grand. Ceci peut être observé sur la figure 2.14 : pour trois des quatre simulations, le regret cumulé $\text{Regret}_n^{LD}(l_{expl})$ avec $l_{expl} = 20$ (ligne pointillée) est petit, tandis que pour la quatrième simulation le regret augmente brutalement et continuement.

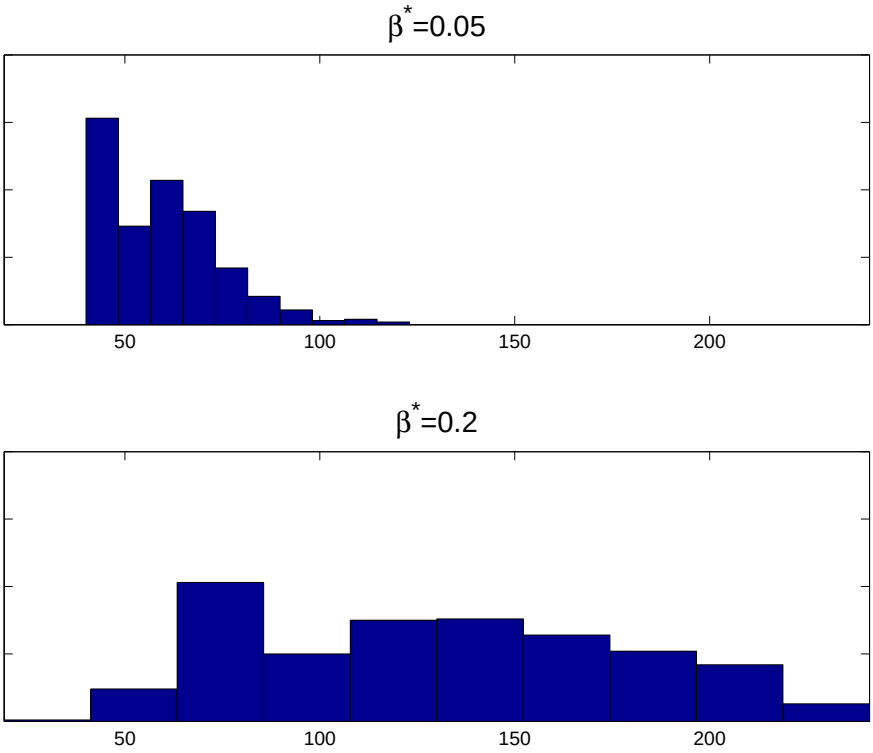

FIGURE 2.13 – Distribution empirique de la longueur de la phase d'exploration en suivant l'algorithme de pavage pour $(\alpha^*, \beta^*) = (0.8, 0.05)$ et pour $(\alpha^*, \beta^*) = (0.8, 0.2)$.

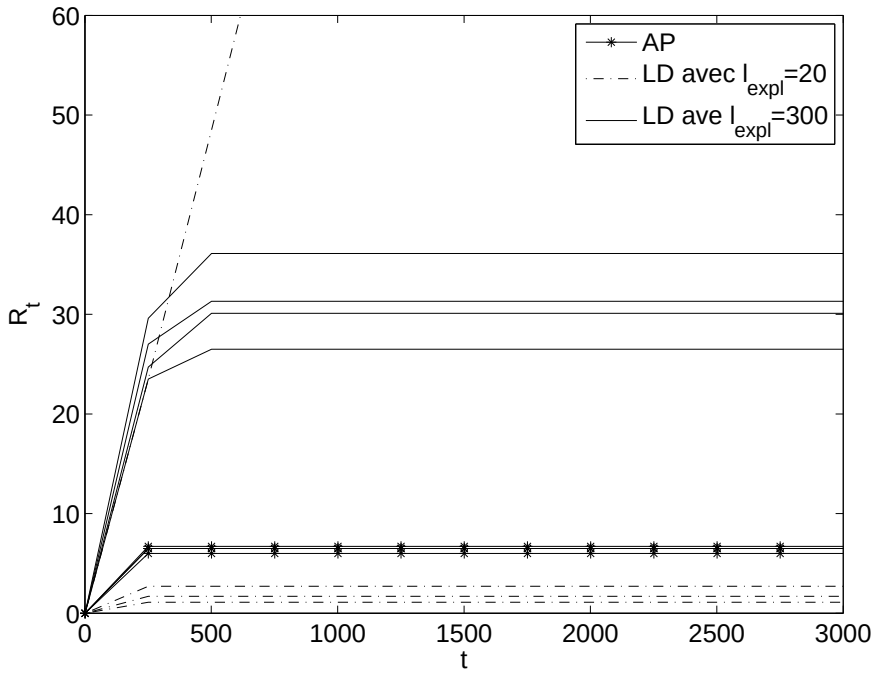

FIGURE 2.14 – Comparaison du regret cumulé pour l'algorithme de pavage (ligne étoilée) et un algorithme avec une phase d'exploration de longueur fixée égale à 20 (ligne pointillée) ou à 300 (ligne continue) pour $(\alpha^*, \beta^*) = (0.8, 0.05)$

2.4.6 Application pour le modèle à N canaux stochastiquement identiques

Construction des zones de politique et de la zone frontière

Comme mentionné dans la section 2.3.3, dans le cas de canaux stochastiquement identiques, pour déterminer la politique optimale l'utilisateur secondaire doit uniquement savoir si α^* est plus petit ou plus grand que β^*. Soient Z_+ et Z_- les zones de politiques correspondant aux politiques optimales π_+ et π_- (voir figure 2.15). Entre ces zones, on introduit une zone frontière $F(n) = \{(\alpha, \beta), |\alpha - \beta| \leq \epsilon(n)\}$. L'estimation du paramètre (α, β) et la

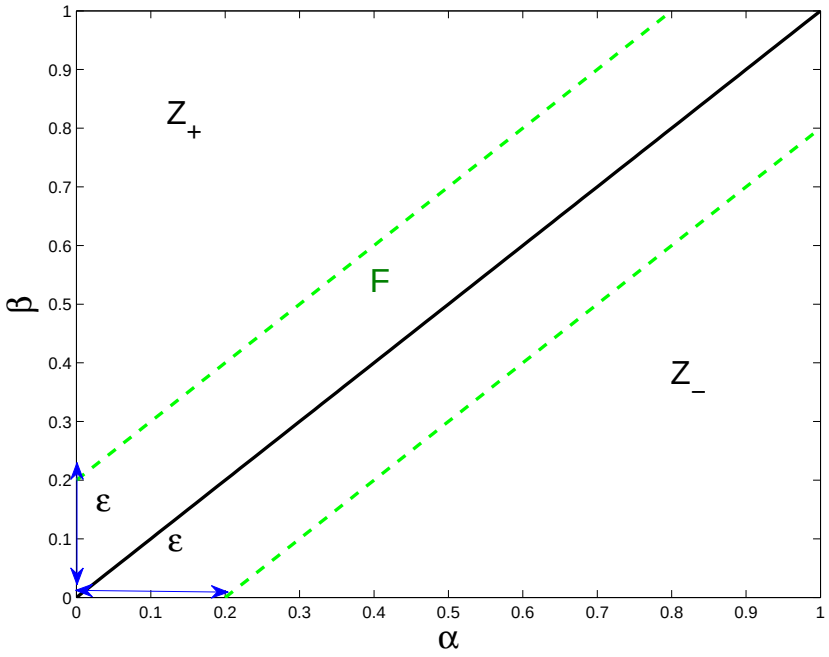

FIGURE 2.15 – Zones de politique et zone frontière pour le modèle à N canaux stochastiquement identiques.

construction de la région de confiance sont similaires au *modèle d'écoute du canal* (voir paragraphe 2.4.5). L'hypothèse 2.1 du théorème 2.2 est alors satisfaite. De plus, étant donné la simplicité de la géométrie de la zone frontière, l'hypothèse 2.2 est facilement vérifiée. En effet, tout rectangle de confiance de longueur inférieure à $\epsilon(n)/2$ est soit contenu dans la zone frontière soit dans une des zones de politique. De plus, pour tout point de la zone frontière, il existe un point d'une zone de politique qui est à une distance inférieure à $\epsilon(n)$ et est aussi dans la zone frontière mais appartient à une autre zone de politique. Pour finir, les approximations de la récompense moyenne égale à $\eta^{\pi_+}_{\alpha,\beta}$ et $\eta^{\pi_-}_{\alpha,\beta}$ définies équation (2.23) sont des fonctions Lipschitziennes et donc la troisième condition de théorème 2.2 est satisfaite [2].

Résultats numériques

Pour illustrer la performance de l'approche, nous utilisons l'algorithme de pavage pour une grille de valeurs de (α^*, β^*) couvrant régulièrement l'ensemble $[\zeta, 1-\zeta]$, avec $\zeta = 0.01$. Pour chaque valeur du paramètre, on procède à 10 réplications Monte-Carlo. L'horizon temporel choisi est $n = 10000$ et la largeur de la zone frontière $\epsilon(n) = 0.15$. La distribution empirique

2. on utilise ici une approximation de la récompense moyenne en suivant une politique plutôt que la valeur exacte.

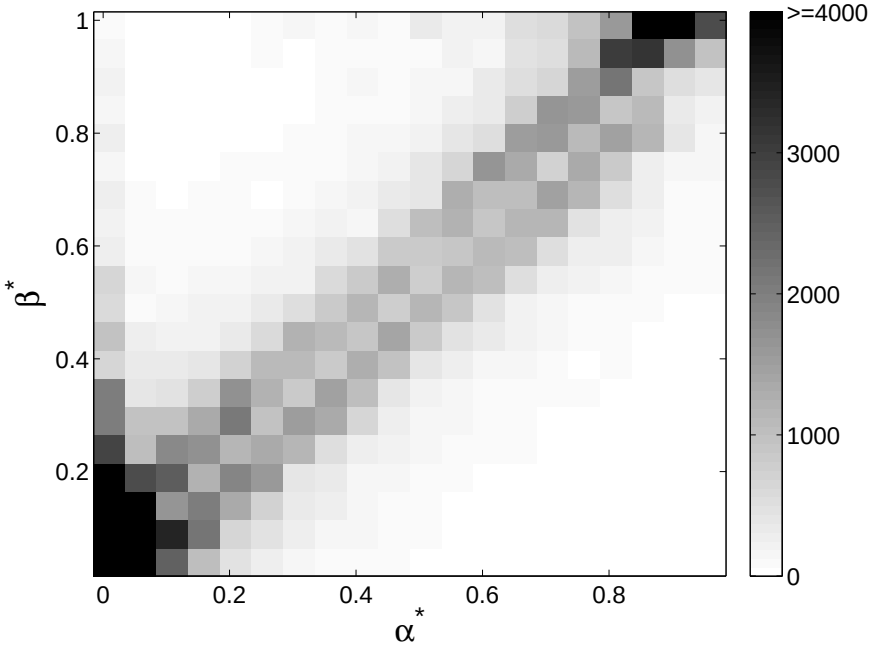

FIGURE 2.16 – Durée de la phase d'exploration de l'algorithme de pavage pour différentes valeurs de (α^*, β^*).

du regret cumulé obtenu ne varie pas beaucoup avec la valeur du paramètre, et vaut en moyenne 90. Cependant, on peut observer que la longueur moyenne de la phase d'exploration T_n, représentée sur la figure 2.16, dépend fortement de la valeur de (α^*, β^*). Tout d'abord, observons que T_n est assez grand pour (α^*, β^*) près de la zone frontière et petit autrement. En effet, quand la vraie valeur du paramètre est loin du bord de la zone de politique, la phase d'exploration dure jusqu'à ce que la région de confiance soit contenue dans la zone de politique correspondante, ce qui est accompli en peu de temps. Remarquons que si la valeur du vrai paramètre est exactement sur le bord des zones frontières, alors les politiques sont équivalentes. Par ailleurs, la phase d'exploration est la plus longue quand (α^*, β^*) est près de $(0, 0)$ ou $(1, 1)$. En effet, quand (α^*, β^*) est autour de $(0, 0)$ (resp. $(1, 1)$), le canal est très souvent occupé (resp. libre) et il est donc très difficile d'estimer β (resp. α). Cet effet est en partie prédit par l'approche asymptotique de [Long et al., 2008]. Dans cet article, les auteurs ont utilisés un *théorème central limite (TCL)* pour montrer que la longueur de la phase d'exploration, pour un canal avec des probabilités de transition (α^*, β^*), doit être égale à

$$l_{expl}(\alpha^*, \beta^*, \delta, P_C) = \frac{(\Phi^{-1}(\frac{P_C+1}{2}))^2}{\delta^2}(1-\alpha^*)\left(\frac{1}{\alpha^*} + \frac{1}{1-\beta^*}\right) \tag{2.32}$$

pour garantir que α soit estimé convenablement (avec un résultat similaire valable pour β). Dans l'équation (2.32), Φ désigne la distribution cumulative de la loi gaussienne centrée réduite et δ et P_C sont des valeurs telles que $P_C = \mathbb{P}(|\hat{\alpha} - \alpha^*| < \delta\alpha^*)$. Cette formule suggère essentiellement que, quand α^* est petit, peu d'observations sont disponibles partant de l'état occupé vers l'état libre et donc l'estimation de α est difficile. Cependant, on peut observer sur la figure 2.16 que la longueur de la phase d'exploration est la plus grande quand *à la fois* α et β sont très petits. La phase d'exploration n'est pas particulièrement longue quand α est petit et β est proche de 1 (coin en haut à gauche sur la figure 2.16). En effet, dans ce deuxième cas, l'état du canal est très persistant, ce qui implique que peu de transitions sont observées. L'estimation de α ou β nécessite donc beaucoup de temps. Cependant, dans ce cas, le canal

est très fortement corrélé et peu d'observations suffisent pour décider quelle est la politique la plus appropriée entre π_+ et π_-.

2.5 Conclusion

L'accès opportuniste aux ressources spectrales pour les radios cognitives peut être modélisé par un processus de décision markovien partiellement observé particulier, aussi appelé « restless bandit », dans lequel une action permet de choisir quelles composantes de l'état du système observer et la transition entre les états est indépendante des actions. Nous avons analysé la tâche de planification dans ce modèle et proposé un algorithme, basé sur *les états de croyance atteignables*, permettant de trouver une politique proche de l'optimale. La proximité entre la fonction de valeur optimale et la fonction de valeur de la politique ainsi obtenue est majorée par un facteur dépendant des probabilités de transition du modèle, et décroissant avec le paramètre de discrétisation choisi.

La complexité de la planification augmentant de manière exponentielle avec le nombre de canaux, des politiques sous-optimales d'indice sont généralement proposées. Ces politiques reposent sur des simplifications du modèle initial et reviennent à se ramener au *modèle d'écoute de canal* dans lequel une chaîne de Markov à deux états évolue dans le temps et l'agent doit payer un coût λ pour observer l'état de la chaîne. Ce modèle est potentiellement intéressant pour d'autres applications tant dans le domaine des télécommunications que dans d'autres domaines où l'utilisateur n'observe pas a priori l'état du système mais peut « payer » pour obtenir une information sur celui-ci.

Dans un deuxième temps, nous nous sommes intéressés au problème d'apprentissage par renforcement dans ce *modèle d'écoute de canal* ainsi que dans un modèle avec N canaux stochastiquement identiques. Nous avons proposé un algorithme original, appelé *algorithme de pavage* composé de deux phases successives : une phase d'exploration puis une phase d'exploitation. L'algorithme adapte la durée respective de ces deux phases en fonction des actions effectuées et des observations passées. Il équilibre de manière adéquate exploration et exploitation afin de garantir une borne de l'espérance du regret en $(\log n)^{1/3}\, n^{2/3}$ dans le pire des cas pour un horizon fini n. De plus, lorsque les probabilités de transition sont suffisamment loin des frontières entre les politiques, l'espérance du regret est logarithmique. Au vu des simulations numériques, il a été observé que l'*algorithme de pavage* est en effet capable d'adapter la longueur de la phase d'exploration selon la séquence d'observation perçues.

L'*algorithme de pavage* tel qu'il a été présenté ne permet pas d'agir dans le modèle général *d'allocation de canal* avec N canaux quelconques. Une perspective intéressante serait d'adapter cette approche de telle manière que les mêmes principes généraux puissent être adaptés au modèle à N canaux.

L'*algorithme de pavage* a été présenté, dans la section 2.4, dans un cadre plus large que le *modèle d'écoute de canal*. Il s'agit d'un POMDP ou MDP à espace d'états et d'actions discrets dans lequel la probabilité de transition entre les états est paramétrée ; ce paramètre doit pouvoir être estimé de manière consistante en suivant une politique connue a priori. De plus, pour construire les zones de politiques, il est nécessaire de pouvoir résoudre la tâche de planification pour toute valeur de paramètre. Ces deux contraintes restreignent fortement l'utilisation possible de cet algorithme. Cependant, dans un modèle de bandit ou dans un MDP à espace d'états et d'actions finis, la politique optimale peut être calculée explicitement à partir du modèle et donc les zones de politique et les zones frontières peuvent être construites. Une extension de ce travail serait alors d'appliquer l'algorithme de pavage à d'autres applications.

CHAPITRE 3

Bandits paramétriques

3.1 Introduction

Nous considérons dans ce chapitre des modèles de bandits paramétriques. Comme mentionné dans le chapitre 1, l'étude des problèmes de bandits est centrale dans l'analyse de la prise de décisions en milieu incertain. Il s'agit d'un des premiers modèles étudiés en apprentissage par renforcement. Rappelons que le problème de bandit classique est un cas particulier de processus de décision markovien avec un seul état. Un agent choisit à chaque instant une action parmi un ensemble $\mathcal{A} = \{1, \ldots, |\mathcal{A}|\}$ de décisions possibles et reçoit une récompense aléatoire tirée selon une distribution déterminée par l'action choisie. L'objectif de l'agent est de choisir les bras à jouer de manière à maximiser la somme des récompenses reçues. La politique optimale, lorsque le modèle est connu, est de jouer le bras ayant la plus grande récompense espérée.

Dans le problème de bandit classique, également appelé problème de bandits indépendants, chaque bras conduit à des récompenses qui sont des réalisations de variables aléatoires distribuées selon une loi spécifique, sans aucun lien d'un bras à l'autre. De nombreux travaux ont été consacrés à ce modèle [Lai and Robbins, 1985; Auer et al., 2002; Cesa-Bianchi and Lugosi, 2006; Audibert et al., 2007]. Récemment, des modèles de bandit structurés dans lesquels les bras sont connectés par un paramètre commun ont suscité un grand intérêt. Ils traduisent des situations dans lesquelles l'éventail des décisions possibles est très large et où le choix d'une action peut permettre de gagner de l'information sur la loi des récompenses associées à d'autres actions. L'interdépendance des bras a été modélisée de différentes manières dans la très récente mais relativement abondante littérature à ce sujet. Certains travaux regroupent les bras en différents *clusters* ayant des caractéristiques communes ; un algorithme, sous un tel modèle, cherche à déterminer le meilleur cluster puis le meilleur bras dans ce cluster [Pandey et al., 2007b; Ortner, 2010]. Des modèles de *bandits paramétriques* où l'espérance de la récompense reçue est une fonction d'un vecteur associé au bras joué et d'un paramètre inconnu ont été proposés par [Auer, 2002; Dani et al., 2008; Rusmevichientong et al., 2009; Rusmevichientong and Tsitsiklis, 2008; Dorard et al., 2009]. Les fonctions qu'ils utilisent sont soit des fonctions linéaires soit des fonctions gaussiennes. D'autres études ayant pour but de modéliser la dépendance entre les bras s'intéressent à une infinité de bras indexés dans un espace métrique [Kleinberg et al., 2008; Bubeck et al., 2009b]. Les bandits dits *contextuels* désignent un autre type de modèle dans lequel l'information disponible varie à chaque instant, la récompense associée à chaque bras dépendant de cette information [Kakade et al.,

2008; Langford and Zhang, 2008; Wang et al., 2005; Li et al., 2010; Pandey et al., 2007a; Slivkins, 2009]. Dans ce cas, contrairement aux autres approches, le bras optimal dépend de l'information contextuelle et est donc susceptible de varier au cours du temps.

Nous nous intéressons ici à des modèles de bandit paramétriques. Nous étudierons ensuite une extension de ces modèles à des bandits contextuels. Dans les modèles de bandits paramétriques, l'agent dispose d'une connaissance a priori sur les bras, celle-ci étant fixe le long de l'expérience. Notons m_a l'information associée au bras a et θ un paramètre inconnu. L'espérance de la récompense reçue lorsque le bras a est joué est :

$$\mathbb{E}\left[R_t \,|\, A_t = a\right] = f_\theta(m_a) \tag{3.1}$$

où f_θ est une fonction paramétrique de l'ensemble $(m_a)_{a\in\mathcal{A}}$ des informations caractéristiques associées à chaque action dans $\mathbb{R}$. [Auer, 2002], [Dani et al., 2008] et [Rusmevichientong and Tsitsiklis, 2008] ont considéré un modèle de bandit paramétrique linéaire où l'information disponible est représentée par un vecteur de $\mathbb{R}^d$. Dans ce cas, pour tout $a \in \mathcal{A}$, $m_a \in \mathbb{R}^d$ et $\theta \in \mathbb{R}^d$ et on réécrit l'équation (3.1) de la manière suivante :

$$\mathbb{E}\left[R_t \,|\, A_t = a\right] = m_a'\theta \ .$$

[Dani et al., 2008] ont montré que la politique proposée par [Auer, 2002] peut être étendue à des modèles ayant un nombre infini de bras et pour lesquels il existe d vecteurs tels que tout vecteur d'information m_a s'écrit comme combinaison linéaire, à coefficients dans $[-1, 1]$, de ces vecteurs. [Rusmevichientong and Tsitsiklis, 2008] ont démontré ce même résultat en utilisant une technique de preuve différente et sous des conditions légèrement plus générales.

Contrairement au cas du bandit linéaire, dans le modèle plus général de bandit paramétrique défini par l'équation (3.1), la tâche d'estimation du paramètre θ à partir des décisions prises et des récompenses reçues dans le passé est complexe, car la fonction f_θ est quelconque. Afin de considérer un modèle assez général dans lequel le paramètre reste estimable, nous avons choisi de nous placer dans le cadre des *modèles linéaires généralisés*. Dans ces modèles, l'espérance de la récompense conditionnellement à l'action a est de la forme

$$\mathbb{E}\left[R_t \,|\, A_t = a\right] = \mu(m_a'\theta)$$

où μ est une fonction croissante non-linéaire appelée *fonction de lien inverse*. Cette généralisation des modèles linéaires permet de considérer une classe plus importante de problèmes. Elle permet, en particulier, d'aborder des cas intéressants où les récompenses sont à valeurs entières ou binaires en considérant, respectivement, des modèles de régression poissonnienne ou logistique.

Le modèle proposé est utile dans de nombreuses applications pour lesquelles une information a priori est disponible pour chaque décision et où une décision engendre une récompense binaire (ou une variable de comptage). De telles situations sont fréquemment rencontrées dans les domaines du marketing et des réseaux sociaux mais aussi en biologie ou en médecine. Prenons l'exemple du problème d'optimisation des ressources publicitaires sur internet, qui a été l'objet de nombreuses recherches ces dernières années (voir par exemple [Abe and Nakamura, 1999; Pandey et al., 2007b; Li et al., 2010]). Dans le modèle de facturation dit *« pay per click »*, le revenu est directement fonction du fait que l'utilisateur consulte ou non l'annonce publicitaire qui lui est présentée [Jank and Shmueli, 2008]. Du point de vue du gestionnaire de site, la sélection de la (ou des) annonce(s) publicitaire(s) à afficher peut être modélisée par un bandit paramétrique, la sélection d'annonces jouant le rôle de bras. L'information spécifique à chaque annonce a, c'est-à-dire le vecteur m_a, correspond alors à des caractéristiques des annonces (par exemple une catégorisation sémantique : « sport », « cinéma », « loisirs », etc.). La récompense est la réaction binaire du visiteur, qui clique ou non sur l'annonce. Dans ce type

d'applications, la dimension d du vecteur de caractéristiques est typiquement petit devant le nombre $|\mathcal{A}|$ d'annonces publicitaires. Cet écart serait d'autant plus grand si l'action ne correspondait non pas à la sélection d'une annonce publicitaire mais au choix de composition d'une page impliquant plusieurs annonces publicitaires choisies dans un panel d'annonces. Dans ce contexte, la régression logistique semble s'imposer pour modéliser la loi des récompenses, et paraît en tout cas plus satisfaisante que l'utilisation d'un simple modèle de régression linéaire, qui ignore la nature binaire des récompenses. Un exemple similaire d'application peut être trouvé dans le domaine de l'évaluation de traitements médicaux, où l'information correspond à la description des différents composants chimiques présents dans les traitements et les récompenses associées sont le résultat du traitement pour chaque sujet, qui est typiquement modélisé par une variable catégorielle.

Pour ce cadre de bandit linéaire généralisé, nous proposons un nouvel algorithme optimiste, appelé GLM-UCB, inspiré de l'approche *Upper Confidence Bound* (UCB) de [Auer et al., 2002], et qui généralise les algorithmes étudiés par [Auer, 2002], [Dani et al., 2008] et [Rusmevichientong and Tsitsiklis, 2008]. Nous présentons une analyse théorique des performances de cet algorithme en terme de regret. En particulier, nous montrons que ces performances dépendent de la dimension du vecteur de paramètres mais pas du nombre de bras, résultat qui n'était connu jusqu'à présent que dans le cas linéaire. On peut souligner le fait que l'approche GLM-UCB utilise la structure particulière de l'estimateur du paramètre dans les modèles linéaires généralisés. Contrairement à l'approche adoptée dans le modèle linéaire, basée sur une région de confiance dans l'espace des paramètres, l'approche GLM-UCB repose sur des intervalles de confiance autour des récompenses espérées pour chacun des bras, ce qui semble être l'approche adéquate lorsque les fonctions considérées sont non-linéaires. Par ailleurs, nous avons remarqué qu'en pratique la performance des algorithmes UCB paramétriques proposés jusqu'à présent est assez décevante sur des horizons modérés lorsque les paramètres garantissant des regrets théoriques faibles sont utilisés. Ceci est dû aux difficultés théoriques apparaissant dans l'analyse de ces algorithmes et aux approximations mathématiques qui semblent inévitables. Pour surmonter cette difficulté, nous expliquons comment régler la largeur de la borne de confiance pour optimiser les performances en pratique. Ce réglage, basé sur des arguments de statistique asymptotique, est renforcé par une analogie avec l'algorithme UCB classique pour lequel les paramètres théoriques fonctionnent bien en pratique.

Nous considérerons dans un deuxième temps des *modèles de bandit paramétriques contextuels*. La différence avec les bandits paramétriques est qu'une nouvelle information est donnée à l'agent à chaque instant. Cette information est appelée contexte. On suppose alors que la loi des récompenses reçues dépend à la fois du bras choisi et du contexte. Plus précisément, on a

$$\mathbb{E}\left[R_t \mid X_t = x, A_t = a\right] = \mu(\Phi(x,a)'\theta) \ ,$$

où Φ est une fonction qui associe un vecteur à un contexte x et une action a. On s'attend à ce que ce modèle soit particulièrement utile pour l'optimisation des ressources publicitaires sur internet. Pour chaque instant, le contexte peut par exemple correspondre à la page requise par l'utilisateur tandis que l'action représente l'annonce publicitaire à afficher. Pour cette application, ce modèle est plus adapté qu'un bandit paramétrique non contextuel car il prend en compte le fait que le taux de clics des publicités dépend de la page visitée et que la publicité la plus cliquée n'est pas forcément la même d'une page à l'autre. Il semble en effet plausible que les visiteurs de deux pages internet ayant des contenus très différents soient intéressés par cliquer sur des publicités différentes. Nous proposons pour ce modèle de bandit contextuel une extension de l'algorithme GLM-UCB, appelée GLM-UCB *Contextuel*, ayant un regret logarithmique. Nous illustrons les performances de l'algorithme en utilisant des données réelles concernant l'activité d'utilisateurs internet fournies par Orange.

Nous formalisons le problème de bandit paramétrique et présentons les modèles linéaires généralisés dans la section 3.2. La section 3.3 est consacrée au descriptif de l'algorithme proposé, qui est comparé en section 3.4 à d'autres approches de référence. La section 3.5.1 contient des garanties théoriques de performance pour notre algorithme ainsi que des éléments permettant d'utiliser en pratique l'algorithme. Dans la section 3.6, nous illustrons l'algorithme sur des exemples simulés et des données réelles. Une extension de l'algorithme est ensuite proposé dans la section 3.7 permettant de considérer des bandits paramétriques contextuels.

3.2 Modèle de bandit linéaire généralisé

Nous considérons un modèle de bandit paramétrique ayant un nombre de bras $|\mathcal{A}|$ fini, mais potentiellement très grand. A chaque instant t, l'agent choisit de jouer un bras $A_t \in \mathcal{A} \stackrel{\text{def}}{=} \{1 \dots |\mathcal{A}|\}$. On suppose que l'agent dispose d'une information a priori qui consiste en une collection de vecteurs de caractéristiques $(m_a)_{a\in\mathcal{A}}$ spécifiques à chaque bras.

Le modèle de bandit linéaire généralisé considéré dans ce travail est basé sur l'hypothèse que, conditionnellement à la séquence des bras sélectionnés, les récompenses R_t sont des variables aléatoires indépendantes qui satisfont

$$\mathbb{E}_{\theta_*}\left[R_t \,|\, A_t = a\right] = \mu(m_a'\theta_*)\ , \tag{3.2}$$

où $\theta_* \in \Theta \subset \mathbb{R}^d$ est un vecteur de paramètres inconnu, et μ est une fonction réelle connue, potentiellement non-linéaire. Ce cadre généralise le modèle de régression linéaire considéré par [Auer, 2002], [Dani et al., 2008] et [Rusmevichientong and Tsitsiklis, 2008] et permet d'aborder des cas de récompenses ayant des structures plus spécifiques tout en profitant du cadre statistique bien connu des modèles linéaires généralisés. Il permet par exemple d'aborder des cas intéressants où les récompenses sont des variables catégorielles ou binaires. Typiquement, pour des variables binaires, un choix convenable de μ est $\mu(x) = \exp(x)/(1+\exp(x))$, conduisant au modèle de régression logistique. Pour des valeurs de récompense entières, le choix $\mu(x) = \exp(x)$ conduit au modèle de régression poissonnien. Ce cadre peut également être étendu au cas de la régression logistique multinomiale (ou polytomique), appropriée pour des situations dans lesquelles les récompenses associées sont des variables catégorielles.

3.2.1 Modèles linéaires généralisés

Avant de présenter plus en détails le modèle considéré, nous rappelons les principales propriétés des modèles linéaires généralisés (GLM) [McCullagh and Nelder, 1989]. On dit qu'une variable aléatoire réelle suit une distribution dans la *famille exponentielle canonique* si sa densité par rapport à une mesure de référence sur $\mathbb{R}$ est donnée par

$$p_\zeta(r) = \exp\left(r\zeta - b(\zeta) + c(r)\right)\ . \tag{3.3}$$

où ζ désigne un paramètre réel, $c(.)$ est une fonction réelle et la fonction $b(.)$ est supposée deux fois continûment différentiable. Cette famille comprend notamment les distributions Gaussiennes et Gamma lorsque la mesure de référence est la mesure de Lebesgue, ainsi que les distributions de Poisson et de Bernoulli lorsque la mesure de référence est la mesure de comptage sur l'ensemble des entiers. Pour une variable aléatoire R de densité définie par l'équation (3.3), $\mathbb{E}(R) = \dot{b}(\zeta)$ et $\mathrm{Var}(R) = \ddot{b}(\zeta)$, où $\dot{b}$ et $\ddot{b}$ désignent, respectivement, la dérivée première et la dérivée seconde de b. De plus, $\ddot{b}(\zeta)$ est égale à la matrice d'information de Fisher pour le paramètre ζ. Ceci implique en particulier que b est une fonction strictement convexe dès que le paramètre ζ est identifiable.

Supposons maintenant qu'en plus de la variable cible R on dispose d'un vecteur de covariables $X \in \mathbb{R}^d$. On définit $p_\theta(r|x)$ la probabilité de la variable R conditionnellement au

vecteur de covariable X. Dans les modèles linéaires généralisés canoniques associés à (3.3), on a $p_\theta(r|x) = p_{x'\theta}(r)$, où $\theta \in \mathbb{R}^d$ est un vecteur de paramètres. Soit $\mu = \dot{b}$ la fonction dite *fonction de lien inverse*. Lorsque la variable R suit une loi de Bernoulli conditionnellement à $\{X = x\}$, alors μ est la fonction Logit inverse $z \mapsto \exp(z)/(1+\exp(z))$; dans le cas d'une distribution de Poisson, $\mu : z \mapsto \exp(z)$, pour une loi gamma, $\mu : z \mapsto 1/z$ et pour une loi gaussienne μ est l'identité. On déduit des propriétés de b que μ est continûment différentiable et strictement croissante. L'estimateur du maximum de vraisemblance $\hat{\theta}_t$, calculé à partir des observations $(X_0, R_0), \dots (X_{t-1}, R_{t-1})$, est défini comme étant la valeur maximisant $\sum_{k=0}^{t-1} \log p_\theta(R_k|X_k)$, c'est-à-dire

$$\hat{\theta}_t = \operatorname*{argmax}_\theta \sum_{k=0}^{t-1} R_k X_k'\theta - b(X_k'\theta) + c(R_k) \ ,$$

qui est strictement concave en θ. En dérivant, on obtient que $\hat{\theta}_t$ est l'unique solution de l'équation suivante

$$\sum_{k=0}^{t-1} \big(R_k - \mu(X_k'\theta)\big) X_k = 0 \ , \tag{3.4}$$

puisque $\mu = \dot{b}$. En pratique, la solution de (3.4) peut être trouvée en utilisant, par exemple, l'algorithme de Newton.

3.2.2 Modèle de bandit linéaire généralisé

Dans le modèle de bandit linéaire généralisé, aucune hypothèse n'est faite sur la distribution conditionnelle de R étant donné le vecteur de caractéristique m_a associé au bras a joué. En particulier, contrairement au modèle linéaire généralisé présenté ci-dessus, on ne suppose pas que la distribution appartient à la famille exponentielle canonique.

A chaque instant t, l'estimateur $\hat{\theta}_t$ que nous considérons est l'unique solution de l'équation suivante

$$\sum_{k=0}^{t-1} \Big(R_k - \mu(m_{A_k}'\hat{\theta}_t)\Big) m_{A_k} = 0 \ , \tag{3.5}$$

où $A_0, \dots, A_{t-1}$ désignent les bras joués jusqu'à l'instant t et $R_0, \dots, R_{t-1}$ sont les récompenses reçues. L'équation (3.5) a été justifiée dans le paragraphe ci-dessus en se plaçant dans un cadre de modèle linéaire généralisé où les distributions de R conditionnellement aux actions choisies appartiennent à la famille exponentielle canonique, mais le même estimateur est également utile dans le contexte où on suppose simplement que $\mathbb{E}_\theta\left[R_t \,|\, A_t\right] = \mu(m_{A_t}'\theta)$. Cet estimateur est alors appelé *estimateur du maximum de quasi-vraisemblance*. Il est connu pour être consistant sous des conditions très générales dès que la matrice $\sum_{k=0}^{t-1} m_{A_k} m_{A_k}'$ est telle que le rapport entre sa plus petite valeur propre et le logarithme de sa plus grande valeur propre tend vers l'infini avec t [Chen et al., 1999]. Comme nous le verrons dans la suite, cette matrice joue un rôle crucial dans l'algorithme que nous proposons dans le cadre de bandits linéaires généralisés.

3.3 L'algorithme GLM-UCB

Selon (3.2), l'espérance de la récompense reçue en jouant un bras $a \in \mathcal{A}$ vaut $\mu(m_a'\theta_*)$ où θ_* est inconnu. Nous appelons *bras optimal* un bras, noté a^*, dont la récompense espérée est la plus grande :

$$a^* \in \operatorname*{argmax}_{a \in \mathcal{A}} \mu(m_a'\theta_*) \ .$$

L'objectif de l'agent est de trouver le plus rapidement possible un bras optimal et de jouer celui-ci de manière à maximiser les récompenses qu'il reçoit. L'agent ne connaissant pas la loi des récompenses, il doit accumuler suffisamment d'expériences en jouant les différents bras avant d'exploiter celui qui lui semble donner la récompense la plus grande. Lorsque le modèle est paramétrique, l'exploration de tous les bras n'est pas nécessaire puisque il suffit d'estimer le paramètre θ pour pouvoir évaluer l'espérance des récompenses reçues $\mu(m'_a\theta)$ pour chaque bras a. Soit $\hat{\theta}_t$ un estimateur du paramètre à l'instant t. On sait que jouer à chaque instant l'*action gloutonne* par rapport au paramètre estimé, c'est-à-dire ici $\operatorname{argmax}_{a\in\mathcal{A}} \mu(m'_a\hat{\theta}_t)$, conduit en général à un algorithme non robuste. En effet, cet algorithme est entièrement basé sur l'exploitation au détriment de l'exploration. Une méthode souvent utilisée pour équilibrer exploration et exploitation consiste à suivre une approche *optimiste* (voir section 1.3.2). Comme décrit dans [Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2008] pour le cas linéaire, un algorithme optimiste consiste à sélectionner, à l'instant t, le bras

$$A_t = \operatorname*{argmax}_a \max_{\theta,\, \|\theta-\hat{\theta}_t\|_{M_t}\leq\rho(t)} \mathbb{E}_\theta\left[R_t \,|\, A_t = a\right] \ , \tag{3.6}$$

où

$$M_t = \sum_{k=0}^{t-1} m_{A_k} m'_{A_k} \tag{3.7}$$

est la matrice de « design » correspondant aux $t-1$ premiers pas de temps et $\|v\|_M = \sqrt{v'Mv}$ est la norme de v pondérée par la matrice M. La fonction ρ est, quant à elle, une fonction faiblement croissante du temps t. L'ensemble des paramètres θ tels que $\|\theta - \hat{\theta}_t\|_{M_t} \leq \rho(t)$ est une région de confiance ellipsoïdale autour du paramètre estimé $\hat{\theta}_t$. La généralisation de cette approche à des fonctions de liens non linéaires soulève deux sérieux problèmes. Tout d'abord, la maximisation de $\mu(m'_a\theta)$ sur une région de confiance ellipsoïdale pour une fonction μ non convexe n'est pas évidente tant en théorie qu'en pratique. Or la fonction μ peut tout à fait être non convexe : par exemple, la fonction logit-inverse (utilisée dans le cas de variables de Bernoulli) est log-concave. De plus, dans les modèles linéaires généralisés, les régions de confiance appropriées peuvent parfois avoir des géométries dans l'espace des paramètres beaucoup plus compliquées que de simples ellipsoïdes. Cette formulation de l'algorithme optimiste paraît donc périlleuse.

Une approche alternative consiste à déterminer directement une borne de confiance pour la récompense espérée de chaque bras, en choisissant l'action a qui maximise

$$\mathbb{E}_{\hat{\theta}_t}\left[R_t \,|\, A_t = a\right] + \rho(t)\|m_a\|_{M_t^{-1}} \ .$$

Dans le cas linéaire, ces deux approches sont équivalentes [Rusmevichientong and Tsitsiklis, 2008]. En effet, pour une action a fixée, la valeur de θ qui maximise $\mathbb{E}_\theta\left[R_t \,|\, A_t = a\right] = \theta' m_a$ sous la contrainte que $\left\|\theta - \hat{\theta}_t\right\|_{M_t} \leq \rho(t)$ vérifie le système d'équation suivant

$$\begin{cases} \dfrac{dL(\theta,\lambda)}{d\theta} \overset{\text{def}}{=} m_a + 2\lambda M_t(\theta - \hat{\theta}_t) = 0 & (3.8) \\ \left\|\theta - \hat{\theta}_t\right\|_{M_t} \leq \rho(t) & (3.9) \\ \lambda \geq 0 & (3.10) \end{cases}$$

où le lagrangien L s'écrit $L(\theta,\lambda) \overset{\text{def}}{=} m'_a\theta + \lambda\left(\left\|\theta - \hat{\theta}_t\right\|^2_{M_t} - \rho(t)^2\right)$. On déduit des équations (3.8) et (3.9) que

$$\lambda = \frac{(\theta - \hat{\theta}_t)' m_a}{2\rho(t)^2} \ .$$

En injectant ce résultat dans l'équation (3.8) réecrite, en multipliant les deux membres par $m_a' M_t^{-1}$, de la manière suivante $m_a' M_t^{-1} m_a + 2\lambda m_a'(\theta - \hat{\theta}_t) = 0$ on obtient que θ vérifie l'équation

$$\theta' m_a = \hat{\theta}_t' m_a + \rho(t)\|m_a\|_{M_t^{-1}} = \mathbb{E}_{\hat{\theta}_t}\left[R_t \,|\, A_t = a\right] + \rho(t)\|m_a\|_{M_t^{-1}} \,.$$

Dans le cas non-linéaire, la seconde approche, qui cherche directement à maximiser la récompense espérée semble plus appropriée. On propose donc d'utiliser cette seconde approche pour le modèle de bandit linéaire généralisé défini par l'équation (3.2) en incluant un bonus d'exploration de la forme $\rho(t)\|m_a\|_{M_t^{-1}}$. L'estimateur du maximum de quasi-vraisemblance dans le modèle linéaire généralisé est l'unique solution de l'équation (3.5).

Soit $\mathcal{M} = \mathrm{Vect}(m_a, a \in \mathcal{A})$ l'espace vectoriel engendré par les vecteurs d'information associés aux actions. Sans perte de généralité, on peut supposer que la dimension de $\mathcal{M}$ est égale à d. On note $a_0, \ldots, a_{d-1} \in \mathcal{A}$ un ensemble de d actions telles que les vecteurs $m_{a_0} \ldots, m_{a_{d-1}}$ forment une base de $\mathcal{M}$. Pour garantir que la matrice M_t soit inversible pour tout $t \geq d$, l'agent commence par jouer les actions $a_0, \ldots, a_{d-1}$. L'algorithme proposé, nommé dans la suite GLM-UCB, est décrit ci-dessous.

Algorithme 3.1 Algorithme GLM-UCB

1: Jouer les actions $a_0, \ldots, a_{d-1}$
2: Réception des récompenses $R_0, \ldots, R_{d-1}$.
3: **Pour** $t \geq d$ **faire**
4: Calculer $\hat{\theta}_t$
5: Jouer l'action $A_t = \mathrm{argmax}_a\, \mu(m_a'\hat{\theta}_t) + \beta_t^a(\delta)$
6: Réception de la récompense R_t
7: **fin Pour**

A l'instant t, pour chaque bras a, une borne supérieure de $\mu(m_a'\hat{\theta}_t)$ égale à $\mu(m_a'\hat{\theta}_t) + \beta_t^a$ est calculée. Le bonus d'exploration $\beta_t^a = \rho(t)\|m_a\|_{M_t^{-1}}$ est le produit de deux termes. La quantité $\rho(t)$ est une fonction faiblement croissante ; nous montrons dans la section 3.5.1 que $\rho(t)$ peut être fixée de manière à garantir une borne sur le regret espéré. Cependant, comme nous le verrons dans la section suivante, le terme principal dans β_t^a est $\|m_a\|_{M_t^{-1}}$, qui décroît vers 0 quand t croît.

3.4 Discussion

Pour comprendre le comportement de l'algorithme GLM-UCB, il est primordial d'analyser le rôle joué par le terme $\|m_a\|_{M_t^{-1}}$ présent dans le bonus d'exploration β_t^a. Nous commençons par examiner le cas général où le nombre de bras est beaucoup plus grand que la dimension d du problème puis nous analysons le comportement de l'algorithme GLM-UCB dans le cadre plus simple des bandits indépendants.

3.4.1 Influence du nombre de bras sur le regret

L'algorithme que nous proposons exploite l'interdépendance des bras modélisée par le paramètre $\theta \in \mathbb{R}^d$. Il est alors principalement intéressant dans des modèles où le nombre de bras $|\mathcal{A}|$ est grand devant la dimension d de l'espace des paramètres. Notons en particulier que, contrairement à un algorithme tel que UCB, l'algorithme 3.1 ne nécessite pas forcément de jouer tous les bras.

Pour comprendre ce phénomène, observons que $M_{t+1} = M_t + m_{A_t} m'_{A_t}$. Donc, d'après le lemme d'inversion matricielle [Horn and Johnson, 1990]

$$M_{t+1}^{-1} = (M_t + m_{A_t} m'_{A_t})^{-1} = M_t^{-1} - \frac{M_t^{-1} m_{A_t} m'_{A_t} M_t^{-1}}{1 + \|m_{A_t}\|^2_{M_t^{-1}}} .$$

On en déduit que, pour tout bras a,

$$\|m_a\|^2_{M_{t+1}^{-1}} = \|m_a\|^2_{M_t^{-1}} - \frac{\left(m'_a M_t^{-1} m_{A_t}\right)^2}{1 + \|m_{A_t}\|^2_{M_t^{-1}}} .$$

Ainsi, à chaque instant t et pour chaque bras a, $\|m_a\|_{M_{t+1}^{-1}}$ est inférieur ou égal à $\|m_a\|_{M_t^{-1}}$. Ceci est vrai que le bras a ait été joué à l'instant t ou pas. Les seuls bras pour lesquels le terme du bonus d'exploration $\|m_a\|_{M_{t+1}^{-1}}$ ne décroît pas strictement sont les bras orthogonaux à m_{A_t} dans la métrique induite par M_t^{-1}, c'est-à-dire les bras tels que $m'_a M_t^{-1} m_{A_t} = 0$. Inversement, la décroissance est d'autant plus significative que la valeur absolue du produit scalaire $|m'_a M_t^{-1} m_{A_t}|$ est grande. Ceci permet de comprendre pourquoi les bornes de regret présentées dans les théorèmes 3.1 et 3.2 ci-dessous dépendent de d mais pas de $|\mathcal{A}|$.

3.4.2 Généralisation de l'algorithme UCB

L'algorithme UCB pour $|\mathcal{A}|$ bras [Auer et al., 2002] (voir algorithme 1.8) peut être vu comme un cas particulier de l'algorithme GLM-UCB quand les bras sont linéairement indépendants. En effet, pour $d = |\mathcal{A}|$, on peut définir les vecteurs $\{m_a\}_{a\in\mathcal{A}}$ comme étant la base canonique de $\mathbb{R}^d$, et poser $\theta \in \mathbb{R}^d$ le vecteur dont chaque composante θ_a correspond à la récompense espérée en jouant le bras a.

Sous ces conditions, la matrice M_t est une matrice diagonale $d \times d$ dont le a-ième élément diagonal est $N_t(a)$, c'est-à-dire le nombre de fois où le bras a a été joué avant l'instant t. La matrice M_t est donc inversible dès que tous les bras ont été joués une fois. Et, pour tout $a \in \mathcal{A}$,

$$\|m_a\|_{M_t^{-1}} = \sqrt{m'_a M_t^{-1} m_a} = 1/\sqrt{N_t(a)} .$$

Le bonus d'exploration est donc de la forme

$$\beta_t^a = \rho(t)/\sqrt{N_t(a)} .$$

De plus, l'estimateur $\hat{\theta}_t$ du maximum de quasi-vraisemblance vérifie $\bar{R}_t^a = \mu(\hat{\theta}_t(a))$ pour tout $a \in \mathcal{A}$, où $\bar{R}_t^a = \frac{1}{N_t(a)} \sum_{k=0}^{t-1} \mathbb{1}_{\{A_k=a\}} R_k$ est la moyenne estimée des récompenses reçues en jouant le bras a.

Lorsque les bras sont linéairement indépendants, l'algorithme 3.1 est donc exactement l'algorithme UCB [Auer et al., 2002]. Dans ce cas, on sait que le regret espéré peut être contrôlé en prenant $\rho(t)$ égal à $\sqrt{\log(t)/2}$ lorsque les récompenses sont majorées par 1.

3.5 Résultats théoriques

3.5.1 Analyse du regret

Pour évaluer l'algorithme proposé, nous calculons le regret espéré à horizon fini qui, pour tout horizon n, compare l'espérance des récompenses accumulées en suivant l'algorithme à l'espérance des récompenses reçues par un agent *oracle* qui joue en permanence un bras

optimal. Notons $\theta_* \in \Theta$ la vraie valeur du paramètre, inconnue de l'agent mais connue de l'oracle. Nous nous intéressons ici à la notion suivante de regret :

$$\text{Regret}_n = \sum_{t=0}^{n} \mu(m'_{a_*}\theta_*) - \mu(m'_{A_t}\theta_*)$$

où on rappelle que $a_* = \text{argmax}_{a\in\mathcal{A}}\, \mu(m'_a\theta_*)$ désigne un bras optimal.

Certaines hypothèses de régularité sur la fonction de lien, ainsi que sur les vecteurs d'informations caractéristiques et sur les récompenses, sont nécessaires pour analyser le regret de l'algorithme GLM-UCB.

Hypothèse 3.1. *L'ensemble $\Theta \subset \mathbb{R}^d$ est un ensemble compact et convexe contenant le paramètre θ_*.*

Hypothèse 3.2. *La fonction de lien $\mu : \mathbb{R} \to \mathbb{R}$ est continûment différentiable, lipschitzienne de constante de Lipschitz k_μ. L'espace des paramètres Θ et les vecteurs d'information sont tels que $c_\mu = \inf_{\theta\in\Theta, a\in\mathcal{A}} \dot{\mu}(m'_a\theta) > 0$.*

Cette hypothèse est satisfaite par les fonctions de lien usuellement utilisées dans les modèles linéaires généralisés. Pour la fonction logistique, par exemple, $k_\mu = 1/4$ et c_μ dépend de $\sup_{\theta\in\Theta, a\in\mathcal{A}} |m'_a\theta|$ qui sera fini (compte tenu de l'hypothèse 3.3 sur les vecteurs m_a ci-dessous) dès que Θ est un ensemble compact.

Hypothèse 3.3. *Les vecteurs d'informations $\{m_a, a \in \mathcal{A}\}$ sont de norme bornée : il existe $c_m < \infty$ tel que pour tout $a \in \mathcal{A}$, $\|m_a\|_2 \leq c_m$.*

Hypothèse 3.4. *Soit $\epsilon_t = R_t - \mu(m'_{A_t}\theta_*)$. Pour tout t, les variables aléatoires $(\epsilon_t)_t$ sont indépendantes et centrées. De plus, il existe $R_{\max} > 0$ tel que, pour tout t, les récompenses aléatoires R_t sont positives et bornées par $R_{\max}$.*

Légère modification de l'algorithme

Avant de présenter des bornes de regret en suivant l'algorithme GLM-UCB, nous proposons une version modifiée de celui-ci qui permet de résoudre un détail technique dans les preuves ci-dessous. Afin de pouvoir obtenir des bornes de regret qui soient valables pour tout horizon n, il est nécessaire de s'assurer que le paramètre $\hat{\theta}_t$ appartient, pour tout instant t, à l'ensemble Θ, supposé compact, dans lequel on recherche les paramètres. A priori, rien ne garantit que la solution de l'équation (3.5) appartienne à Θ. Nous exposons donc ci-dessous une seconde version de l'algorithme GLM-UCB dans laquelle le paramètre estimé $\hat{\theta}_t$ est projeté dans l'ensemble Θ. Introduisons tout d'abord la fonction inversible

$$g_t : \theta \mapsto \sum_{k=0}^{t-1} \mu(m'_{A_k}\theta) m_{A_k} \ .$$

On remarque que l'équation (3.5) peut se réécrire

$$g_t(\theta) = \sum_{k=0}^{t-1} R_k m_{A_k} \ . \tag{3.11}$$

On définit $\widetilde{\theta}_t$ comme étant la projection de $\hat{\theta}_t$ sur l'espace Θ selon :

$$\widetilde{\theta}_t = \underset{\theta\in\Theta}{\text{argmin}} \left\| g_t(\theta) - g_t(\hat{\theta}_t) \right\|_{M_t^{-1}} . \tag{3.12}$$

En pratique les deux versions de l'algorithme donnent les mêmes résultats. De plus, l'ensemble Θ peut être choisi assez grand pour que $\hat{\theta}_t$ soit toujours inclu dans Θ. Cette deuxième version de l'algorithme ayant un intérêt principalement théorique, dans les expériences numériques présentées dans la suite, nous utilisons l'algorithme 3.1.

Algorithme 3.2 Algorithme GLM-UCB -version 2-

1: Jouer les actions $a_0, \ldots, a_{d-1}$
2: Réception des récompenses $R_0, \ldots, R_{d-1}$.
3: **Pour** $t > d$ **faire**
4: Calculer $\hat{\theta}_t$, solution de l'équation (3.5)
5: Calculer $\widetilde{\theta}_t$ défini par l'équation (3.12)
6: Jouer l'action $A_t = \text{argmax}_a \, \mu(m'_a \widetilde{\theta}_t) + \beta_t^a(\delta)$
7: Réception de la récompense R_t
8: **fin Pour**

Bornes de regret

Nous exposons maintenant une première borne supérieure du regret accumulé en suivant l'algorithme GLM-UCB avec

$$\rho(t) = \frac{4k_\mu \kappa^2 R_{\max}}{c_\mu} \sqrt{2d \log(t) \log(2\, t\ n/\delta)}\ , \tag{3.13}$$

où $\kappa = \sqrt{3 + 2\log(1 + 2c_m^2/\lambda_0)}$ et λ_0 désigne la plus petite valeur propre de la matrice $\sum_{i=0}^{d-1} m_{a_i} m'_{a_i}$. La variable n est l'horizon que l'on s'est fixé. Cette borne dépend du vrai paramètre θ_* à travers la différence entre l'espérance de la récompense reçue en jouant le bras optimal et celle correspondant au meilleur bras sous-optimal. Posons :

$$\Delta(\theta_*) = \min_{a, \mu(m'_a\theta_*) \neq \mu(m'_{a_*}\theta_*)} \mu(m'_{a_*}\theta_*) - \mu(m'_a\theta_*)\ .$$

Théorème 3.1 (Borne supérieure du regret dépendant de θ_*)**.** *Sous les hypothèses 3.1, 3.2, 3.3 et 3.4, et pour tout horizon n suffisamment grand, le regret de l'algorithme est contrôlé par*

$$\mathbb{P}\left(\mathit{Regret}_n \leq (d+1) R_{\max} + \frac{C\, d^2}{\Delta(\theta_*)} \log^2 [s\, n] \log \left\lceil \frac{2n^2}{\delta} \right\rceil \right) \geq 1 - \delta$$

avec $C = \frac{256 \kappa^4 R_{\max}^2 k_\mu^2}{c_\mu^2}$.

Le théorème suivant donne une borne supérieure du regret indépendante du vrai paramètre θ_*.

Théorème 3.2 (Borne supérieure du regret indépendante de θ_*)**.** *Sous les hypothèses 3.1, 3.2, 3.3 et 3.4, et pour tout horizon n suffisamment grand, le regret de l'algorithme vérifie*

$$\mathbb{P}\left(\mathit{Regret}_n \leq (d+1) R_{\max} + C d \log [s\, n] \sqrt{n \log \left\lceil \frac{2n^2}{\delta} \right\rceil} \right) \geq 1 - \delta$$

avec $C = \frac{16 R_{\max} k_\mu \kappa^2}{c_\mu}$.

Les deux bornes de regret ci-dessus ne dépendent pas du nombre de bras $|\mathcal{A}|$ mais uniquement de la dimension d de l'espace des paramètres. Un tel résultat n'avait, jusqu'à présent, été prouvé que pour des bandits linéaires. De plus, la dépendance de ces bornes, à la fois en l'horizon n et en la dimension de l'espace des paramètres d, est similaire à celles des bornes de regret construites dans le cas linéaire [Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2008]. La seule différence importante entre les bornes de regret dans le cas linéaire et dans le cas linéaire généralisé est l'ajout d'une constante multiplicative k_μ/c_μ dépendant de la fonction de lien μ.

La suite de cette section est dédiée à la démonstration des théorèmes 3.1 et 3.2. Ces démonstrations s'inspirent des preuves faites par [Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2008] pour des bandits linéaires.

Avant de démontrer les théorèmes 3.1 et 3.2, nous présentons deux résultats préliminaires. En particulier, nous démontrons une inégalité exponentielle sur des martingales vectorielles, inspirée du résultat A.8 présenté en annexe.

Résultats préliminaires

Lemme 3.1. *Soit $(\mathcal{F}_k)_{k\geq 0}$ une filtration, $(m_k)_{k\geq 0}$ un processus stochastique à valeur dans $\mathbb{R}^d$ adapté à $(\mathcal{F}_k)$, $(\eta_k)_{k\geq 1}$ une suite d'incréments de martingales à valeurs réelles adaptée à $(\mathcal{F}_k)$. Supposons qu'il existe $R > 0$ tel que pour tout $\gamma \geq 0$, $k \geq 1$,*

$$\mathbb{E}[\exp(\gamma\eta_k) \,|\, \mathcal{F}_{k-1}] \leq \exp\left(\frac{\gamma^2 R^2}{2}\right) \quad \text{p.s.} \tag{3.14}$$

Soit $\xi_t = \sum_{k=1}^{t} m_{k-1}\eta_k$ [1] et $M_t = \sum_{k=1}^{t} m_{k-1}m'_{k-1}$. Supposons que la plus petite valeur propre de la matrice M_d soit minorée par une constante λ_0 positive avec probabilité 1 et que $\|m_k\|_2 \leq c_m$ presque sûrement pour tout $k \geq 0$.

Soit

$$\kappa = \sqrt{3 + 2\log(1 + 2c_m^2/\lambda_0)}. \tag{3.15}$$

Alors, pour tout $x \in \mathbb{R}^d$, $0 < \delta \leq 1/e$, $t \geq \max(d, 2)$, avec probabilité au moins égale à $1 - \delta$,

$$|x'\xi_t| \leq \kappa R\sqrt{2\log t}\,\sqrt{\log(1/\delta)}\;\|x\|_{M_t}. \tag{3.16}$$

De plus, pour tout $0 < \delta < 1$, $t \geq \max(d, 2)$, avec probabilité au moins égale à $1 - \delta$,

$$\|\xi_t\|_{M_t^{-1}} \leq 2\kappa^2 R\sqrt{2\,d\log t}\,\sqrt{\log(t/\delta)}. \tag{3.17}$$

La preuve de (3.16) est basée sur une inégalité exponentielle de [De La Pena et al., 2004] et s'inspire du lemme B.4 de [Rusmevichientong and Tsitsiklis, 2008]. L'inégalité (3.17) découle de (3.16) en utilisant une borne de l'union et quelques résultats algébriques.

Démonstration. Afin de démontrer (3.16), nous utilisons le Corollaire 2.2 de [De La Pena et al., 2004] dont le résultat est le suivant. Pour toutes variables aléatoires A et $B \geq 0$ telles que

$$\mathbb{E}\left[\exp\left\{\gamma A - \frac{\gamma^2}{2}B^2\right\}\right] \leq 1 \quad \text{pour tout } \gamma \in \mathbb{R}\,. \tag{3.18}$$

Alors, pour tout $c \geq \sqrt{2}$, et tout $y > 0$,

$$\mathbb{P}\left(|A| \geq c\sqrt{(B^2 + y)\left(1 + \frac{1}{2}\log\left(\frac{B^2}{y} + 1\right)\right)}\right) \leq \exp\left\{-\frac{c^2}{2}\right\}\,. \tag{3.19}$$

Nous appliquons ce résultat aux variables aléatoires $A = x'\xi_t/R$ et $B = \|x\|_{M_t}$, où $x \in \mathbb{R}^d$ est un vecteur fixé. Vérifions en premier lieu que les variables aléatoires A et B satisfont (3.18). Soit $\gamma \in \mathbb{R}$. Étudions la quantité $\gamma A - (\gamma B)^2/2$. On a

$$\gamma A - (\gamma B)^2/2 = \frac{\gamma\, x'\xi_t}{R} - \frac{\gamma^2\, x' M_t x}{2} = \sum_{k=1}^{t} D_k\,,$$

1. Notons que ce lemme utilise une indexation différentes de celle utilisée dans le reste du chapitre.

où

$$D_k = \frac{\gamma}{R}\, x' m_{k-1}\eta_k - \frac{\gamma^2}{2} x'\, m_{k-1} m_{k-1}'\, x = \frac{\gamma}{R}\, x' m_{k-1}\eta_k - \frac{\gamma^2}{2}(x' m_{k-1})^2 \,.$$

Observons maintenant que, d'après (3.14), $\mathbb{E}\left[\exp(D_k)\,|\,\mathcal{F}_{k-1}\right] \leq 1$. Soit $P_k = \exp(D_k)$. En remarquant que P_k est $\mathcal{F}_k$-adaptée, on a

$$\begin{aligned}
\mathbb{E}\left[\exp(\gamma A - \gamma B^2/2)\right] &= \mathbb{E}\left[P_1 \cdots P_{t-1} P_t\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[P_1 \cdots P_{t-1} P_t \,|\, \mathcal{F}_{t-1}\right]\right] = \mathbb{E}\left[P_1 \cdots P_{t-1}\, \mathbb{E}\left[P_t \,|\, \mathcal{F}_{t-1}\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[P_1 \cdots P_{t-1} \,|\, \mathcal{F}_{t-2}\right]\right] = \mathbb{E}\left[P_1 \cdots P_{t-2} \mathbb{E}\left[P_{t-1} \,|\, \mathcal{F}_{t-2}\right]\right] \\
&\vdots \\
&\leq \mathbb{E}\left[\mathbb{E}\left[P_1 \,|\, \mathcal{F}_0\right]\right] \leq 1
\end{aligned}$$

ce qui termine la vérification de (3.18). Choisissons maintenant $y = \lambda_0 \|x\|_2^2$ et utilisons (3.19) pour montrer que, pour tout $0 < \delta \leq 1/e$, $t \geq 1$, avec probabilité $1 - \delta$, on a

$$|x'\xi_t| \leq R\sqrt{\left(\|x\|_{M_t}^2 + \lambda_0\|x\|_2^2\right)\left(1 + \frac{1}{2}\log\left(1 + \frac{\|x\|_{M_t}^2}{\lambda_0\|x\|_2^2}\right)\right)}\sqrt{2\log\left(\frac{1}{\delta}\right)}\,. \tag{3.20}$$

En remarquant que pour $t \geq \max(d, 2)$, $\lambda_0\|x\|_2^2 \leq \|x\|_{M_t}^2 \leq t\|x\|_2^2 c_m^2$, on a

$$\|x\|_{M_t}^2 + \lambda_0\|x\|_2^2 \leq 2\|x\|_{M_t}^2$$

et

$$1 + \frac{1}{2}\log\left(1 + \frac{\|x\|_{M_t}^2}{\lambda_0\|x\|_2^2}\right) \leq 1 + \frac{1}{2}\log\left(1 + \frac{tc_m^2}{\lambda_0}\right) \leq \kappa^2 \log(t)/2\,,$$

si

$$\kappa \geq \sqrt{\frac{2 + \log(1 + 2c_m^2/\lambda_0)}{\log 2}}.$$

En majorant $2/\log 2$ par 3 et $1/\log 2$ par 2, nous obtenons que la valeur de κ définie dans l'énoncé, vérifie la condition ci-dessus.

Donc, quand (3.20) est vérifiée, on a

$$|x'\xi_t| \leq \kappa R\|x\|_{M_t}\sqrt{\log(t)}\sqrt{2\log\left(\frac{1}{\delta}\right)} \tag{3.21}$$

ce qui est exactement (3.16).

Démontrons maintenant (3.17). La preuve est basée sur un simple argument géométrique. Soit $\mathbb{S}^d = \left\{x \in \mathbb{R}^d \,:\, \|x\|_2 = 1\right\}$ la sphère unité de $\mathbb{R}^d$. Un ensemble $\mathcal{N}$ est dit être un ϵ-réseau d'un ensemble $\mathcal{U} \subset \mathbb{R}^d$ si $\mathcal{N} \subset \mathcal{U}$ et pour tout $x, y \in \mathcal{N}$, $\|x - y\|_2 \geq \epsilon$. Un ϵ-réseau est dit maximal si ce n'est pas un sous-ensemble d'un autre ϵ-réseau (du même ensemble U). Un ϵ-réseau maximal $\mathcal{N}$ crée un recouvrement U. Si $\mathcal{N}$ est un tel ϵ-réseau maximal alors pour tout $x \in U$ il existe un élément $y \in \mathcal{N}$ tel que $\|x - y\| \leq \epsilon$. Le nombre d'élément $k = |\mathcal{N}|$ d'un tel ensemble est inférieur ou égal à $(2/\epsilon + 1)^d$. Pour tout $1 \leq i \leq d$, soit $\mathbf{e}_i$ le i-ème vecteur unité. Posons

$$\epsilon = \frac{1}{2}\sqrt{\frac{\lambda_0}{tc_m^2}}$$

et choisissons un ϵ-réseau maximal $\mathcal{N}$ de $\mathbb{S}^d$ tel que $\{\mathbf{e}_1, \dots \mathbf{e}_d\} \subset \mathcal{N}$. D'après (3.16) et une borne de l'union, avec probabilité $1-\delta$,

$$\max_{x \in \mathcal{N}} \frac{|x'\xi_t|}{\|x\|_{M_t^{-1}}} \leq \kappa R \sqrt{2\log(t)}\, \sqrt{\log\left(k/\delta\right)}\ .$$

D'après le choix de la valeur de ϵ, et des applications successives de l'inégalité triangulaire ainsi que la majoration de k, avec probabilité $1-\delta$, on a

$$\begin{aligned}
\sup_{x \neq 0} \frac{|x'\xi_t|}{\|x\|_{M_t^{-1}}} &\leq 2\kappa R \sqrt{2\log(t)}\, \sqrt{\log(k/\delta)} \\
&= 2\kappa R \sqrt{2\log(t)}\, \sqrt{d\log\left(2/\epsilon + 1\right) + \log\left(1/\delta\right)} \\
&\leq 2\kappa R \sqrt{2\log(t)}\, \sqrt{d\log\left(\sqrt{16 t c_m^2/\lambda_0} + 1\right) + \log\left(1/\delta\right)} \\
&\leq 2\kappa R \sqrt{2\log(t)}\, \sqrt{d\left(1 + \kappa^2/2\ \log(t)\right) + \log\left(1/\delta\right)} \\
&\leq 2\kappa^2 R \sqrt{2\log(t)}\, \sqrt{d\log(t) + \log\left(1/\delta\right)}\ .
\end{aligned}$$

De plus, on remarque que

$$\|\xi_t\|_{M_t^{-1}} = \sup_{x \neq 0} \frac{|x'\xi_t|}{\|x\|_{M_t^{-1}}}$$

ce qui termine la preuve.

□

Remarque 2. Notons que s'il existe α_k, une variable aléatoire $\mathcal{F}_{k-1}$-mesurable, telle que $\eta_k \in [\alpha_k - R, \alpha_k + R]$ presque sûrement alors, en utilisant l'inégalité d'Hoeffding (voir A.1 en annexe), on a, pour tout $\gamma \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left\{\gamma\eta_k\right\} | \mathcal{F}_{k-1}\right] \leq \exp\left\{\gamma \mathbb{E}\left[\eta_k \,|\, \mathcal{F}_{k-1}\right]\right\} \exp\left\{\frac{4R^2\gamma^2}{8}\right\} = \exp\left\{\frac{\gamma^2\, R^2}{2}\right\}\ ,$$

ce qui montre que (η_k) satisfait la condition (3.14). En particulier, ceci est vrai si $|\eta_k| \leq R$ presque sûrement.

Nous démontrons maintenant des majoration de l'erreur de prédiction de la moyenne des récompenses. Nous commençons avec le résultat suivant :

Proposition 3.1. *Soient δ et t quelconques tels que $0 < \delta < 1$, $1 + \max(d, 2) \leq t \leq n$. Soit $\tilde{A}_t$ une variable aléatoire à valeur dans $\mathcal{A}$. Soit*

$$\beta_t^a(\delta) = \frac{4\, k_\mu \kappa^2 R_{\max}}{c_\mu} \|m_a\|_{M_t^{-1}}\, \sqrt{2\, d \log t}\, \sqrt{\log(t/\delta)}\ , \tag{3.22}$$

où κ est défini équation (3.15)*. Alors, avec probabilité au moins égale à $1-\delta$, on a que*

$$\left|\mu(m'_{\tilde{A}_t}\theta_*) - \mu(m'_{\tilde{A}_t}\tilde{\theta}_t)\right| \leq \beta_t^{\tilde{A}_t}(\delta)\ .$$

Démonstration. Soit t tel que $d+1 \leq t \leq n$ et une action $a \in \mathcal{A}$. Commençons par majorer $\left|\mu(m'_a\theta_*) - \mu(m'_a\tilde{\theta}_t)\right|$. Puisque μ est Lipschitzienne, on a

$$|\mu(m'_a\theta_*) - \mu(m'_a\tilde{\theta}_t)| \leq k_\mu |m'_a(\theta_* - \tilde{\theta}_t)|\ .$$

D'après l'hypothèse 3.2, ∇g_t est continue, [2] donc, d'après le Théorème Fondamental de Calcul,

$$g_t(\theta_*) - g_t(\tilde{\theta}_t) = G_t(\theta_* - \tilde{\theta}_t) \ ,$$

où

$$G_t = \int_0^1 \nabla g_t \left(s\theta_* + (1-s)\tilde{\theta}_t\right) ds \ .$$

Pour tout $\theta \in \Theta$, $\nabla g_t(\theta) = \sum_{k=1}^{t-1} m_{A_k} m'_{A_k} \dot{\mu}(m'_{A_k}\theta)$. Pour toute matrice A et B, on note $A \succeq B$ si la matrice $A - B$ est semi-définie positive et $A \succ B$ si la matrice $A - B$ est définie positive. D'après l'hypothèse 3.2, on a $G_t \succeq c_\mu M_t \succeq c_\mu M_d \succ 0$, où l'on utilise le fait que les d premières actions sont telles que $M_d \succeq \lambda_0 I \succ 0$. Donc, G_t est définie positive et donc c'est une matrice non-singulière. Ainsi,

$$\left|\mu(m'_a\theta_*) - \mu(m'_a\tilde{\theta}_t)\right| \leq k_\mu \left|m'_a G_t^{-1}(g_t(\theta_*) - g_t(\tilde{\theta}_t))\right| \ .$$

Puisque G_t^{-1} est également définie, on a

$$\left|\mu(m'_a\theta_*) - \mu(m'_a\tilde{\theta}_t)\right| \leq k_\mu \|m_a\|_{G_t^{-1}} \left\|g_t(\theta_*) - g_t(\tilde{\theta}_t)\right\|_{G_t^{-1}} \ . \tag{3.23}$$

Puisque $G_t \succeq c_\mu M_t$ implique que $G_t^{-1} \preceq c_\mu^{-1} M_t^{-1}$, $\|x\|_{G_t^{-1}} \leq \frac{1}{\sqrt{c_\mu}} \|x\|_{M_t^{-1}}$ est satisfaite pour tout $x \in \mathbb{R}^d$. Donc,

$$\left|\mu(m'_a\theta_*) - \mu(m'_a\tilde{\theta}_t)\right| \leq \frac{k_\mu}{c_\mu} \|m_a\|_{M_t^{-1}} \left\|g_t(\theta_*) - g_t(\tilde{\theta}_t)\right\|_{M_t^{-1}} \ .$$

On a

$$\begin{aligned}\left\|g_t(\theta_*) - g_t(\tilde{\theta}_t)\right\|_{M_t^{-1}} &\leq \left\|g_t(\theta_*) - g_t(\hat{\theta}_t)\right\|_{M_t^{-1}} + \left\|g_t(\hat{\theta}_t) - g_t(\tilde{\theta}_t)\right\|_{M_t^{-1}} \\ &\leq 2\left\|g_t(\theta_*) - g_t(\hat{\theta}_t)\right\|_{M_t^{-1}} \ ,\end{aligned}$$

où la première inégalité découle de l'inégalité triangulaire et la deuxième de l'hypothèse que $\theta_* \in \Theta$ et du fait que $\left\|g_t(\hat{\theta}_t) - g_t(\tilde{\theta}_t)\right\|_{M_t^{-1}} \leq \left\|g_t(\theta_*) - g_t(\hat{\theta}_t)\right\|_{M_t^{-1}}$.

D'après la définition de $\hat{\theta}_t$, et en utilisant le fait que $\epsilon_k = R_k - \mu(m'_{A_k}\theta_*)$, on a $\xi_t \stackrel{\text{def}}{=} g_t(\hat{\theta}_t) - g_t(\theta_*) = \sum_{k=1}^{t-1} m_{A_k}\epsilon_k$. Donc,

$$\left|\mu(m'_a\theta_*) - \mu(m'_a\tilde{\theta}_t)\right| \leq \frac{2\,k_\mu}{c_\mu} \|m_a\|_{M_t^{-1}} \|\xi_t\|_{M_t^{-1}} \ .$$

Etant donné que cette inégalité est satisfaite simultanément pour toute action $a \in \mathcal{A}$, elle l'est aussi lorsque a est remplacé par *n'importe quelle* variable aléatoire $\tilde{A}_t$ à valeur dans $\mathcal{A}$:

$$\left|\mu(m'_{\tilde{A}_t}\theta_*) - \mu(m'_{\tilde{A}_t}\tilde{\theta}_t)\right| \leq \frac{2\,k_\mu}{c_\mu} \left\|m_{\tilde{A}_t}\right\|_{M_t^{-1}} \|\xi_t\|_{M_t^{-1}} \ . \tag{3.24}$$

Utilisons maintenant le lemme 3.1 pour majorer $\|\xi_t\|_{M_t^{-1}}$. Soit $m_k = m_{A_{k+1}}$ $(k = 0, 1, \ldots)$, $\eta_k = \epsilon_k$ $(k = 1, 2, \ldots)$, $\mathcal{F}_k = \sigma(m_s, \eta_s; s \leq k)$. D'après l'hypothèse 3.4, on a

$$\mathbb{E}\left[\eta_k | \mathcal{F}_{k-1}\right] = \mathbb{E}\left[\eta_k | m_{k-1}, \eta_{k-1}, \ldots, m_1, \eta_1, m_0\right] = \mathbb{E}\left[\epsilon_k | m_{A_k}, \epsilon_{k-1}, \ldots, m_{A_2}, \epsilon_1, m_{A_1}\right] = 0 \ .$$

2. Pour tout $x \in \mathbb{R}^d$, $\nabla g_t(x)$ désigne la matrice Jacobienne de g_t au point x.

Puisque, $|\epsilon_k| \leq R_{\max}$, nous pouvons choisir $R = R_{\max}$ d'après la remarque 2. De plus, d'après l'hypothèse 3.3,

$$\|m_k\|_2 = \|m_{A_{k+1}}\|_2 \leq \max_{a \in \mathcal{A}} \|m_a\|_2 \leq c_m \ ,$$

et, étant donné le choix des d premières actions,

$$\sum_{k=1}^{d} m_{k-1} m'_{k-1} = \sum_{k=1}^{d} m_{A_k} m'_{A_k} \succeq \lambda_0 I \ .$$

Donc, toutes les hypothèses du lemme sont satisfaites et l'on peut conclure que, pour tout $0 < \delta < 1$, $t \geq 1 + \max(d, 2)$, avec probabilité au moins égale à $1 - \delta$,

$$\|\xi_t\|_{M_t^{-1}} \leq 2\kappa^2 R_{\max} \sqrt{2\, d \log t}\, \sqrt{\log(t/\delta)}, \tag{3.25}$$

où κ est défini par (3.15).

En enchaînant (3.24) et (3.25), on a que, dès que l'évènement (3.25) est vérifié, alors

$$\left|\mu(m'_{\tilde{A}_t}\theta_*) - \mu(m'_{\tilde{A}_t}\tilde{\theta}_t)\right| \leq \frac{4\, k_\mu \kappa^2 R_{\max}}{c_\mu} \left\|m_{\tilde{A}_t}\right\|_{M_t^{-1}} \sqrt{2\, d \log t}\, \sqrt{\log(t/\delta)} \ ,$$

ce qui termine la preuve. □

La proposition 3.1 permet de démontrer la majoration suivante sur le regret moyen instantané :

Proposition 3.2. *Pour tout $\delta \in (0,1)$, simultanément pour tout $t \in \{1 + \max(d,2), \ldots, n\}$,*

$$\mu(m'_{a_*}\theta_*) - \mu(m'_{A_t}\theta_*) \leq 2\, \beta_t^{A_t}\left(\tfrac{\delta}{2n}\right) .$$

avec probabilité $1 - \delta$.

Démonstration. Soit $t \in \{1+\max(d,2), \ldots, n\}$ et soit δ fixé comme dans l'énoncé. Considérons la décomposition

$$\begin{aligned}\mu(m'_{a_*}\theta_*) - \mu(m'_{A_t}\theta_*) = &\left(\mu(m'_{a_*}\theta_*) - \mu(m_{a_*}\tilde{\theta}_t)\right) \\ &+ \left(\mu(m_{a_*}\tilde{\theta}_t) - \mu(m_{A_t}\tilde{\theta}_t)\right) + \left(\mu(m_{A_t}\tilde{\theta}_t) - \mu(m'_{A_t}\theta_*)\right) .\end{aligned}$$

D'après la proposition 3.1, avec probabilité $1 - \delta/(2n)$, on a

$$\mu(m'_{a_*}\theta_*) - \mu(m'_{a_*}\tilde{\theta}_t) \leq \beta_t^{a_*}(\delta/(2n)) \ .$$

De plus, avec probabilité $1 - \delta/(2n)$, on a

$$\mu(m'_{A_t}\theta_*) - \mu(m'_{A_t}\tilde{\theta}_t) \leq \beta_t^{A_t}(\delta/(2n)) \ .$$

Et, par définition de A_t,

$$\begin{aligned}\mu(m_{a_*}\tilde{\theta}_t) - \mu(m_{A_t}\tilde{\theta}_t) &= \mu(m_{a_*}\tilde{\theta}_t) + \beta_t^{a_*}(\delta/(2n)) - \mu(m_{A_t}\tilde{\theta}_t) - \beta_t^{a_*}(\delta/(2n)) \\ &\leq \mu(m_{A_t}\tilde{\theta}_t) + \beta_t^{A_t}(\delta/(2n)) - \mu(m_{A_t}\tilde{\theta}_t) - \beta_t^{a_*}(\delta/(2n)) \\ &= \beta_t^{A_t}(\delta/(2n)) - \beta_t^{a_*}(\delta/(2n)).\end{aligned}$$

En enchaînant les inégalités et en utilisant une borne de l'union, on obtient le résultat final. □

D'après la proposition précédente, le comportement du regret immédiat à l'instant t est majoré par $2\beta_t^{A_t}(\delta/2n) = 2\rho(t)\|m_{A_t}\|_{M_t^{-1}} \leq 2\rho(n)\|m_{A_t}\|_{M_t^{-1}}$ (voir. (3.13) pour la définition de $\rho(t)$.) Donc, avec $t_0 = 1 + \max(d, 2)$, on peut majorer le regret cumulé jusqu'à l'instant n, avec probabilité $1 - \delta$, par

$$\text{Regret}_n \leq (t_0 - 1)R_{\max} + \sum_{t=t_0}^{n} \min\left\{\mu(m'_{a_*}\theta_*) - \mu(m'_{A_t}\theta_*), R_{\max}\right\} \tag{3.26}$$

$$\leq (t_0 - 1)R_{\max} + 2\,\rho(n) \sum_{t=t_0}^{n} \min\left\{\|m_{A_t}\|_{M_t^{-1}}, 1\right\} , \tag{3.27}$$

où la dernière inégalité découle du fait que $R_{\max} \leq 2\rho(n)$ par définition de $\rho(n)$. Remarquons que $\|m_{A_t}\|_{M_t^{-1}}$ devient petit lorsque t grandit. Ceci motive la majoration de la somme des $\|m_{A_t}\|^2_{M_t^{-1}}$. Pour des raisons techniques qui deviendront claires par la suite, nous majorons $\sum_{t=d}^{n} \min\left\{\|m_{A_t}\|^2_{M_t^{-1}}, 1\right\}$.

Proposition 3.3. *Soit $t_0 \geq d + 1$. Alors,*

$$\textstyle\sum_{t=t_0}^{n} \min\left\{\|m_{A_t}\|^2_{M_t^{-1}}, 1\right\} \leq 2\,d\,\log\left(\frac{c_m^2 n}{\lambda_0}\right) \quad \text{p.s.} .$$

Démonstration. La preuve suit les pas de la preuve du lemme 9 de [Dani et al., 2008]. Par définition de M_{t+1}, on a

$$\begin{aligned}\det(M_{t+1}) &= \det\left(M_t + m_{A_t}m'_{A_t}\right) = \det(M_t)\det\left(I + M_t^{-1/2}m_{A_t}(M_t^{-1/2}m_{A_t})'\right) \\ &= \det(M_t)\left(1 + \|m_{A_t}\|^2_{M_t^{-1}}\right) = \det(M_{t_0})\prod_{k=t_0}^{t}\left(1 + \|m_{A_k}\|^2_{M_k^{-1}}\right) ,\end{aligned}$$

où la dernière ligne découle du fait que $1 + \|m_{A_t}\|^2_{M_t^{-1}}$ est une valeur propre de la matrice $I + M_t^{-1/2}m_{A_t}(M_t^{-1/2}m_{A_t})'$ et que toutes les autres valeurs propres sont égales à 1. Donc, en utilisant le fait que $x \leq 2\log(1 + x)$ ce qui est vrai pour tout $0 \leq x \leq 1$, on a

$$\begin{aligned}\sum_{t=t_0}^{n} \min\left\{\|m_{A_t}\|^2_{M_t^{-1}}, 1\right\} &\leq 2\sum_{t=t_0}^{n} \log\left(1 + \|m_{A_t}\|^2_{M_t^{-1}}\right) \\ &= 2\log\prod_{t=t_0}^{n}\left(1 + \|m_{A_t}\|^2_{M_t^{-1}}\right) \\ &= 2\log\left(\frac{\det(M_{n+1})}{\det(M_{t_0})}\right) .\end{aligned}$$

Notons que la trace de M_{t+1} est majorée par $t\,c_m^2$. Donc, puisque la trace de la matrice définie positive M_{t+1} est égale à la somme de ses valeurs propres et $\det(M_{t+1})$ est le produit de ses valeurs propres, on a $\det(M_{t+1}) \leq (tc_m^2)^d$. De plus, $\det(M_{t_0}) \geq \lambda_0^d$ lorsque $t_0 \geq d + 1$. Donc,

$$\sum_{t=t_0}^{n} \min\left\{\|m_{A_t}\|^2_{M_t^{-1}}, 1\right\} \leq 2\,d\log\left(\frac{c_m^2 n}{\lambda_0}\right) .$$

□

Preuve des théorèmes

Preuve du théorème 3.1. Nous partons de (3.26), où $t_0 = 1 + \max(d, 2)$. D'après la définition de $\Delta(\theta_*)$ quand A_t est une action sous-optimale, $\mu(m'_{a_*}\theta_*) - \mu(m'_{A_t}\theta_*) \geq \Delta(\theta_*)$, alors que dans l'autre cas $\mu(m'_{a_*}\theta_*) - \mu(m'_{A_t}\theta_*) = 0$. Dans les deux cas, on peut écrire

$$\mu(m'_{a_*}\theta_*) - \mu(m'_{A_t}\theta_*) \leq \frac{(\mu(m'_{a_*}\theta_*) - \mu(m'_{A_t}\theta_*))^2}{\Delta(\theta_*)} \ .$$

D'après la proposition 3.2, avec probabilité $1 - \delta$, simultanément pour tout $t \in \{t_0, \dots, n\}$,

$$\mu(m'_{a_*}\theta_*) - \mu(m'_{A_t}\theta_*) \leq 2\beta_t^{A_t}(\delta/(2n)) = 2\rho(t)\,\|m_{A_t}\|_{M_t^{-1}} \ .$$

Donc, lorsque toutes ces inégalités sont satisfaites, on a

$$\begin{aligned}\sum_{t=t_0}^{n} \min\left\{\mu(m'_{a_*}\theta_*) - \mu(m'_{A_t}\theta_*), R_{\max}\right\} &\leq \sum_{t=t_0}^{n} \min\left\{4\frac{\rho(t)^2}{\Delta(\theta_*)}\|m_{A_t}\|^2_{M_t^{-1}}, R_{\max}\right\} \\ &\leq 4\frac{\rho(n)^2}{\Delta(\theta_*)} \sum_{t=t_0}^{n} \min\left\{\|m_{A_t}\|^2_{M_t^{-1}}, 1\right\}\end{aligned}$$

où la dernière inégalité découle du fait que $\Delta(\theta_*) \leq R_{\max} \leq 4\rho(n)^2/R_{\max}$ et que $\rho(.)$ est une fonction croissante. En utilisant de plus la majoration de la proposition 3.3, on obtient

$$\sum_{t=t_0}^{n} \min\left\{\mu(m'_{a_*}\theta_*) - \mu(m'_{A_t}\theta_*), R_{\max}\right\} \leq 8\,d\,\frac{\rho(n)^2}{\Delta(\theta_*)}\,\log\left(\frac{c_m^2 n}{\lambda_0}\right) \ .$$

En ajoutant la définition de $\rho(n)$, on a, avec probabilité $1 - \delta$, que

$$\begin{aligned}\text{Regret}_n &\leq (t_0 - 1)R_{\max} + \sum_{t=t_0}^{n} \min\left\{\mu(m'_{a_*}\theta_*) - \mu(m'_{A_t}\theta_*), R_{\max}\right\} \\ &\leq (t_0 - 1)R_{\max} + \frac{256\,d^2\,\kappa^4 R_{\max}^2 k_\mu^2}{c_\mu^2 \Delta(\theta_*)}\,\log(n)\log(2n^2/\delta)\log\left(\frac{c_m^2 n}{\lambda_0}\right) \ .\end{aligned}$$

□

Preuve du thèorème 3.2. Soit $t_0 = 1 + \max(d, 2)$. D'après la proposition 3.2, (3.27) est satisfaite avec probabilité $1 - \delta$, donc il reste à majorer

$$\sum_{t=t_0}^{n} \min\left\{\|m_{A_t}\|_{M_t^{-1}}, 1\right\}.$$

En utilisant l'inégalité de Cauchy-Schwarz et la proposition 3.3, on a

$$\begin{aligned}\sum_{t=t_0}^{n} \min\left\{\|m_{A_t}\|_{M_t^{-1}}, 1\right\} &\leq \sqrt{n}\sqrt{\sum_{t=t_0}^{n} \min\left\{\|m_{A_t}\|^2_{M_t^{-1}}, 1\right\}} \\ &\leq \sqrt{n}\,\sqrt{2d\log(c_m^2 n/\lambda_0)} \ .\end{aligned}$$

En utilisant (3.27) ainsi que la définition de $\rho(\cdot)$ on obtient

$$\begin{aligned}\text{Regret}_n &\leq (t_0 - 1)R_{\max} + 2\,\rho(n)\,\sqrt{2\,d\,n\log(c_m^2 n/\lambda_0)} \\ &= (t_0 - 1)R_{\max} + 16\,d\,\frac{k_\mu \kappa^2 R_{\max}}{c_\mu}\,\sqrt{n\,\log(n)\,\log(c_m^2 n/\lambda_0)\log(2n^2/\delta)} \\ &\leq (d+1)R_{\max} + 16\,d\,\frac{k_\mu \kappa^2 R_{\max}}{c_\mu}\,\log(s\,n)\,\sqrt{n\,\log(2n^2/\delta)},\end{aligned}$$

où $s = \max\left(\frac{c_m^2}{\lambda_0}, 1\right)$, ce qui finit donc la preuve. □

3.5.2 Borne de confiance asymptotique

Des expériences préliminaires faites en utilisant les bornes de confiance considérées dans les théorèmes 3.1 et 3.2, c'est-à-dire en utilisant la fonction ρ définie par l'équation (3.13), produisent de mauvais résultats, excepté pour des horizons très grands. Dans le cas où μ est la fonction identité, les algorithmes proposés par [Auer, 2002], [Dani et al., 2008] et [Rusmevichientong and Tsitsiklis, 2008] ont un comportement similaire. En effet, certaines approximations mathématiques inévitables conduisent à des bornes de confiance très pessimistes visant à garantir que la vraie valeur du paramètre appartient avec grande probabilité aux régions de confiance et ainsi garantir des bornes de regret assez faibles. Bien qu'aucune preuve ne soit disponible pour le moment, une comparaison avec l'algorithme UCB ainsi que les arguments asymptotiques présentés ci-dessous suggèrent un choix de paramètre $\rho(t)$ du bonus d'exploration significativement plus petit. C'est ce dernier qui sera utilisé dans les expériences numériques.

Considérons le modèle linéaire généralisé canonique associé à la fonction de lien inverse μ et supposons que les vecteurs de covariables X sont générés de manière indépendante sous une distribution fixée. Cette situation décrit, par exemple, le cas où les bras seraient tirés en suivant une politique aléatoire fixée. Des arguments statistiques classiques montrent que la matrice d'information de Fisher correspondant à ce modèle est donnée par $I = \mathbb{E}[\dot{\mu}(X'\theta_*)XX']$ et que l'estimateur de maximum de vraisemblance $\hat{\theta}_t$ est tel que $t^{1/2}(\hat{\theta}_t - \theta_*) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1})$, où $\xrightarrow{\mathcal{L}}$ désigne la convergence en loi. De plus, $t^{-1}M_t \xrightarrow{\text{p.s.}} \Sigma$ où $\Sigma = \mathbb{E}[XX']$. Donc, en utilisant la delta-méthode et le lemme de Slutsky [Casella and Berger, 1990; Billingsley, 1979]

$$\|m_a\|_{M_t^{-1}}^{-1}(\mu(m_a'\hat{\theta}_t) - \mu(m_a'\theta_*)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \dot{\mu}(m_a'\theta_*)^2 \|m_a'\|_{\Sigma^{-1}}^{-2} \|m_a'\|_{I^{-1}}^2\right) .$$

La variance du terme de droite est plus petite que k_μ^2/c_μ puisque $I \succeq c_\mu \Sigma$. Donc, pour toute distribution d'échantillonnage telle que I et Σ sont des matrices définies positives, pour tout t suffisamment grand et tout δ suffisamment petit,

$$\mathbb{P}\left(\|m_a\|_{M_t^{-1}}^{-1}(\mu(m_a'\hat{\theta}_t) - \mu(m_a'\theta_*)) > \sqrt{2\frac{k_\mu^2}{c_\mu}\log(1/\delta)}\right)$$

est asymptotiquement borné par δ.

Une différence notable avec le cas linéaire réside dans le fait que la matrice d'information de Fisher dépend également de θ_* et pas uniquement de la distribution des covariables. Néanmoins, nous recommandons d'utiliser la norme $\|m_a\|_{M_t^{-1}}$ dans la construction des bornes de confiance car, comme on l'a vu ci-dessus et dans les Théorèmes 3.1–3.2, cette norme fournit une mesure robuste de l'incertitude pour des fonctions μ générales. En se fondant sur ces arguments asymptotiques, nous proposons d'utiliser en pratique des bornes de confiance augmentées d'un facteur $\sqrt{k_\mu^2/c_\mu}$ par rapport à celles utilisées dans l'algorithme UCB. Cela conduit à un algorithme beaucoup moins prudent que celui analysé dans la section précédente.

3.6 Expériences

Dans cette section, nous présentons différents résultats numériques illustrant les performances de l'algorithme GLM-UCB. Nous nous plaçons d'abord dans le cadre favorable à notre algorithme que sont des environnements simulés où la distribution des récompenses reçues appartient à la famille exponentielle canonique. Un tel environnement permet de comprendre plus précisément le comportement de l'algorithme. Ensuite, pour tester la robustesse de l'algorithme, nous nous confrontons à des applications réelles. A notre connaissance, il n'existe

pas de système de référence disponible pour tester des méthodes de bandits paramétriques sur des données réelles. Nous avons donc mis en oeuvre deux expériences à partir de données réelles.

Dans chacun des cas, l'algorithme GLM-UCB sera comparé à l'algorithme UCB, connu pour équilibrer de manière optimale exploration et exploitation dans des problèmes de bandit. L'algorithme UCB n'étant pas dédié aux bandits paramétriques, il ne prend pas en compte l'information associée à chacun des bras. Il semblerait donc normal que, lorsque le modèle est réellement paramétrique, le regret soit plus faible sous l'algorithme GLM-UCB que sous l'algorithme UCB. En revanche, si le modèle paramétrique que l'on a conjecturé décrit trop mal les récompenses reçues, alors l'algorithme UCB est vraisemblablement plus performant que l'algorithme GLM-UCB, au moins sur le long terme. En plus de l'algorithme UCB, nous comparons notre algorithme à un algorithme ϵ-glouton connaissant les vecteurs de caractéristique de chaque bras et jouant, avec probabilté $1-\epsilon$, le meilleur bras par rapport au paramètre estimé

$$A_t = \operatorname*{argmax}_a \mu(m_a' \hat{\theta}_t)$$

et avec probabilité ϵ un bras tiré au hasard.

3.6.1 Données simulées

Pour illustrer le comportement de l'algorithme, nous nous plaçons dans un environnement simulé à 2 dimensions dans lequel les récompenses sont des variables de Bernoulli de paramètre $\mu(m_a'\theta_*)$ où $\theta^* = (-1, 1)'$. Les informations caractérisant chaque bras sont générées de manière à être équiréparties et sont telles que, pour tout a, $\|m_a\|_2 = 1$. La fonction de lien étant la fonction logistique, on a $k_\mu = 1/4$ et, $c_\mu \simeq 0.1$; on prend alors $\rho(t) = 3\log(t)$. Pour estimer le paramètre $\hat{\theta}_t$, nous utilisons quelques itérations de l'algorithme de Newton [Boyd and Vandenberghe, 2004].

Nous appliquons les algorithmes GLM-UCB et UCB sur 20 trajectoires indépendantes pour un horizon $n = 1000$ et différents nombres $|\mathcal{A}|$ de bras. Le regret moyen reçu en fonction du nombre de bras est affiché sur la figure 3.1. Comme prévu, on observe que, contrairement à ce qu'il se passe pour l'algorithme UCB, augmenter le nombre de bras ne dégrade pas la performance de l'algorithme GLM-UCB.

De plus, nous représentons sur la figure 3.2 le nombre de fois où chaque bras est joué en suivant chacun des deux algorithmes. Les bras sont classés par ordre croissant de récompense moyenne. Contrairement à l'algorithme UCB qui nécessite de jouer plusieurs fois tous les bras -les bras ayant une récompense moyenne grande étant joués plus souvent que les autres-, on remarque que sous l'algorithme GLM-UCB peu de bras distincts sont tirés. En effet, l'algorithme utilise les vecteurs de caractéristiques pour estimer la moyenne des récompenses associées aux bras non joués.

Nous profitons de cet environnement simulé pour analyser le comportement de l'estimateur du maximum de quasi-vraisemblance en suivant l'algorithme GLM-UCB. Pour cela, nous considérons un environnement à 3 dimensions dans lequel 30 vecteurs de caractéristiques de norme 1 sont générés de manière à être équirépartis dans un plan contenant le paramètre $\theta^* = (1, -1, 0)'$. On ajoute un vecteur de caractéristique orthogonal à ce plan. Notons que, dans un tel environnement, plusieurs bras sont très proches de l'optimal. Représentons, sur la figure 3.3, l'évolution de l'estimateur $\hat{\theta}_t$ pour 10 trajectoires indépendantes. On remarque que, au bout d'un horizon $n = 20000$, les deux premières composantes du paramètre sont estimées de manière plus précise que la troisième composante.

On représente sur la figure 3.4 le nombre de fois où, en suivant l'algorithme GLM-UCB, l'agent joue respectivement un bras très largement sous-optimal dont le vecteur de caractéristique est dans le même plan que le paramètre θ_*, un bras proche de l'optimal (dont le

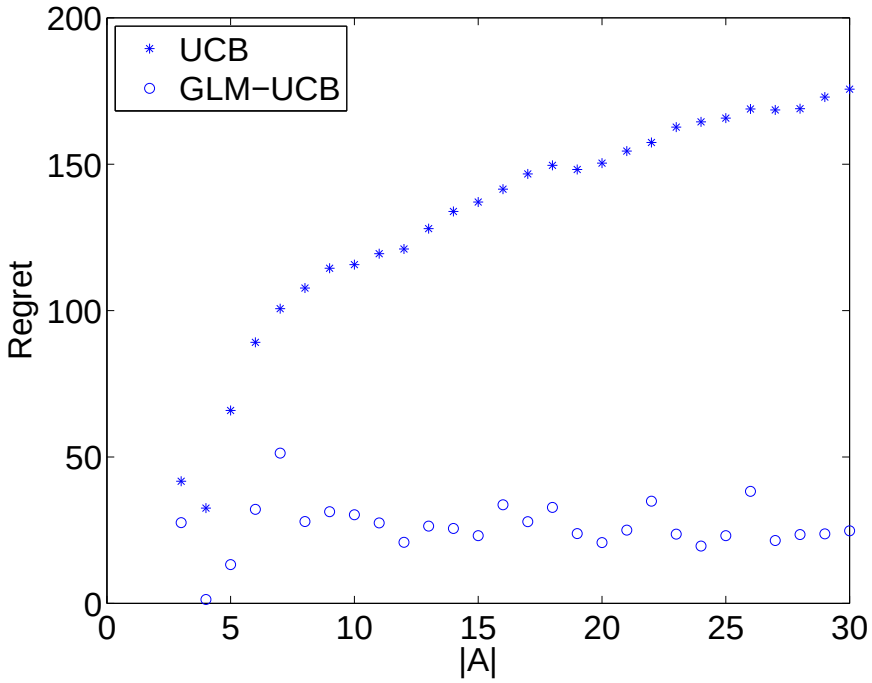

FIGURE 3.1 – Regret des algorithmes UCB et GLM-UCB pour $n = 1000$ en fonction du nombre de bras.

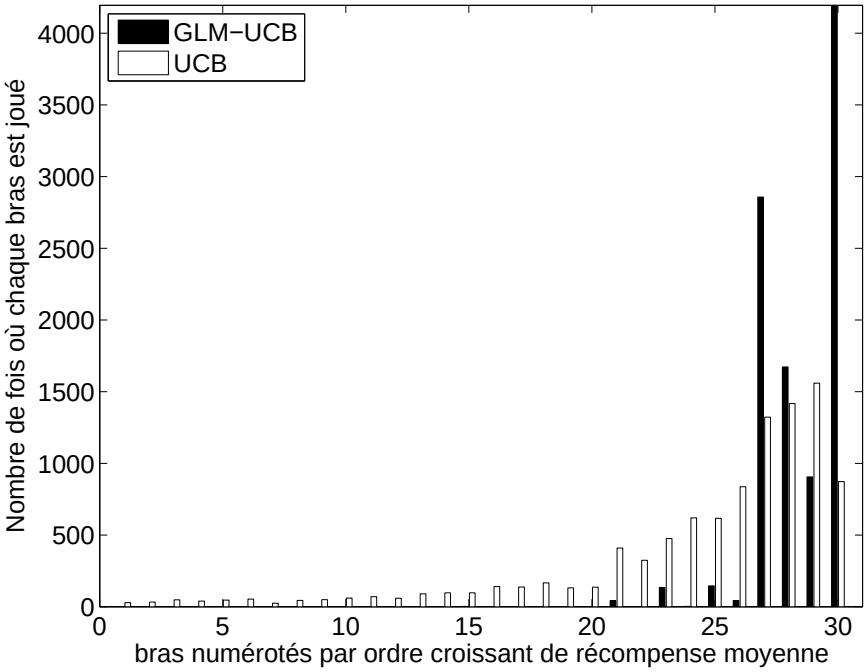

FIGURE 3.2 – Nombre de fois où chaque bras a été joué en suivant respectivement les algorithmes UCB et GLM-UCB sur une trajectoire et un horizon $n = 10000$.

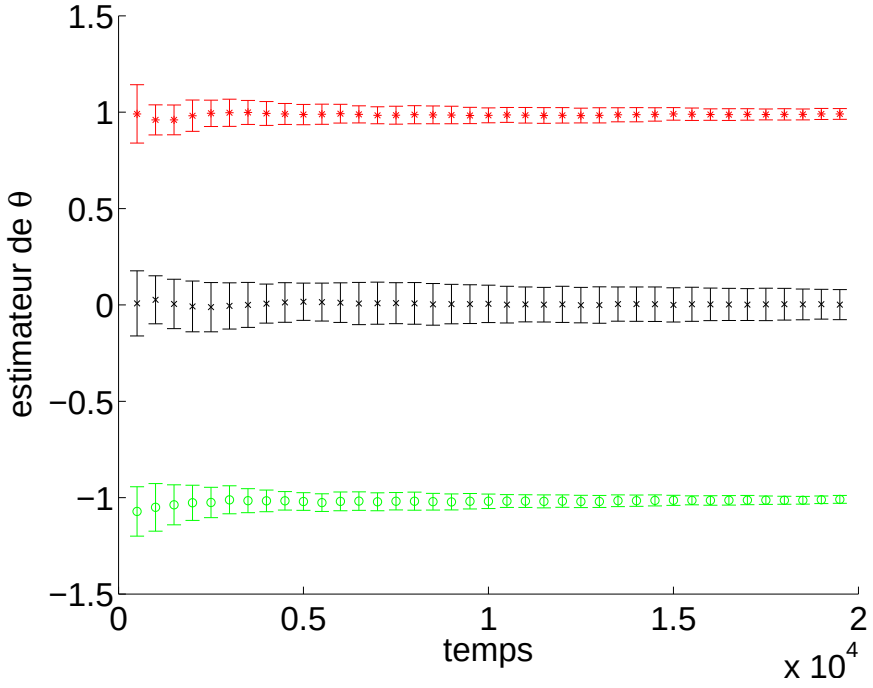

FIGURE 3.3 – Estimateur $\hat{\theta}_t$ en fonction du temps t.

vecteur de caractéristique se trouve aussi dans le plan) et le bras orthogonal a_o. On remarque que le bras orthogonal est joué un nombre proportionnel à $\log(t)$ de fois. En effet, ce bras doit être suffisamment joué pour s'assurer que le bras optimal ne se trouve pas dans cette direction. Parmi les bras qui se trouvent dans le même plan que θ_*, certains ne sont jamais joués, d'autres très peu souvent. Les bras proches de l'optimal sont quant à eux joués de plus en plus souvent. Ceci peut être observé sur la figure 3.4 : le bras proche de l'optimal est peu joué au début puis est ensuite joué de l'ordre de t fois.

3.6.2 Données réelles publiques

Pour évaluer la robustesse de notre méthode, nous avons choisi dans un premier temps d'utiliser une base de donnée publique. La base de donnée « Forest CoverType dataset » du répertoire de données UCI est issu de l'inventaire des arbres de forêts des Etats-Unis. Elle contient une liste d'arbres de 6 espèces différentes accompagnés de leurs caractéristiques (longueur, aspect, éloignement de la source d'eau la plus proche...). Une fois centrés, réduits et après l'ajout d'une covariable (et lorsque l'on ignore toutes les variables catégorielles), ces vecteurs de caractéristiques sont de dimension 11. Ce nuage de points dans l'espace $\mathbb{R}^{11}$ a été partitionné en $|\mathcal{A}| = 32$ clusters à l'aide d'une méthode non-supervisée de quantification vectorielle (*k-means*).

On considère chaque cluster comme étant un bras. Les valeurs des variables de réponse, c'est-à-dire l'espèce des arbres, pour les données associées à chaque cluster sont vues comme les récompenses associées à ce bras. Les centroïdes des clusters sont considérés comme étant les vecteurs de caractéristique de chaque bras. Pour rendre le problème similaire au cas précédent, chaque variable cible est rendue binaire en associant les points de la première classe (l'espèce « Spruce/Fir ») à la récompense $R = 1$ et ceux de toutes les autres classes à $R = 0$. Les proportions de réponses égales à 1 dans chaque cluster (en d'autres termes, la récompense espérée associée à chaque bras) varient de 0.354 à 0.992, tandis que la proportion sur l'ensemble des 581 012 points vaut 0.367. Nous cherchons donc à localiser le plus vite possible

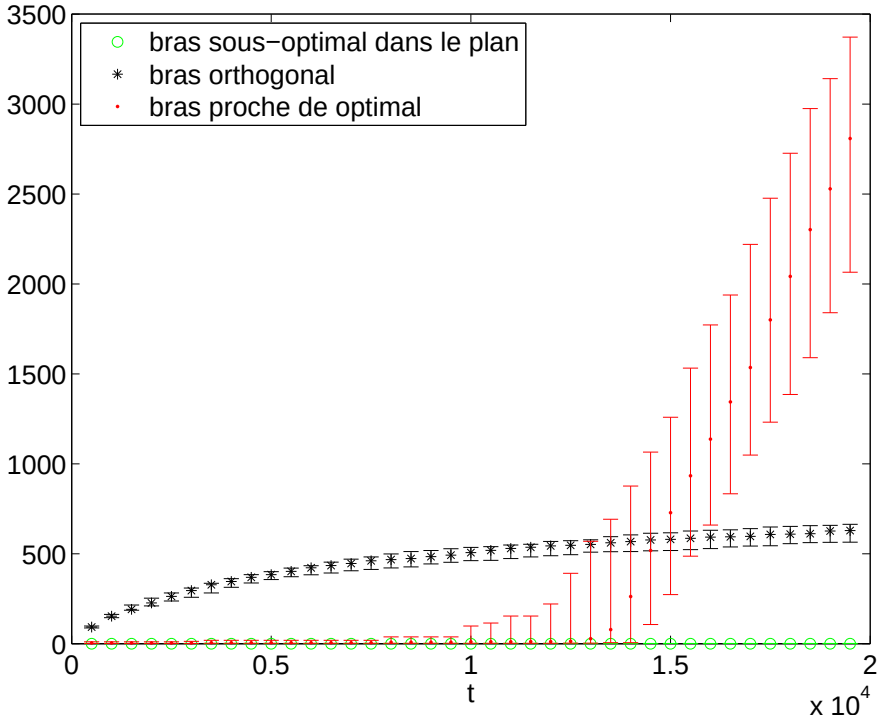

FIGURE 3.4 – Nombre de fois où le bras orthogonal a_o, un bras sous-optimal dans le plan et un bras proche de l'optimal dans le plan sont joués le long de l'algorithme GLM-UCB.

le bras/cluster qui contient la plus grande proportion d'arbres d'une espèce donnée. Ce problème est clairement un « problème jouet », mais il ressemble, par exemple, aux applications mentionnées en introduction, comme l'évaluation d'un médicament ou le problème d'optimisation de l'affichage des publicités sur internet. L'agent est donc confronté à un problème de bandit à 32 bras ayant des vecteurs de caractéristique dans un espace de dimension 11 avec des récompenses binaires. Le modèle de régression logistique n'est pas exactement satisfait puisque les données sont réelles. Cependant, nous espérons certaines régularités par rapport à la position du centroïde du cluster puisque, par exemple, la régression logistique entraînée sur la totalité des données atteint un taux d'erreur de 0.293 (donc inférieur à la performance par défaut de 0.367).

Nous appliquons les algorithmes GLM-UCB, UCB et ϵ-glouton sur 10 trajectoires indépendantes. On observe sur la figure 3.5 que l'algorithme GLM-UCB obtient le plus petit regret moyen. Quand le paramètre est bien estimé, l'algorithme glouton peut trouver un bras optimal en très peu de temps et conduit donc à un faible regret. Cependant, par manque d'exploration, l'algorithme glouton obtient parfois des regrets qui sont considérablement grands (ce phénomène est étudié dans [Audibert et al., 2007]).

3.6.3 Données de publicité sur internet

Pour cette troisième expérience, nous avons utilisé l'enregistrement de l'activité des utilisateurs internet durant une semaine fournie par Orange. Il s'agit d'un fichier contenant environ 5.10^8 visites de 1222 pages internet. Pour chaque visite, le fichier permet de déterminer si l'utilisateur a cliqué ou non sur la publicité qui lui a été présentée. Nous nous sommes restreints à un sous-ensemble de 208 publicités et de 3.10^5 utilisateurs. Les ensembles des pages et des publicités ont été partitionnés respectivement en 10 et 8 catégories en utilisant l'algorithme LDA (*Latent Dirichlet Allocation* [Blei et al., 2002]) appliqué au contenu textuel des pages et

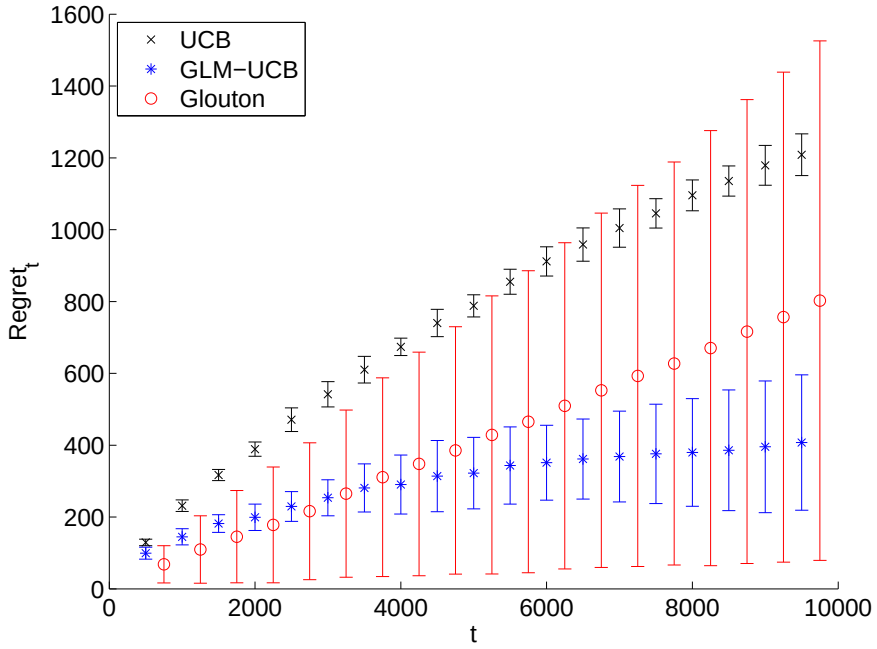

FIGURE 3.5 – Comparaison du regret obtenu en suivant les algorithmes UCB, GLM-UCB et ϵ-gloutons sur les données « Forest CoverType dataset ».

des annonces publicitaires (dans ce dernier cas, il s'agit du contenu textuel de la page vers la quelle le lien publicitaire pointe).

L'espace d'action est composé de 80 paires de catégories de page et catégories de publicité. Quand un couple page-publicité est choisi, il est présenté à un groupe de 50 utilisateurs de la base de donnée, et la récompense est le nombre de personnes ayant cliqué sur la publicité. La récompense moyenne étant typiquement de l'ordre de 0.15, on utilise une fonction logarithmique correspondant à une régression poissonnienne. Le vecteur de covariables pour chaque couple est de dimension 19 : il est composé d'une constante suivie de la concaténation de deux vecteurs de dimensions 10 et 8 représentant respectivement les catégories des pages et des publicités. Dans ce cas, les vecteurs de covariables n'engendrent pas l'espace entier. Pour résoudre ce problème, il est suffisant de considérer la pseudo-inverse de M_t au lieu de l'inverse.

Sur ces données, nous avons comparé l'algorithme GLM-UCB avec les deux algorithmes alternatifs décrits ci-dessus. La figure 3.6 montre que l'algorithme GLM-UCB surpasse une fois de plus les deux autres même si l'écart entre l'algorithme UCB et l'algorithme GLM-UCB est moins important que dans l'exemple précédent. Étant donné le pouvoir peu prédictif des covariables dans cet exemple, il permet d'illustrer le potentiel de notre algorithme pour des applications de la vie réelle. Cependant, cette première expérience étant assez naïve et artificielle pour l'optimisation des ressources sur internet, nous considérons ci-dessous un modèle mieux adapté à ces données.

3.7 Bandits contextuels paramétriques

Dans la première partie de ce chapitre, nous nous sommes intéressés à un modèle de bandit paramétrique avec un nombre $|\mathcal{A}|$ potentiellement très grand de bras et dans lequel l'agent dispose d'une information a priori pour chacun des bras. Nous avons supposé que cette information ne varie pas avec le temps et que le bras optimal est le même à chaque instant. Un autre modèle assez similaire, appelé *bandit contextuel* a été l'objet de nombreuses

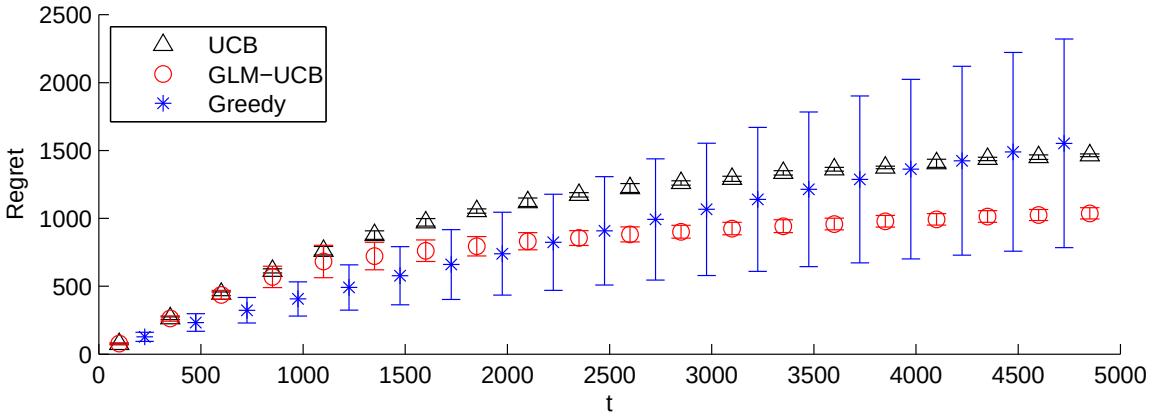

FIGURE 3.6 – Comparaison du regret obtenu en suivant les algorithmes UCB, GLM-UCB et ϵ-gloutons sur les données de publicité internet.

recherches ces dernières années [Wang et al., 2005; Langford and Zhang, 2008; Kakade et al., 2008]. Il s'agit d'un modèle de bandit où l'agent dispose d'une information dite *contextuelle* potentiellement différente à chaque instant. La loi des récompenses reçues en jouant chaque bras dépend de cette information. Le bras optimal peut donc a priori être différent à chaque instant.

Nous proposons d'étendre l'algorithme GLM-UCB a un *modèle de bandit contextuel paramétrique*. Dans ce modèle, on suppose non seulement que l'agent dispose d'un vecteur de caractéristique m_a associée à chaque bras, mais qu'en plus il détient une information sur le contexte à chaque instant. Notons X_t cette information contextuelle à l'instant t et $\mathcal{X}$ l'ensemble de ces informations. Dans la suite, on supposera que l'ensemble $\mathcal{X}$ est fini. Comme précédemment, nous considérons un modèle linéaire généralisé. Ainsi, la récompense reçue à l'instant t si le bras a a été choisi a pour moyenne

$$\mathbb{E}\left[R_t \,|\, X_t = x, A_t = a\right] = \mu(\Phi(x,a)'\theta^*)$$

où $\Phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ est une fonction qui, à une information contextuelle et une action, associe un vecteur d'information de dimension d. La fonction μ et le paramètre $\theta \in \mathbb{R}^d$ sont définis comme précédemment.

Ce modèle est pertinent notamment dans l'exemple de l'optimisation des ressources publicitaires sur internet. En effet, à chaque fois qu'un utilisateur accède à une page internet, le gestionnaire de site sélectionne une (ou plusieurs) publicité(s) à afficher sur la page parmi un ensemble de publicités. Dans cet exemple, l'information contextuelle est la page internet requise par l'utilisateur. A priori, la publicité qui aura le plus grand taux de clic varie d'une page à l'autre. En effet, les utilisateurs accédant à certaines pages sont généralement plus intéressés par certains domaines que par d'autres. Typiquement, dans cette application, le gestionnaire de site a à la fois accès à une information contextuelle qui est la page demandée par l'utilisateur, et à des vecteurs de caractéristiques liés à chaque action, qui sont les informations associées aux publicités. Il parait donc opportun de considérer un modèle de bandit contextuel paramétrique.

3.7.1 Algorithme GLM-UCB*Context*

L'algorithme que nous proposons dans ce modèle est une extension de l'algorithme GLM-UCB. Ce nouvel algorithme, appelé GLM-UCB*Context*, est présenté ci-dessous. A l'instant t, pour chaque bras a, une borne supérieure de $\mu(\Phi(X_t,a)'\hat{\theta}_t)$, égale à $\mu(\Phi(X_t,a)'\hat{\theta}_t) + \beta_t^a$,

est calculée ; β_t^a désigne un bonus d'exploration valant $\rho(t)\|m_a\|_{\widetilde{M}_t^{-1}}$. La matrice de design $\widetilde{M}_t$ est légèrement différente de la matrice M_t précédente puisqu'elle doit tenir compte des informations contextuelles :

$$\widetilde{M}_t = \sum_{k=0}^{t-1} \Phi(X_k, A_k)\Phi(X_k, A_k)' \, .$$

De manière similaire l'estimateur $\hat{\theta}_t$ est solution de l'équation

$$\sum_{k=0}^{t-1} \Big(R_k - \mu(\Phi(X_k, A_k)'\hat{\theta}_t) \Big) \Phi(X_k, A_k) = 0 \, , \tag{3.28}$$

où $X_0, \ldots, X_{t-1}$ et $A_0, \ldots, A_{t-1}$ désignent respectivement les informations contextuelles et les bras joués jusqu'à l'instant t et $R_0, \ldots, R_{t-1}$ sont les récompenses reçues.

Algorithme 3.3 Algorithme GLM-UCB*Context*

1: **Pour** $t \geq 0$ **faire**
2: Observation de l'information contextuelle X_t
3: Calculer $\hat{\theta}_t$
4: Jouer l'action $A_t = \text{argmax}_a \, \mu(\Phi(X_t, a)'\hat{\theta}_t) + \beta_t^a(\delta)$
5: Réception de la récompense R_t
6: **fin Pour**

A la différence de l'algorithme GLM-UCB, jouer un ensemble de d actions au début de l'interaction ne pourra pas garantir que la matrice $\widetilde{M}_d$ soit inversible puisque la matrice dépend également des informations contextuelles. Pour résoudre ce problème technique, nous introduisons un léger biais dans la matrice $\widetilde{M}_t$ en lui ajoutant la matrice identité :

$$\widetilde{M}_t = I_d + \sum_{k=0}^{t-1} \Phi(X_k, A_k)\Phi(X_k, A_k)' \, .$$

Cet artifice est également utilisé par [Dani et al., 2008] dans leur algorithme pour des bandits linéaires.

3.7.2 Résultats théoriques

Les preuves exposées section 3.5 peuvent être aisément adaptées au modèle de bandit contextuel paramétrique en remplaçant les vecteurs de caractéristique m_{A_t} par $\Phi(X_t, A_t)$ et la matrice M_t par la matrice $\widetilde{M}_t$. La seule différence est due à l'ajout de la matrice identité dans la définition de $\widetilde{M}_t$ mais les conséquences de ce changement sont faibles. Il est donc possible de majorer le regret accumulé durant l'algorithme. Les deux hypothèses suivantes sont des variantes des hypothèses 3.3 et 3.4 :

Hypothèse 3.5. *Il existe $c_m < \infty$ tel que pour tout $a \in \mathcal{A}$ et tout $x \in \mathcal{X}$, $\|\Phi(x, a)\|_2 \leq c_m$.*

Hypothèse 3.6. *Soit $\epsilon_t = R_t - \mu(\Phi(A_t, X_t)'\theta_*)$. Pour tout t, les variables aléatoires $(\epsilon_t)_t$ sont indépendantes et centrées. De plus, il existe $R_{\max} > 0$ tel que, pour tout t, les récompenses aléatoires R_t sont positives et bornées par $R_{\max}$.*

Théorème 3.3. *Sous les hypothèses 3.1, 3.2, 3.5 et 3.6 et pour tout horizon n suffisamment grand, le regret de l'algorithme vérifie*

$$\mathbb{P}\left(\mathit{Regret}_n \leq \frac{Cd^2}{\Delta(\theta_*)} R_{\max} \log(n)^2 \log(d\, n/\delta) \right) \geq 1 - \delta \, ,$$

et

$$\mathbb{P}\left(Regret_n \leq dR_{\max} + Cd\log(n)\sqrt{n\log(n/\delta)R_{\max}}\right) \geq 1-\delta\ ,$$

pour des constantes C dépendant de μ et où $\Delta(\theta_*)$ est défini par

$$\Delta(\theta_*) = \min_{x\in\mathcal{X}} \min_{a\in\mathcal{A},\mu(\Phi(m_a,x)'\theta_*)\neq\mu(\Phi(m_{a_*},x)'\theta_*)} \mu(\Phi(m_{a_*},x)'\theta_*) - \mu(\Phi(m_a,x)'\theta_*)\ .$$

3.7.3 Résultats numériques

Comme pour l'algorithme précédent, nous testons l'algorithme GLM-UCB *Contextuel* tout d'abord dans un environnement simulé puis à l'aide de données réelles permettant ainsi de tester la robustesse de l'algorithme.

Données simulées

Nous considérons un modèle de bandit à $|\mathcal{A}| = 100$ bras. A chaque instant t, l'agent observe un contexte $X_t \in \mathcal{X} \overset{\text{def}}{=} \{1,\dots,50\}$ et sélectionne un bras parmi les 100 possibles. Le vecteur de caractéristique $\Phi(x,a)$ associé au contexte x et à l'action a est un vecteur de dimension 3 tiré sous une loi uniforme parmi un ensemble de 150 vecteurs. Ainsi, différents couples état-action partagent les mêmes vecteurs de caractéristique. Les récompenses reçues sont des variables de Bernoulli de paramètres $\mu(\Phi(X_t,A_t)'\theta_*)$ où $\theta^* = (1,-1,0.5)'$ et où μ est la fonction logistique inverse. Comme précédemment, on prend $\rho(t) = 3\log(t)$.

Nous appliquons les algorithmes GLM-UCB *Context* et UCB sur 20 trajectoires indépendantes pour un horizon $n = 10000$. Nous observons sur la figure 3.7 que le regret en suivant l'algorithme GLM-UCB *Context* est beaucoup plus petit qu'en suivant l'algorithme UCB. En effet, l'algorithme GLM-UCB *Context* exploite sa connaissance du modèle. Le paramètre θ étant de dimension 3, il peut être très rapidement estimé et l'algorithme GLM-UCB *Context* connaît ainsi très vite le meilleur bras dans chaque contexte. Le nombre de combinaisons possibles étant $|\mathcal{X}||\mathcal{A}| = 5000$, on pouvait s'attendre à ce que l'algorithme UCB nécessite un temps beaucoup plus long pour déterminer le bras optimal associé à chaque contexte.

Données de publicité sur internet

Nous appliquons maintenant l'algorithme GLM-UCB *Context* à des données réelles. Il s'agit des mêmes données que celles utilisées dans la section 3.6.3. A chaque instant, une page internet est requise par un utilisateur. Pour cette page, l'agent (c'est-à-dire le gestionnaire de site) sélectionne une des publicités dont il dispose et l'affiche. On suppose que l'agent présente la même annonce publicitaire à 50 utilisateurs consultant cette page. La récompense reçue est alors le nombre d'utilisateurs, parmi ces 50, ayant cliqué sur la publicité.

La base de donnée contient 373 publicités et 1751 pages différentes. Pour chacune des publicités (resp. des pages), nous disposons de 4 (resp. 3) variables descriptives. Ces variables sont des entiers désignant le numéro du cluster auquel la publicité (resp. la page) appartient en suivant différentes méthodes de classification de données appliquées au contenu textuel des annonces publicitaires (resp. pages), ou effectuées manuellement. De plus, pour les publicités, une des variables descriptives correspond au format de l'annonce publicitaire.

Les données dont nous disposons étant l'enregistrement de la réaction d'utilisateurs internet face aux publicités qui ont réellement été affichées sur la page qu'ils visitaient, elles sont fortement dépendantes de la stratégie de sélection des publicités utilisée durant la semaine d'enregistrement. En particulier, cette stratégie ne proposant pas toutes les annonces publicitaires sur chacune des pages, nous n'avons à notre disposition la probabilité de clic que de certains couples page-annonce publicitaire. Pour pouvoir illustrer notre algorithme avec ces

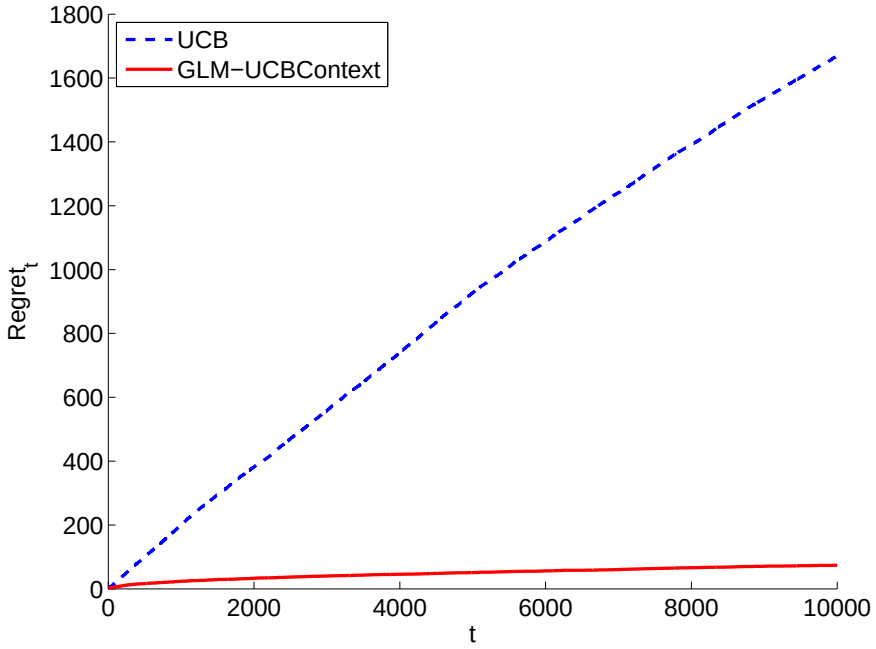

FIGURE 3.7 – Comparaison du regret obtenu en suivant les algorithmes UCB, GLM-UCB*Context* sur un environnement simulé.

données, nous avons sélectionné 6 pages et 63 annonces publicitaires telles que chacune des annonces a été proposée un nombre conséquent de fois sur chacune des pages. Pour 3 des pages, la publicité la plus cliquée est la numéro 7 alors que, pour les 3 autres, il s'agit de la numéro 25.

Pour apprendre la fonction Φ qui, à chaque couple page-annonce publicitaire (x, a) associe un vecteur de caractéristique $\Phi(x, a)$, nous avons utilisé un *algorithme d'analyse en composante principale logistique* [Schein et al., 2003] sur différents ensembles de 5000 couples choisis au hasard parmi tous les 373×1751 couples pages-annonces publicitaires. Nous avons alors sélectionné la fonction Φ qui permet au mieux d'expliquer les probabilités des 6 pages et 63 annonces publicitaires sélectionnées. Pour tout couple (x, a), $\Phi(x, a)$ est un vecteur de dimension $d = 90$.

Pour évaluer le pouvoir prédictif du modèle linéaire généralisé sur ces données, nous comparons la probabilité de clic prédite par le modèle à la « vraie » probabilité de clic d'un couple page-annonce publicitaire calculée sur les données. Cette dernière est définie comme étant le nombre de fois où les utilisateurs visitant une page donnée ont cliqué sur l'annonce publicitaire donnée divisé par la nombre de fois où cette publicité a été affichée sur cette page. Notons θ_* la valeur du paramètre qui explique au mieux les probabilités de clic, c'est-à-dire telle que, pour toute page $1 \leq x \leq 6$ et toute publicité $1 \leq a \leq 63$, $\mu(\Phi(x, a)'\theta_*)$ est proche de la « vraie » probabilité de clic du couple (x, a). Nous utilisons la fonction exponentielle comme fonction de lien μ. Une étude préliminaire montre que, pour chacune des 6 pages x, $\operatorname{argmax}_a \mu(\Phi(x, a)'\theta_*)$ est égale à la publicité la plus cliquée sur la base de donnée. Ainsi le modèle paramétrique permet de prédire la publicité la plus cliquée. La figure 3.8 représente les vraies probabilités de clic pour une page x fixée ainsi que $\mu(\Phi(x, a)'\theta_*)$, où les publicités a sont classées dans l'ordre croissant de vraie probabilité de clic. On observe que les publicités les plus cliquées sont classées dans le bon ordre par le modèle de régression poissonnienne. Néanmoins, l'adéquation du modèle aux données est insuffisante pour permettre au modèle de réellement prédire les vraies probabilités de clic.

Nous avons comparé l'algorithme GLM-UCB*Context* à l'algorithme UCB. Le regret en

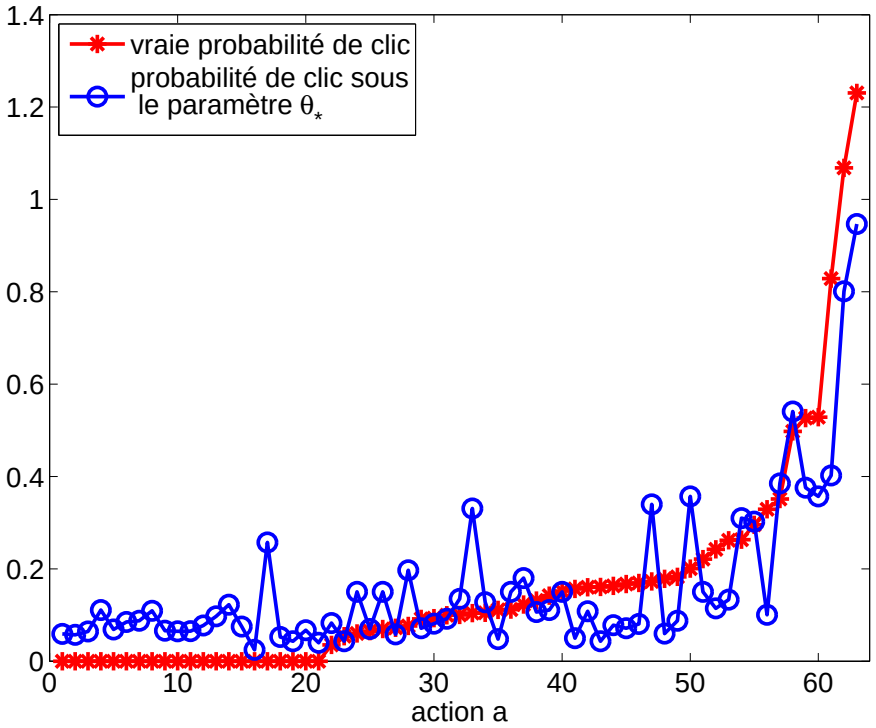

FIGURE 3.8 – Comparaison de la vraie probabilité de clic et celle sous le paramètre θ_* pour une page fixée en fonction des 63 actions classées par ordre croissant de la vraie probabilité de clic.

suivant ces deux algorithmes sur 10 trajectoires indépendantes est représenté sur la figure 3.9. On observe que le regret en suivant l'algorithme GLM-UCB *Context*, bien que plus faible pour les 20000 premiers instants, est parfois nettement plus grand que celui obtenu en suivant l'algorithme UCB quand le nombre d'instants est grand. En effet, notre algorithme étant fondé sur une modélisation paramétrique des récompenses reçues, il est naturel qu'un algorithme non paramétrique tel que UCB soit plus robuste. Ce dernier permet donc d'obtenir, à long-terme, un regret plus faible dans le cas où le modèle n'explique pas assez bien les données.

En analysant la séquence des publicités sélectionnées par l'algorithme GLM-UCB *Context* pour chacun des contextes le long des trajectoires ayant un regret assez élevé, on observe que l'agent réussit à détecter la meilleur publicité sur certaines des pages mais, pour d'autres pages, sélectionne une publicité qui, bien qu'ayant une assez grande probabilité de clic, est sous-optimale. La figure 3.10 représente les vraies probabilités de clic pour une page x pour laquelle l'agent choisit une publicité sous-optimale. On représente sur le même graphe $\mu(\Phi(x,a)'\hat{\theta}_t)$. On remarque que la prédiction de la probabilité de clic de la publicité optimale sous le paramètre estimé est très loin de la vraie probabilité de clic calculée sur toutes les données. Cependant, la prédiction de clic de la publicité ayant la plus grande probabilité de clic sous le paramètre estimé est très proche de la vraie probabilité de clic. En effet, cette annonce publicitaire a été affichée un tellement grand nombre de fois que le paramètre estimé $\hat{\theta}_t$ explique parfaitement la vraie probabilité de clic de cette publicité mais assez mal les autres. On en déduit que, dans un cas tel que celui-ci où le modèle est assez mal spécifié, l'algorithme GLM-UCB *Context* ne permet pas forcément de trouver la valeur $\hat{\theta}_t$ qui prédit au mieux les probabilités de clic de chacune des publicités dans chacun des contextes.

On observe néanmoins sur la figure 3.9 que le regret de l'algorithme GLM-UCB *Context*

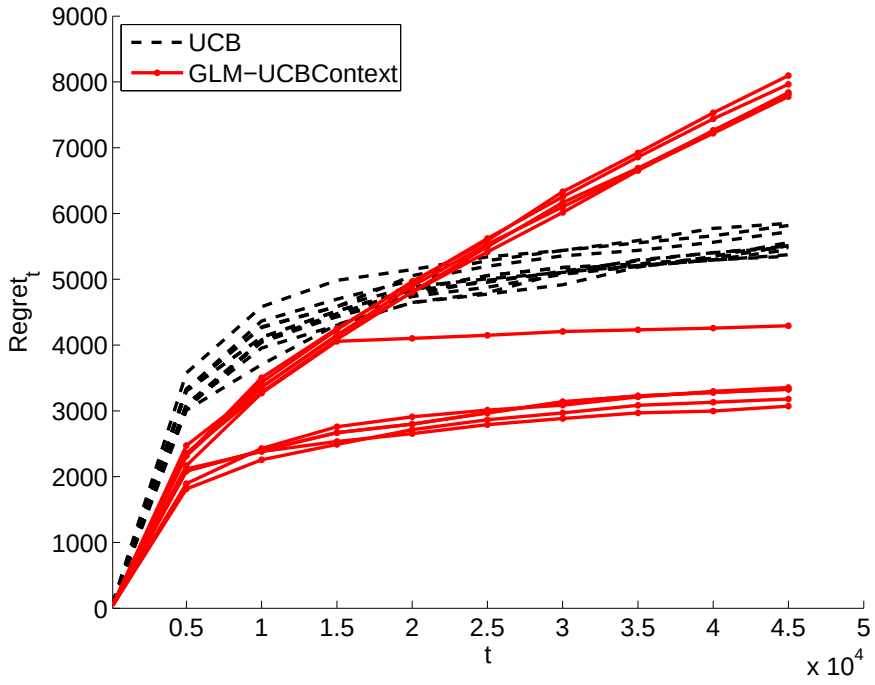

FIGURE 3.9 – Comparaison du regret obtenu en suivant les algorithmes UCB, GLM-UCB*Context* sur un système de séléction de publicité avec des données réelles.

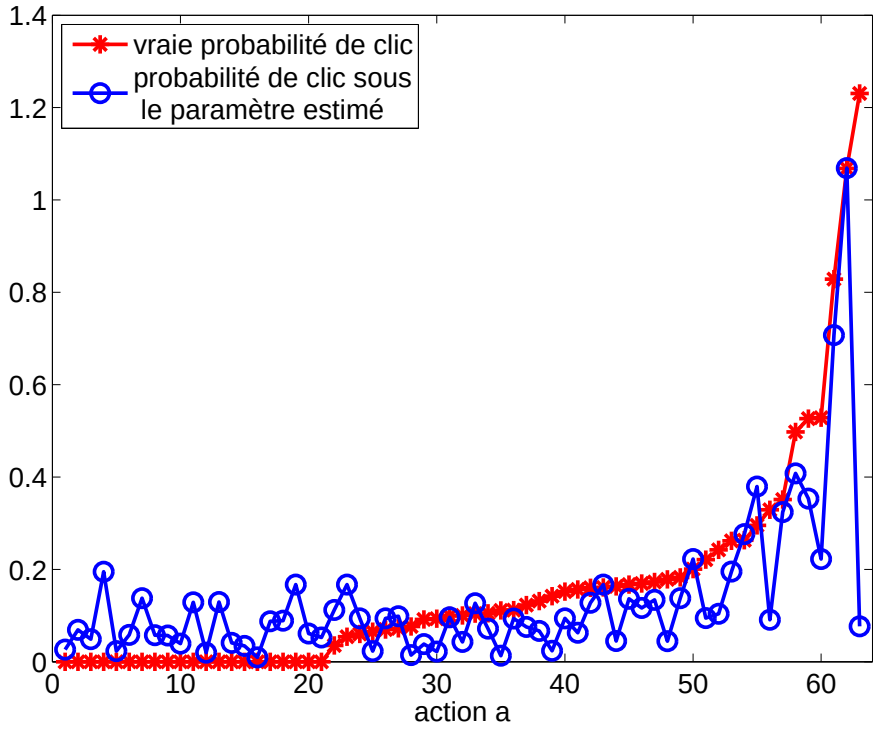

FIGURE 3.10 – Comparaison de la vraie probabilité de clic et celle sous le paramètre $\hat{\theta}_t$ pour une page fixée en fonction des 63 actions classée par ordre croissant de la vraie probabilité de clic.

est plus faible que celui de l'algorithme UCB pour les 20000 premiers instants, même dans un tel cas où le modèle n'explique pas assez bien les données. Or, la publicité sélectionnée par le gestionnaire de site à chaque instant étant proposée à 50 utilisateurs, au bout de 20000 instants, 1 million d'utilisateurs ont visités le site. Ainsi, il peut être intéressant pour un gestionnaire de site d'utiliser un algorithme tel que GLM-UCB*Context* qui permet d'obtenir un plus faible regret que l'algorithme UCB a court et un moyen terme même dans un tel cas.

3.8 Conclusion

Nous nous sommes intéressés à un modèle de bandit paramétrique où l'agent dispose d'une information sur la moyenne des récompenses de chacun des bras. Nous nous sommes placés dans un cadre plus général que le modèle linéaire, proposé par [Auer, 2002], et étudié par [Dani et al., 2008] et [Rusmevichientong and Tsitsiklis, 2008], en considérant un modèle linéaire généralisé qui couvre en particulier le cas, important en pratique, de récompenses binaires à travers la régression logistique. Comme l'algorithme UCB, l'algorithme GLM-UCB, que nous avons proposé, opère directement dans l'espace des récompenses. Nous avons prouvé que l'algorithme obtient un regret d'au plus $O(d^2 \log(n)^3/\Delta(\theta_*))$ avec grande probabilité -où $\Delta(\theta_*)$ dépend du vrai paramètre θ^*- et un regret indépendant de θ_* d'au plus $O\left(d\sqrt{n}\log(n)^{3/2}\right)$. Notons que ces regrets dépendent de la dimension du vecteur de paramètres mais pas du nombre de bras, résultat qui n'avait été prouvé jusqu'à présent que pour des modèles linéaires. De plus, nous avons exposé une méthode permettant de régler les paramètres de l'algorithme afin d'éviter un pessimisme exagéré dans le choix du bonus d'exploration. Dans les simulations numériques, il a été observé que, lorsque les récompenses sont binaires, l'algorithme proposé est efficace et suffisamment robuste pour agir dans des problèmes réels.

Nous avons ensuite étendu l'algorithme au cas des bandits contextuels paramétriques dans lequel l'agent dispose d'une information contextuelle différente à chaque instant qui influe l'espérance de la récompense reçue. Ce modèle est plus adapté à l'optimisation des ressources publicitaires sur internet. Après avoir remarqué que l'algorithme GLM-UCB*Context* admet des bornes de regret similaires à l'algorithme GLM-UCB, nous avons illustré ses performances sur les données réelles fournies par Orange. Ces expériences nous ont permis d'observer les limitations introduites par le caractère paramétrique de notre algorithme. Lorsque le modèle n'est pas en adéquation avec les données, un algorithme non paramétrique tel que UCB permet d'obtenir, à très long terme, des regrets nettement plus faibles.

Nous avons considéré dans ce travail un modèle linéaire généralisé, où la fonction μ est monotone. Ce type de modèle, bien que plus général que le modèle linéaire, reste limitant. Il semblerait intéressant de pouvoir considérer un modèle plus général, cependant, une difficulté majeure pour aborder des modèles plus complexes réside dans l'estimation du paramètre. Un autre problème intéressant qui reste ouvert, et qui est ambitieux même dans le cas linéaire, consisterait à diminuer la différence entre les bornes de confiance pessimistes qui permettent de garantir des regrets logarithmiques et celles suggérées par les arguments asymptotiques présentés section 3.5.2, qui semblent donner de bons résultats en pratique.

CHAPITRE 4

Utilisation de la divergence de Kullback-Leibler dans les algorithmes optimistes

4.1 Introduction

Tout au long de cette thèse nous nous sommes intéressés à des algorithmes par renforcement *« model-based »* dits optimistes. Le principe de *l'optimisme face à l'incertain* est d'agir comme si l'agent se trouvait dans le meilleur des mondes possibles : il suit la politique optimale pour le modèle lui permettant de recevoir la plus grande récompense moyenne parmi les modèles assez proches du modèle estimé. La performance de telles approches, connues pour équilibrer très efficacement exploration et exploitation, peut être analysée en terme de *regret* (voir chapitre 1 pour une définition du regret). Dans les modèles de bandit [Lai and Robbins, 1985] ont prouvé qu'un algorithme optimiste peut induire un regret logarithmique. Ce résultat a été étendu pour des MDP à espace d'états et d'actions finis par [Burnetas and Katehakis, 1997]. Les travaux de [Auer et al., 2002; Audibert et al., 2007] et de [Auer and Ortner, 2007; Jaksch et al., 2010; Bartlett and Tewari, 2009] ont introduit des algorithmes qui garantissent des regrets logarithmiques non asymptotiques respectivement pour des modèles de bandit et pour une grande classe de MDP. Dans ces travaux, le modèle optimiste est calculé en utilisant la norme L^1 (ou variation totale) comme mesure de proximité entre le modèle estimé et les éléments de la classe des modèles compatibles avec les actions et récompenses passées. Il est cependant possible d'utiliser d'autres mesures de proximité et nous nous proposons d'utiliser la divergence de Kullback-Leibler. Cette dernière est en effet connue pour être beaucoup plus adaptée à la géométrie des simplexes de probabilité.

Dans ce chapitre, nous étudions deux algorithme optimistes, appelés KL-UCB et KL-UCRL, respectivement pour un modèle de bandit avec des récompenses de Bernoulli et pour un MDP fini en utilisant la pseudo-distance de Kullback-Leibler (KL) au lieu de la métrique L^1, comme dans [Lai and Robbins, 1985; Burnetas and Katehakis, 1997]. En adaptant l'analyse théorique de [Auer et al., 2002, 2009a; Bartlett and Tewari, 2009; Jaksch et al., 2010], nous obtenons des bornes de regret logarithmiques pour ces deux algorithmes. Les preuves de ces résultats sont basées sur des inégalités de concentration nouvelles pour la divergence de KL, qui ont des propriétés intéressantes comparées à celles traditionnellement utilisées pour la borne L^1. Bien que les bornes de regret auxquelles nous parvenons soient similaires à celles des algorithmes de la littérature en terme de dépendance en le nombre d'états et d'actions, nous avons observé, en pratique, des améliorations significatives quant aux performances des algorithmes. Cette observation est illustrée dans la suite à l'aide de deux exemples classiques

proposés par [Strehl and Littman, 2008] ainsi qu'à travers une discussion sur les propriétés géométriques des voisinages KL.

Ce chapitre est organisé autour de trois sections principales. Les modèles considérés et les principes importants des approches optimistes seront brièvement rappelés dans la section 4.2. Cette section permet, de plus, de présenter les inégalités de concentration pour la divergence de KL qui seront à la base des algorithmes proposés. Les section 4.3 et 4.4 sont réservées à la description et à l'analyse des algorithmes KL-UCB et KL-UCRL. Elles contiennent également des résultats numériques et une discussion sur les avantages de l'utilisation de voisinages de confiance utilisant la divergence de KL plutôt qu'une métrique L^1.

4.2 Modèles et approche optimiste

4.2.1 Modèles considérés

Dans ce chapitre nous utiliserons les modèles ci-après :

Modèle de bandit à récompense binaire

Dans le modèle de bandit à $|\mathcal{A}|$ bras (présenté paragraphe 1.1.1), à chaque instant t, l'agent choisit un bras $A_t \in \{1 \dots |\mathcal{A}|\}$ et reçoit une récompense aléatoire R_t. On suppose, dans ce chapitre, que les récompenses sont des variables de Bernoulli de moyenne inconnue qui dépend du bras sélectionné. Plus précisément, si le bras a est joué, l'agent reçoit une récompense égale à 1 avec probabilité $r(a)$ et égale à 0 avec probabilité $1-r(a)$. La récompense moyenne espérée en jouant le bras a est alors égale à $r(a)$. On appelle *bras optimal* et l'on note a^* tout bras dont la récompense moyenne est la plus grande.

Processus de décision markovien à espace d'états fini

Le deuxième modèle auquel nous nous intéressons est un processus de décision markovien $\mathbf{M} = (\mathcal{X}, \mathcal{A}, P, r)$ où l'espace d'états $\mathcal{X}$, et l'espace d'actions $\mathcal{A}$ sont finis. Dans ce chapitre, nous considérons des MDP *communicant* (voir paragraphe 1.2.2) qui sont tels que pour tous états $x, x' \in \mathcal{X}$, il existe une politique $\pi : \mathcal{X} \mapsto \mathcal{A}$ sous laquelle x' peut être atteint en partant de l'état x avec une probabilité strictement positive. Pour ces MDP, la récompense moyenne $\eta^\pi(\mathbf{M})$ reçue en suivant une politique stationnaire π est indépendante de l'état initial. Dans ce chapitre, nous noterons $\pi^*(\mathbf{M}) : \mathcal{X} \to \mathcal{A}$ et $\eta^*(\mathbf{M})$ respectivement la politique optimale et la récompense moyenne optimale dans le modèle $\mathbf{M}$. Ces notations permettent de souligner le fait que la politique optimale et la récompense moyenne dépendent du modèle $\mathbf{M}$.

4.2.2 Approches « model-based » optimistes

En apprentissage par renforcement, le modèle est supposé inconnu de l'agent. Plus précisément, pour un modèle de bandit, la loi des récompenses est inconnue ; et, pour un processus de décision markovien, les probabilités de transition entre les états sont également inconnues de l'agent. Néanmoins, on suppose en général que les espaces $\mathcal{X}$ et $\mathcal{A}$ sont connus à l'avance. Dans ce paragraphe, nous présentons les approches d'intérêt dans le cadre de MDP. Les modèles de bandit étant un cas particulier de MDP avec un seul état, toutes les notions abordées ici pourront leur être appliquées.

Nous considérons des méthodes d'apprentissage par renforcement dites *« model-based »* : tout au long de l'interaction, l'agent estime le modèle à partir des observations passées et sélectionne l'action à jouer en conséquence. Notons $\hat{P}_t(x, a; x')$ l'estimateur à l'instant t de la probabilité de transition de l'état x à l'état x' conditionnellement à l'action a, et, $\hat{r}_t(x, a)$

l'estimation de la moyenne de la récompense reçue dans l'état x quand l'action a a été choisie. On a :

$$\hat{P}_t(x,a;x') = \frac{N_t(x,a,x')}{\max(N_t(x,a),1)}$$
$$\hat{r}_t(x,a) = \frac{\sum_{k=0}^{t-1} R_k \mathbb{1}_{\{X_k=x,A_k=a\}}}{\max(N_t(x,a),1)} \ , \tag{4.1}$$

où

$$N_t(x,a,x') = \sum_{k=0}^{t-1} \mathbb{1}_{\{X_k=x,A_k=a,X_{k+1}=x'\}}$$

est le nombre de visites, jusqu'à l'instant t, à l'état x suivis par une visite à l'état x' si l'action a a été choisie, et de manière similaire,

$$N_t(x,a) = \sum_{k=0}^{t-1} \mathbb{1}_{\{X_k=x,A_k=a\}} \ .$$

Notons $\hat{\mathbf{M}}_t = (\mathcal{X}, \mathcal{A}, \hat{P}_t, \hat{r}_t)$ le MDP estimé à l'instant t.

Que ce soit dans un modèle de bandit ou dans un MDP, la politique qui consiste à choisir les actions optimales pour le modèle estimé, appelée *politique gloutonne* par rapport au modèle estimé, peut être trompeuse. Nous verrons dans la section consacrée aux résultats numériques que pour certains MDP jouer constamment la politique gloutonne par rapport au modèle estimé permet parfois d'atteindre la politique optimale. Cependant, dans de nombreux modèles, une très bonne politique d'exploration est nécessaire pour la déterminer. Les approches dites optimistes sont connues pour équilibrer efficacement exploration et exploitation. Au lieu de s'intéresser uniquement au modèle estimé, ces algorithmes déterminent un ensemble de modèles cohérents avec les observations passées, c'est-à-dire avec les récompenses reçues étant donné les couples état-action visités. L'ensemble $\mathcal{M}_t$ de modèles compatibles avec les observations à l'instant t inclut le modèle estimé $\hat{\mathbf{M}}_t$. Cet ensemble est défini comme suit :

$$\mathcal{M}_t = \{\mathbf{M} = (\mathcal{X}, \mathcal{A}, P, r) : \forall x \in \mathcal{X}, \forall a \in \mathcal{A},\ d_R(\hat{r}_t(x,a), r(x,a)) \leq \frac{C_R}{g_R(N_t(x,a))}$$
$$\text{et } d_P(\hat{P}_t(x,a;.), P(x,a;.)) \leq \frac{C_P}{g_P(N_t(x,a))}\} \ . \tag{4.2}$$

Il s'agit de l'ensemble des modèles tels que la loi des récompenses et les probabilités de transition soient dans un voisinage respectivement de $\hat{r}_t$ et de $\hat{P}_t$. Les fonctions réelles d_R et d_P mesurent respectivement la proximité entre les lois de récompenses et les probabilités de transition ; les constantes C_P et C_R permettent de contrôler le *rayon des voisinages* ; les fonctions $g_R : \mathbb{R} \mapsto \mathbb{R}$ et $g_P : \mathbb{R} \mapsto \mathbb{R}$ sont strictement croissantes.

Un algorithme optimiste consiste à sélectionner, parmi l'ensemble $\mathcal{M}_t$, le modèle qui permet d'obtenir la plus grande récompense moyenne. Plus précisément, l'agent choisit le modèle $\mathbf{M}_t$ tel que

$$\mathbf{M}_t = \underset{\mathbf{M} \in \mathcal{M}_t}{\operatorname{argmax}}\, \eta^*(\mathbf{M}) \ .$$

Les rayons des voisinages autour des lois de récompenses et des probabilités de transitions sont déterminés par les constantes C_R et C_P ainsi que par les fonctions g_R et g_P. Ces valeurs sont choisies de manière à ce que le vrai modèle $\mathbf{M}$ appartienne, à chaque instant t, à l'ensemble $\mathcal{M}_t$ avec grande probabilité, ce qui est rendu possible par l'utilisation d'inégalités de concentration. La plupart des algorithmes, notamment les algorithmes UCB (pour les modèles de bandit [Auer et al., 2002]) et UCRL (pour les MDP [Auer and Ortner, 2007; Auer

et al., 2009a; Jaksch et al., 2010]) utilisent les inégalités de Hoeffding et de Hoeffding-Azuma, décrites en annexe (voir théorèmes A.1 et A.2). Dans ces inégalités, la mesure de proximité utilisée est la distance L^1. Par exemple, en utilisant l'inégalité de concentration exposée en annexe dans l'équation (A.1), on a, pour tout état x, toute action a, et tout instant t

$$\left\|\hat{P}_t(x,a;.) - P(x,a;.)\right\|_1 \stackrel{\text{def}}{=} \sum_{x' \in \mathcal{X}} \left|\hat{P}_t(x,a;x') - P(x,a;x')\right| \\ \leq |\mathcal{X}| \frac{C_P}{\sqrt{N_t(x,a)}} \quad \text{avec probabilité } 1 - 4\log(t)\exp\left\{-1,99C_P^2\right\}. \tag{4.3}$$

lorsque l'on considère des voisinages L^1, il est également possible d'utiliser l'inégalité de Weissman, plus adaptée si le nombre d'états est grand, ou celle de Bernstein qui utilise une estimation de la variance (voir annexe A.1).

Cependant, pour mesurer la proximité entre deux probabilités, il nous paraît plus opportun de se fier à la divergence de Kullback-Leibler (KL) plutôt qu'à une distance L^1. La divergence de KL est en effet très souvent utilisée pour différencier deux probabilités car elle est bien adaptée à la géométrie du simplexe de probabilité. Les premiers articles traitant de modèles optimistes ont d'ailleurs proposé d'utiliser cette mesure [Lai and Robbins, 1985; Burnetas and Katehakis, 1997]. Cependant, aucun des algorithmes optimistes garantissant des bornes de regret non asymptotiques n'ont continué à utiliser la divergence de KL, peut être par faute d'inégalités de concentration dédiées à cette mesure de proximité. Nous présentons dans le paragraphe suivant des inégalités de concentration proposées très récemment pour la divergence de KL. Elles nous serviront à construire des modèles optimistes basés sur la divergence de Kullback-Leibler plutôt que sur une distance L^1.

4.2.3 Inégalités de concentration utilisant la divergence de KL

Nous présentons ici les deux inégalités de concentration spécifiques à la divergence de Kullback-Leibler que nous utiliserons dans ce chapitre. Ces dernières ont été proposées très récemment par [Garivier and Leonardi, 2010]. On rappelle que la divergence de Kullback-Leibler entre deux vecteurs de probabilités p et q de $\mathbb{S}^N = \{p \in [0,1]^N,\ \sum_{i=1}^N p_i = 1\}$ est définie par

$$KL(p;q) = \sum_{i=1}^{N} p_i \log\left(\frac{p_i}{q_i}\right) .$$

Nous souhaitons utiliser la divergence de KL dans deux contextes assez différents. Le premier, qui correspondra par la suite au modèle de bandit avec des récompenses binaires, a pour but de mesurer la proximité entre deux lois de Bernoulli. Le deuxième consiste à quantifier la différence entre deux vecteurs de probabilité de dimension n. Nous énonçons ici les résultats en utilisant directement les notations des MDP et des bandits. Une version plus standard de l'inégalité de concentration est énoncée en annexe (voir théorème A.9).

Pour le modèle de bandit

Nous considérons le modèle de bandit présenté dans le paragraphe 4.2.1. Soient a un bras fixé et $(R_k)_{k\geq 0}$ une séquence de variables aléatoires indépendantes et identiquement distribuées selon une loi de Bernoulli de paramètre $r(a)$. Ces variables correspondent à la suite des récompenses reçues lorsque le bras a a été joué. Nous rappelons qu'entre l'instant 0 et l'instant t, le bras a a été joué $N_t(a)$ fois, où $N_t(a)$ est une quantité aléatoire dépendant

de la politique suivie pour sélectionner les actions. L'estimateur $\hat{r}_t(a)$ de $r(a)$ à l'instant t est défini par

$$\hat{r}_t(a) = \frac{\sum_{k=0}^{t-1} R_k \mathbb{1}_{\{A_k=a\}}}{N_t(a)}$$

avec la convention que $\hat{r}_t(a) = R_{\max}$ si $N_t(a) = 0$.

A chaque instant t et pour chaque bras a, on détermine l'ensemble des lois de Bernoulli de paramètre $r(a)$ ayant un niveau de vraisemblance supérieur ou égal à un seuil fixé. La divergence de Kullback-Leibler entre la loi de Bernoulli de paramètre $\hat{r}_t(a)$ et toutes les lois de Bernoulli de paramètre $r(a)$ est alors inférieure ou égale à un seuil noté β_t^a. Notons $kl(p;q)$ la divergence de KL entre deux lois de Bernoulli de paramètres p et q :

$$kl(p;q) \stackrel{\text{def}}{=} p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right) \ .$$

On construit ainsi l'ensemble des lois de probabilité vraisemblables à l'instant t de la manière suivante

$$\mathcal{M}_t = \left\{ r \in [0,1]^{|\mathcal{A}|} : \forall a \in \mathcal{A},\ kl(\hat{r}_t(a); r(a)) \leq \beta_t^a \right\} \ .$$

Une attention particulière doit être portée au sens dans lequel on considère la divergence de Kullback-Leibler. Nous avons considéré la divergence de KL de $r(a)$ par rapport à $\hat{r}_t(a)$, et non l'inverse. C'est le sens qui apparaît naturellement lorsque l'on considère des niveaux de confiance. De plus, nous avons dans ce cas des inégalités pour contrôler ce niveau de confiance (voir théorème ci-dessous). Cette mesure de dissimilarité permet en particulier d'envisager des lois de récompense de paramètre $r(a)$ strictement positif même si l'estimateur $\hat{r}_t(a)$ vaut 0. Cela n'aurait pas été possible si nous avions considéré la divergence de KL de $\hat{r}_t(a)$ par rapport à $r(a)$. Or, il est important dans un modèle optimiste de pouvoir supposer que l'espérance de la récompense est strictement positive même si, jusqu'à l'instant t, seuls des 0 ont été observés.

Le théorème suivant présente une inégalité de concentration permettant de déterminer le seuil β_t^a de manière à ce que l'espérance des récompenses des bras sous le vrai modèle appartienne à l'ensemble $\mathcal{M}_t$ avec grande probabilité.

Théorème 4.1. *Pour tout instant t, pour tout $\epsilon > 0$ et pour toute action a, on a*

$$\mathbb{P}\left[N_t(a)\ kl(\hat{r}_t(a); r(a)) > \epsilon\right] \leq 2e\ \lceil \epsilon \log(t) \rceil \exp\{-\epsilon\} \ ,$$

où $\lceil c \rceil$ est le plus petit entier supérieur ou égal à c.

D'après ce théorème, l'espérance des récompenses des bras sous le vrai modèle appartient à l'ensemble

$$\mathcal{M}_t = \left\{ r \in [0,1]^{|\mathcal{A}|} : \forall a \in \mathcal{A},\ kl(\hat{r}_t(a); r(a)) \leq \frac{C_R}{N_t(a)} \right\}$$

avec une probabilité supérieure à $1 - 2e|\mathcal{A}|\ \lceil C_R \log(t) \rceil \exp\{-C_R\}$.

Pour un MDP

Dans un processus de décision markovien, en plus de la loi des récompenses, l'agent doit estimer les probabilités de transition entre les états conditionnellement aux action effectuées. Une inégalité de concentration permettant de majorer la différence entre les probabilités de transition estimées par comptage des occurrences de saut et la vraie probabilité de transition est exposée dans le théorème suivant.

Théorème 4.2. *Pour tout instant t, pour tout $\epsilon > 0$, pour tout couple $(x,a) \in \mathcal{X} \times \mathcal{A}$, on a*

$$\mathbb{P}\left[N_t(x,a)\ KL(\hat{P}_t(x,a;.);P(x,a;.)) > \epsilon\right] \leq 2e\ (\epsilon \log(t) + |\mathcal{X}|) \exp\left\{-\frac{\epsilon}{|\mathcal{X}|}\right\}\ .$$

Démonstration. le théorème découle aisément du théorème 4.1 en utilisant le résultat suivant (prouvé dans [Garivier and Leonardi, 2010])

$$KL(\hat{P}_t(x,a;.);P(x,a;.)) \leq \sum_{x' \in \mathcal{X}} kl(\hat{P}_t(x,a;x');P(x,a;x'))\ .$$

On a ainsi :

$$\begin{aligned}
\mathbb{P}\left[N_t(x,a)\ KL(\hat{P}_t(x,a;.);P(x,a;.)) > \epsilon\right] &\leq \mathbb{P}\left[\sum_{x' \in \mathcal{X}} N_t(x,a)\ kl(\hat{P}_t(x,a;x');P(x,a;x')) > \epsilon\right] \qquad (4.4)\\
&\leq \sum_{x' \in \mathcal{X}} \mathbb{P}\left[N_t(x,a)\ kl(\hat{P}_t(x,a;x');P(x,a;x')) > \frac{\epsilon}{|\mathcal{X}|}\right]\\
&\leq \sum_{x' \in \mathcal{X}} 2e\ \left\lceil \frac{\epsilon}{|\mathcal{X}|}\log(t)\right\rceil \exp\left\{-\frac{\epsilon}{|\mathcal{X}|}\right\}\\
&\leq 2e|\mathcal{X}|\ \left\lceil \frac{\epsilon}{|\mathcal{X}|}\log(t)\right\rceil \exp\left\{-\frac{\epsilon}{|\mathcal{X}|}\right\}\\
&\leq 2e\ (\epsilon \log(t) + |\mathcal{X}|) \exp\left\{-\frac{\epsilon}{|\mathcal{X}|}\right\} \qquad (4.5)
\end{aligned}$$

□

Au vu de cette preuve, on remarque que l'inégalité de concentration peut être améliorée si la connectivité du MDP est faible. On dit qu'un état x' est *accessible en un coup* à partir de $(x,a) \in \mathcal{X} \times \mathcal{A}$ si $P(x,a;x') > 0$. Notons $\mathcal{X}^1_{(x,a)}$ l'ensemble des états accessibles en un coup à partir du couple état-action (x,a). Si on savait à l'avance que, pour un couple état action (x,a) donné, $|\mathcal{X}^1_{(x,a)}|$ est strictement inférieur à $|\mathcal{X}|$, alors, dans l'équation (4.4), la somme sur tous les états $x' \in \mathcal{X}$ peut être remplacée par la somme sur tous les états appartenant à $\mathcal{X}^1_{(x,a)}$. Ainsi, la dernière majoration (voir équation (4.5)) peut être affinée en remplaçant $|\mathcal{X}|$ par $|\mathcal{X}^1_{(x,a)}|$.

Soit κ un entier mesurant le nombre maximum d'états atteignables en un coup quelque soit l'état de départ :

$$\kappa \stackrel{\text{def}}{=} \max_{x \in \mathcal{X}, a \in \mathcal{A}} \sum_{x' \in \mathcal{X}} \mathbb{1}_{\{P(x,a;x')>0\}} = \max_{x \in \mathcal{X}, a \in \mathcal{A}} |\mathcal{X}^1_{(x,a)}|\ . \qquad (4.6)$$

κ est un entier compris entre 1 et $|\mathcal{X}|$ appelé la *connectivité du MDP*. On dit qu'un MDP est *faiblement connecté* si $\kappa < |\mathcal{X}|$. La proposition suivante permet d'affiner l'inégalité du théorème 4.2 dans le cas d'un MDP faiblement connecté.

Proposition 4.1. *Soit $\mathbf{M} = (\mathcal{X}, \mathcal{A}, P, r)$ un MDP de connectivité κ. Pour tout instant t, pour tout $\epsilon > 0$, pour tout couple $(x,a) \in \mathcal{X} \times \mathcal{A}$, on a*

$$\mathbb{P}\left[N_t(x,a)\ KL(\hat{P}_t(x,a;.);P(x,a;.)) > \epsilon\right] \leq 2e\ (\epsilon \log(t) + \kappa) \exp\left\{-\frac{\epsilon}{\kappa}\right\}\ .$$

Le théorème 4.2 (ou la proposition 4.1 si le MDP est faiblement connecté) permet de construire des ensembles de MDP cohérents avec les observations passées qui contiennent le vrai MDP avec grande probabilité. En utilisant ces bornes de concentration ainsi qu'une inégalité de déviation L^1 pour les récompenses (voir A.1), l'ensemble

$$\mathcal{M}_t = \Bigg\{ \mathbf{M} = (\mathcal{X}, \mathcal{A}, P, r) : \forall x \in \mathcal{X}, \forall a \in \mathcal{A},\ |\hat{r}_t(x,a) - r(x,a)| \leq \frac{C_R}{\sqrt{N_t(x,a)}} \\ \text{et } KL(\hat{P}_t(x,a;.); P(x,a;.)) \leq \frac{C_P}{N_t(x,a)} \Bigg\} \quad (4.7)$$

contient le vrai modèle avec une probabilité plus grande que

$$1 - 2e|\mathcal{X}||\mathcal{A}|\ (C_P \log(t) + |\mathcal{X}|) \exp\left\{ -\frac{C_P}{|\mathcal{X}|} \right\} - 4|\mathcal{X}||\mathcal{A}| \log(t) \exp\left\{ -\frac{1,99 C_R^2}{R_{\max}} \right\} .$$

Si, de plus, il s'agit d'un MDP faiblement connecté dont on connaît la structure, alors, la probabilité que le vrai modèle soit contenu dans $\mathcal{M}_t$ est majorée par

$$1 - 2e|\mathcal{X}||\mathcal{A}|\ (C_P \log(t) + \kappa) \exp\left\{ -\frac{C_P}{\kappa} \right\} - 4|\mathcal{X}||\mathcal{A}| \log(t) \exp\left\{ -\frac{1,99 C_R^2}{R_{\max}} \right\} ,$$

où κ est la connectivité du MDP.

Notons que le diamètre des voisinages de confiance construits en utilisant la divergence de Kullback-Leibler décroît en $1/N_t(x,a)$, tandis que le diamètre des voisinages L^1 diminue en $1/\sqrt{N_t(x,a)}$ (voir équation (4.3) par exemple). En effet, la divergence de Kullback-Leibler a un comportement quadratique comparé à la distance L^1 : pour tous vecteurs de probabilité $p, q \in \mathbb{S}^n$, la divergence de KL est majorée par la distance du Chi-2

$$KL(p;q) \leq \frac{1}{2} \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i} ,$$

et est minorée par le carré d'une distance L^1 (voir annexe A.2)

$$\|p - q\|_1 \leq \sqrt{2 KL(p;q)} .$$

4.3 Algorithme KL-UCB

Nous présentons dans cette section un algorithme optimiste, appelé KL-UCB, pour agir dans un modèle de bandit où les récompenses reçues sont des variables aléatoires suivant des lois de Bernoulli dont les paramètres dépendent du bras joué. L'algorithme est décrit ci-dessous. Il consiste à calculer, à chaque instant t et pour chaque bras a, la récompense moyenne optimiste $r_t(a)$ définie par

$$r_t(a) = \max \left\{ r \in [0,1] ,\ \text{tel que } kl(\hat{r}_t(a); r) \leq \frac{\alpha \log(t \log(t))}{N_t(a)} \right\} , \quad (4.8)$$

où α est une constante strictement supérieure à 1. L'agent sélectionne alors le bras qui a la plus grande moyenne optimiste :

$$A_t = \operatorname*{argmax}_{a \in \mathcal{A}} r_t(a) .$$

L'étape clé de l'algorithme est la recherche du *modèle optimiste* qui consiste à trouver, pour toute action a, la moyenne de la récompense $r_t(a)$ définie par l'équation (4.8) (étape 5). Cette étape consiste à déterminer le zéro de la fonction convexe $q \mapsto kl(\hat{r}_t(a); q)$ sur l'intervalle $[\hat{r}_t(a), 1]$. La solution numérique de ce problème peut donc être calculée très facilement.

Algorithme 4.1 KL-UCB

Entrées: Paramètre $\alpha > 1$

1: Initialisation : Jouer une fois chaque bras
2: Réception de $R_0, \ldots, R_{|\mathcal{A}|-1}$
3: Calculer l'estimateur $\hat{r}(a)$ pour tout $a \in \mathcal{A}$
4: **Pour** chaque $t \geq |\mathcal{A}|$ **faire**
5: Pour chaque bras $a \in \mathcal{A}$, calculer.

$$r_t(a) = \max \left\{ r \in [0,1] \ , \text{ tel que } kl(\hat{r}_t(a); r) \leq \frac{\alpha \log(t \log(t))}{N_t(a)} \right\}$$

6: Jouer l'action $A_t = \text{argmax}_a \, r_t(a)$
7: Réception de la récompense R_t
8: Mise à jour de l'estimateur $\hat{r}(A_t)$
9: **fin Pour**

4.3.1 Analyse théorique

Pour analyser la performance de l'algorithme KL-UCB, nous calculons le regret à l'horizon n :

$$\text{Regret}_n = \sum_{t=0}^{n} r(a^*) - r(A_t) \ .$$

Soit $\Delta(a)$ l'écart entre l'espérance de la récompense reçue en jouant le bras a et celle reçue en jouant le meilleur bras :

$$\Delta_a = r(a^*) - r(a) \ .$$

Le regret peut alors s'écrire comme suit :

$$\text{Regret}_n = \sum_{a \in \mathcal{A}} N_{n+1}(a) \Delta_a \ .$$

Le théorème suivant présente une borne supérieur de l'espérance du regret.

Théorème 4.3. *Pour tout bras a, pour tout $\alpha > 1$ et pour tout instant $n \geq |\mathcal{A}|$, on a*

$$\mathbb{E}(\textit{Regret}_n) \leq \sum_{a \in \mathcal{A}: \Delta_a > 0} \frac{32 \alpha \log(n \log(n))}{\Delta_a} + \frac{\alpha \Delta_a u(n)}{(\alpha - 1)^2} \ ,$$

où $u(n) = o(1)$.

Cette borne du regret espéré est du même ordre de grandeur que celles obtenues précédemment pour les algorithmes UCB et UCB-V [Auer et al., 2002; Audibert et al., 2007] :

$$\text{UCB} : \mathbb{E}(\text{Regret}_n) \leq \sum_{a \in \mathcal{A}: \Delta_a > 0} \frac{4\alpha \log(n)}{\Delta_a} + \Delta_a \left(1 + \frac{4}{\log(\alpha + 1/2)} \left(\frac{\alpha + 1/2}{\alpha - 1/2} \right)^2 \right)$$

$$\text{UCB-V} : \mathbb{E}(\text{Regret}_n) \leq 8\alpha \sum_{a \in \mathcal{A}: \Delta_a > 0} \left(\frac{\sigma_a}{\Delta_a} + 2 \right) \log(n) + \Delta_a \left(2 + \frac{12}{\log(\alpha + 1)} \left(\frac{\alpha + 1}{\alpha - 1} \right)^2 \right) \ .$$

Démonstration. Soit $a \in \mathcal{A}$ un bras sous-optimal, c'est-à-dire tel que $r(a) < r(a^*)$. La preuve de ce théorème consiste à majorer le nombre $N_{n+1}(a)$ de fois que ce bras a été joué avant l'instant $n + 1$. Rappelons que l'on note $r(a)$ la récompense moyenne reçue lorsque l'on joue

le bras a ; $\hat{r}_t(a)$ et $r_t(a)$ désignent respectivement la récompense moyenne associée au bras a sous le modèle estimé et sous le modèle optimiste.

Montrons, tout d'abord, que si, pour une certaine valeur de t, $r(a)$ et $r(a^*)$ vérifient

$$kl(\hat{r}_t(a); r(a)) \leq \frac{\alpha \log(t \log(t))}{N_t(a)} \qquad \text{et} \qquad kl(\hat{r}_t(a^*); r(a^*)) \leq \frac{\alpha \log(t \log(t))}{N_t(a^*)}$$

et si, de plus

$$N_t(a) \geq \frac{32\alpha \log(t \log(t))}{\Delta(a)^2} \tag{4.9}$$

alors $A_t \neq a$. En effet, par définition du modèle optimiste, $r_t(a^*) \geq r(a^*)$. De plus, en utilisant la définition de $\Delta(a)$ et d'après l'équation (4.9), on a

$$r_t(a^*) \geq r(a^*) = \Delta(a) + r(a) \geq r(a) + 4\sqrt{\frac{2\alpha \log(t \log(t))}{N_t(a)}} \ .$$

En utilisant l'inégalité de Pinsker (voir équation (A.4) en annexe) et le fait que $r(a)$ et $r_t(a)$ sont dans un même voisinage autour de $\hat{r}_t(a)$, on a les deux inégalités suivantes :

$$\hat{r}_t(a) - r(a) \leq |\hat{r}_t(a) - r(a)| \leq 2\sqrt{2kl(\hat{r}_t(a); r(a))} \leq 2\sqrt{\frac{2\alpha \log(t \log(t))}{N_t(a)}}$$

et

$$r_t(a) - \hat{r}_t(a) \leq |\hat{r}_t(a) - r_t(a)| \leq 2\sqrt{2kl(\hat{r}_t(a); r_t(a))} \leq 2\sqrt{\frac{2\alpha \log(t \log(t))}{N_t(a)}}$$

D'où,

$$r_t(a^*) \geq r(a) + 4\sqrt{\frac{2\alpha \log(t \log(t))}{N_t(a)}} \geq \hat{r}_t(a) + 2\sqrt{\frac{2\alpha \log(t \log(t))}{N_t(a)}} \geq r_t(a) \ ,$$

ce qui implique que le bras a n'est pas joué à l'instant t puisque, sous le modèle optimiste, sa récompense est inférieure à celle du bras a^*. Notons que ce raisonnement a été fait pour un seul bras fixé et que cela n'implique donc pas que l'agent sélectionne le bras optimal a^*. En effet, la récompense moyenne d'un autre bras, sous ce modèle optimiste, pourrait être plus grande que $r_t(a^*)$. Néanmoins, on a montré que si $A_t = a$, alors, au moins une des inégalités suivantes est vraie :

$$kl(\hat{r}_t(a); r(a)) > \frac{\alpha \log(t \log(t))}{N_t(a)} \tag{4.10}$$

$$kl(\hat{r}_t(a^*); r(a^*)) > \frac{\alpha \log(t \log(t))}{N_t(a^*)} \tag{4.11}$$

$$N_t(a) \leq \frac{32\alpha \log(t \log(t))}{\Delta(a)^2} \ . \tag{4.12}$$

Posons $u = \left\lceil \frac{32\alpha \log(n \log(n))}{\Delta_a^2} \right\rceil$. On a

$$\begin{aligned}
\mathbb{E}\left[N_{n+1}(a)\right] &= \mathbb{E}\left[\sum_{t=0}^{n} \mathbb{1}_{\{A_t=a\}}\right] \leq u + \mathbb{E}\left[\sum_{t=u}^{n} \mathbb{1}_{\{A_t=a \text{ et } N_t(a)>u\}}\right] \\
&\leq u + \mathbb{E}\left[\sum_{t=u}^{n} \mathbb{1}_{\{A_t=a \text{ et (4.12) est fausse}\}}\right] \\
&\leq u + \mathbb{E}\left[\sum_{t=u}^{n} \mathbb{1}_{\{\text{(4.10) ou (4.11) est vraie}\}}\right] \\
&\leq u + \sum_{t=u}^{n} \mathbb{P}(\text{(4.10) est vraie}) + \mathbb{P}(\text{(4.11) est vraie}) \ .
\end{aligned}$$

En utilisant l'inégalité de concentration exposée dans le théorème 4.1, les probabilités $\mathbb{P}$((4.10) est vraie) et $\mathbb{P}$((4.11) est vraie) sont majorées par

$$2e\alpha \log(t\log(t)) \log(t) \exp\left\{-\alpha \log(t\log(t))\right\} \ .$$

Donc,

$$\mathbb{E}(N_{n+1}(a)) \leq u + 4e\alpha \sum_{t=u}^{n} \log(t\log(t)) \log(t) \exp\left\{-\alpha \log(t\log(t))\right\} \ .$$

Pour tout $\alpha > 1$, la fonction $t \to \log(t\log(t)) \log(t) \exp\left\{-\alpha \log(t\log(t))\right\}$ étant décroissante sur $[4, \infty[$, et en remarquant que $u \geq 4$, on a

$$\begin{aligned}
\mathbb{E}(N_{n+1}(a)) &\leq u + 4e\alpha \int_{u-1}^{\infty} \frac{\log(t\log(t)) \log(t)}{(t\log(t))^{\alpha}} dt \\
&\leq u + 4e\alpha \int_{u-1}^{\infty} \frac{\log(t\log(t))(\log(t)+1)}{(t\log(t))^{\alpha}} dt \ .
\end{aligned}$$

En faisant le changement de variable $s = t\log(t)$, on a $ds = (\log(t)+1)dt$ et

$$\mathbb{E}(N_{n+1}(a)) \leq u + 4e\alpha \int_{(u-1)\log(u-1)}^{\infty} \frac{\log(s)}{s^{\alpha}} ds \ .$$

A l'aide d'une intégration par partie, on montre que

$$\begin{aligned}
\int_{(u-1)\log(u-1)}^{\infty} \frac{\log(s)}{s^{\alpha}} ds &= \frac{\log\left((u-1)\log(u-1)\right)}{(\alpha-1)\left((u-1)\log(u-1)\right)^{\alpha-1}} + \frac{1}{\alpha-1} \int_{(u-1)\log(u-1)}^{\infty} \frac{1}{s^{\alpha}} ds \\
&= \frac{(\alpha-1)\log\left((u-1)\log(u-1)\right)+1}{(\alpha-1)^2\left((u-1)\log(u-1)\right)^{\alpha-1}} \ .
\end{aligned}$$

Par définition de u, on a

$$\frac{(\alpha-1)\log\left((u-1)\log(u-1)\right)+1}{(\alpha-1)^2\left((u-1)\log(u-1)\right)^{\alpha-1}} \leq \frac{\log(\alpha\log(n))}{(\alpha-1)^2(\alpha\log(n))^{\alpha-1}} \ .$$

D'où,

$$\mathbb{E}(N_{n+1}(a)) \leq \frac{32\alpha\log(n\log(n))}{\Delta_a^2} + \frac{4e\alpha\log(\alpha\log(n))}{(\alpha-1)^2(\alpha\log(n))^{\alpha-1}}$$

et

$$\mathbb{E}(\text{Regret}_n) \leq \sum_{a\in\mathcal{A}:\Delta_a>0} \frac{32\alpha\log(n\log(n))}{\Delta_a} + \frac{4e\alpha\Delta_a\log(\alpha\log(n))}{(\alpha-1)^2(\alpha\log(n))^{\alpha-1}} \ .$$

□

4.3.2 Performances pratiques

Nous comparons dans ce paragraphe l'algorithme KL-UCB aux algorithmes UCB et UCB-V (voir algorithmes 1.8 et 1.9)[Auer et al., 2002; Audibert et al., 2007]. Ces algorithmes de référence pour les modèles de bandit ont recours aux voisinages L^1 pour calculer le modèle optimiste. L'algorithme UCB-V utilise de plus une estimation de la variance de la loi des récompenses reçues dans la définition des voisinages ce qui permet d'obtenir des regrets plus faibles. L'utilisation de la divergence de Kullback-Leibler pour déterminer les voisinages de confiance est analogue à l'idée de [Audibert et al., 2007] de tirer profit de la variance estimée. En effet, le point commun à ces deux approches est d'adapter le rayon des voisinages autour de la récompense estimée selon que celle-ci est proche ou éloignée d'un bord de l'intervalle $[0,1]$: plus la récompense estimée est proche de 0 ou de 1, plus le voisinage de confiance est petit. La figure 4.1 illustre cela en considérant deux bras respectivement de moyenne $r(a) = 0.6$ (gauche) et $r(a) = 0.95$ (droite). Pour des valeurs de $N_t(a)$ variant de 100 à 10000, on représente

- la borne supérieure en utilisant un voisinage L^1 : $r(a) + \sqrt{\frac{\log(n)}{2N_t(a)}}$,
- la borne supérieure en utilisant un voisinage L^1 et en tenant compte de la variance : $r(a) + \sqrt{\frac{2\log(n)W(a)}{N_t(a)}} + 3\frac{\log(n)}{N_t(a)}$ où $W(a) = r(a)(1-r(a))$,
- et la borne supérieure en utilisant un voisinage KL :

$$\max\{r,\ kl(r(a),r) \leq \log(n\log(n))/N_t(a)\}\ .$$

On a fixé $n = 10000$ et pris la valeur limite de α (qui vaut $1/2$ pour l'algorithme UCB et 1 pour les algorithmes UCB-V et KL-UCB). On observe que la borne supérieure est toujours plus petite lorsque l'on utilise un voisinage de KL et que celle-ci ne dépasse jamais 1. Selon la valeur de $r(a)$ l'utilisation de la variance est plus ou moins pertinente. En effet, plus la variance est faible, plus le diamètre des intervalles de confiance qui tiennent compte de celle-ci est faible devant le diamètre des voisinages L^1 standard.

La borne de Bernstein, qui a donné lieu à la borne supérieure en utilisant un voisinage L^1 en tenant compte de la variance, est plus fine que la borne d'Hoeffding dès que le terme $3\frac{\log(n)}{N_t(a)}$ est petit devant $\sqrt{\frac{2\log(n)W(a)}{N_t(a)}}$, ce qui a lieu pour $N_t(a)$ grand. Mais la borne supérieure de l'intervalle de confiance en utilisant l'inégalité de Bernstein peut être significativement trop grande lorsque $N_t(a)$ est plus petit ou du même ordre de grandeur que $\log(n)$. On observe en effet sur la figure 4.1 que, lorsque $N_t(a)$ est petit, la borne supérieure du voisinage L^1 avec variance est beaucoup plus grande que celle des autres voisinages. Or, pour tous les bras sous-optimaux, le terme $\frac{\log(n)}{N_t(a)}$ ne tend pas vers 0. Ceci explique probablement la suite des résultats.

Procédons maintenant à des simulations numériques pour comparer les performances de ces trois algorithmes en pratique. Le paramètre α des algorithmes UCB, UCB-V et KL-UCB est respectivement fixé à 0.501, 1.01 et 1.01. Dans un premier temps, nous considérons un modèle de bandit avec 5 bras. La récompense reçue en jouant chacun des bras est une variable de Bernoulli de paramètre égal à 0.5, 0.6, 0.7, 0.8 et 0.9. On représente sur la figure 4.2 la moyenne du regret cumulé calculée au cours de 10 réplications Monte-Carlo en suivant chacun des trois algorithmes. Le regret de l'algorithme KL-UCB est plus faible que les autres. Dans ce cas, le regret de l'algorithme UCB-V est plus grand que celui de UCB. La figure 4.3 représente le regret cumulé en suivant les trois mêmes algorithmes dans un modèle de bandit où les récompenses moyennes sont très proches de 1. Les récompenses moyennes des 5 bras considérés sont comprises entre 0.995 et 0.999 avec un écart de 0.001 entre chacune. Il est donc plus difficile dans ce cas de distinguer le meilleur bras des autres bons bras et l'algorithme UCB met beaucoup plus de temps que les deux autres à le trouver. On observe que la différence

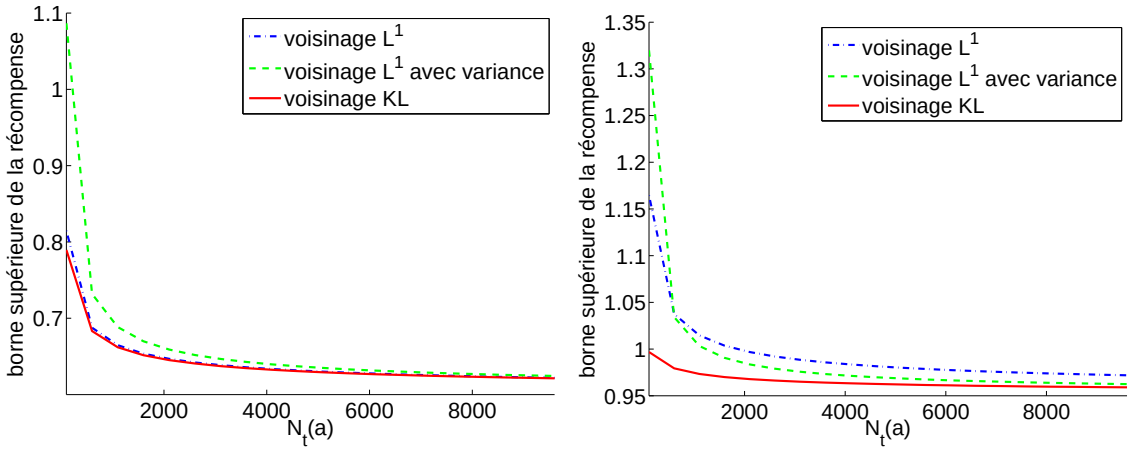

FIGURE 4.1 – Bornes supérieures de l'espérance de la récompense en utilisant différents voisinages pour $r(a) = 0.6$ (gauche) et $r(a) = 0.95$ (droite).

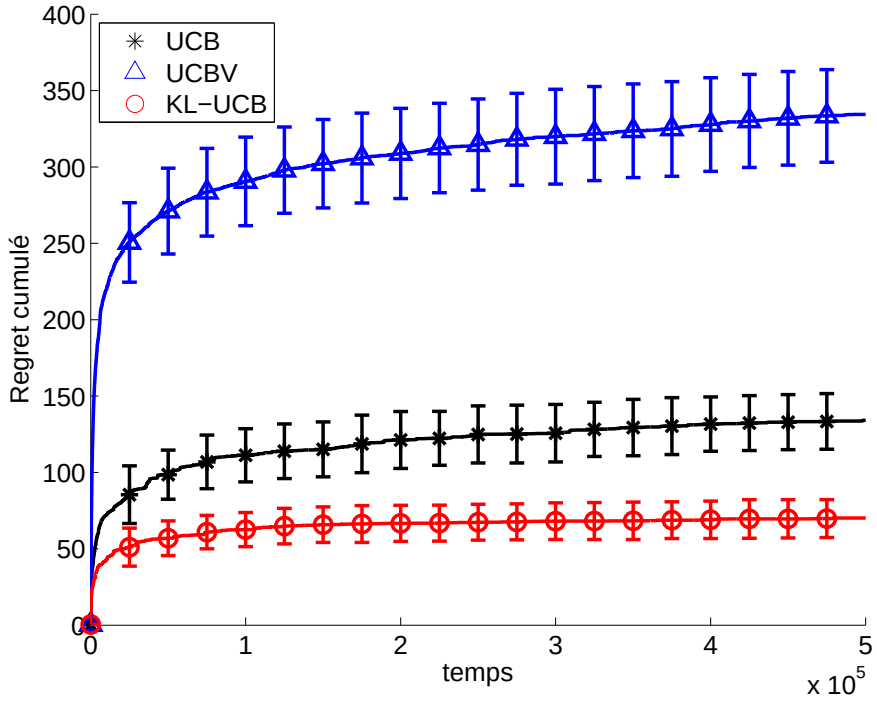

FIGURE 4.2 – Regret en suivant les algorithmes KL-UCB, UCB et UCBV pour un modèle de bandit à 5 bras de moyennes $\{0.5, 0.6, 0.7, 0.8, 0.9\}$.

entre le regret atteint par l'algorithme KL-UCB et celui de l'algorithme UCB-V est du même ordre de grandeur dans les deux exemples.

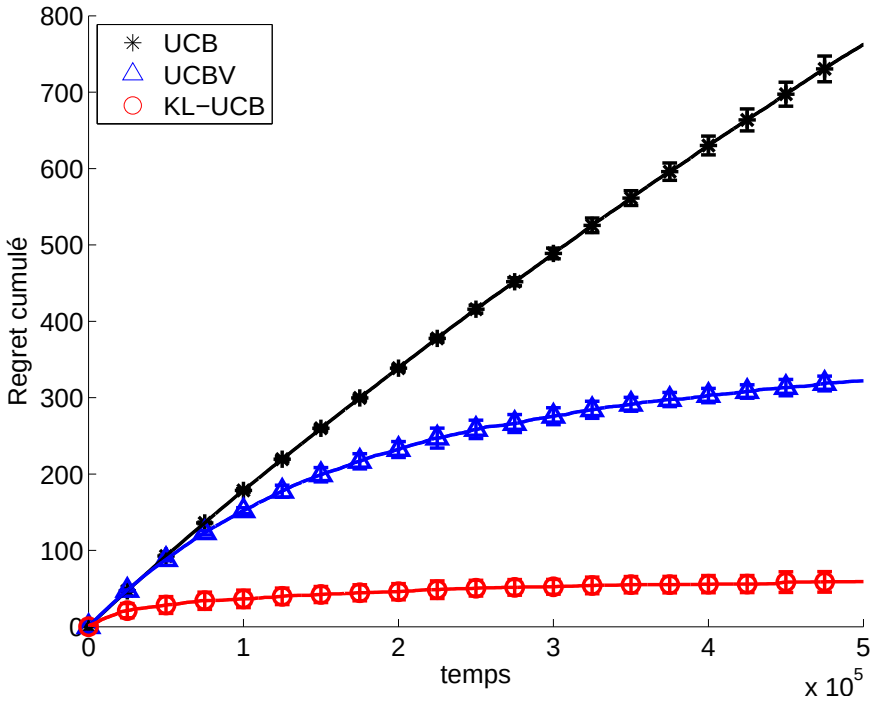

FIGURE 4.3 – Regret en suivant les algorithmes KL-UCB, UCB et UCBV pour un modèle de bandit à 5 bras de moyennes $\{0.995, 0.996, 0.997, 0.998, 0.999\}$.

4.4 L'algorithme KL-UCRL

Après avoir proposé un algorithme optimiste pour des modèles de bandit à récompenses binaires, nous nous intéressons maintenant aux processus de décisions markoviens à espaces d'états et d'actions finis. Nous fournissons un deuxième algorithme optimiste basé sur l'utilisation de la divergence de Kullback-Leibler pour calculer le modèle optimiste. Cet algorithme, appelé KL-UCRL et décrit en détail ci-dessous, est une variante de l'algorithme UCRL2 introduit par [Auer et al., 2009a] après une première version de l'algorithme UCRL présenté par [Auer and Ortner, 2007] et revu par [Bartlett and Tewari, 2009]. La différence entre l'algorithme KL-UCRL et l'algorithme UCRL2 réside dans la construction du modèle optimiste.

L'algorithme KL-UCRL procède en épisodes. Durant chaque épisode, une seule politique est suivie, ce qui permet de jouer chaque politique choisie assez longtemps afin de pouvoir évaluer sa performance moyenne sur différents couples état-action. La longueur de chaque épisode dépend du nombre de fois où chaque couple état-action a été visité pendant cet épisode. Plus précisément un épisode s'arrête dès que le nombre de visites d'au moins un couple état-action a été doublé pendant l'épisode. Soit t_j l'instant du début de l'épisode j et $\tilde{N}_j(x,a)$ le nombre de visites au couple (x,a) pendant l'épisode j. Rappelons que $N_{t_j}(x,a)$ est le nombre de visites au couple (x,a) avant l'instant t_j (c'est-à-dire avant le début de l'épisode j). Pour tout (x,a), $\sum_{j=1}^{k} \tilde{N}_j(x,a) = N_{t_{k+1}}(x,a)$. On peut écrire la condition d'arrêt d'un épisode en utilisant ces notations : l'épisode j se termine dès que $\tilde{N}_j(x,a) \geq N_{t_j}(x,a)$ pour au moins un couple (x,a).

Le MDP dit optimiste et noté $\mathbf{M}_j$ est calculé au début de chaque épisode j, c'est-à-dire à l'instant t_j. Il s'agit du MDP appartenant à l'ensemble $\mathcal{M}_{t_j}$ sous lequel la récompense

Algorithme 4.2 KL-UCRL

Initialisation : $j = 0$; $\forall a \in \mathcal{A}, \forall x \in \mathcal{X}, \tilde{N}_j(x,a) = 0$, $N_0(x,a) = 0$; politique initiale π_0.
Pour tout $t \geq 0$ **faire**
  Observation de X_t.
  si $\tilde{N}_j(X_t, \pi_j(X_t)) \geq \max(N_{t_j}(X_t, \pi_j(X_t)), 1)$ **alors**
    Début d'un nouvel épisode : $j = j+1$, $t_j = t$,
    Re-initialisation : $\forall a \in \mathcal{A}, \forall x \in \mathcal{X}$, $\tilde{N}_j(x,a) = 0$
    Estimation de $\hat{P}_t$ et de $\hat{r}_t$ en utilisant les équations (4.1).
    Recherche du modèle optimiste $\mathbf{M}_j \in \mathcal{M}_t$ et de la politique associée π_j en résolvant l'équation (4.13) et en utilisant l'algorithme 4.5.
  fin si
  Choix de l'action $A_t = \pi_j(X_t)$
  Réception de la récompense R_t.
fin Pour

moyenne optimale est la plus grande :

$$\mathbf{M}_j \in \operatorname*{argmax}_{\mathbf{M} \in \mathcal{M}_{t_j}} \eta^*(\mathbf{M}) \ .$$

Rappelons que $\mathcal{M}_t$ est l'ensemble des MDP vraisemblables à l'instant t défini dans l'équation (4.7). Notons $\eta^*(\mathbf{M}_j)$ la récompense moyenne optimale reçue dans le modèle $\mathbf{M}_j$ et $h^*(\mathbf{M}_j)$ un vecteur de biais associé. Le réel $\eta^*(\mathbf{M}_j)$ et le vecteur $h^*(\mathbf{M}_j)$ vérifient le *système d'équations d'optimalité étendu* suivant : pour tout état $x \in \mathcal{X}$

$$h^*(\mathbf{M}_j, x) + \eta^*(\mathbf{M}_j)\mathbf{e} = \max_{P,r} \max_{a \in \mathcal{A}} \left(r(x,a) + \sum_{x' \in \mathcal{X}} P(x,a;x') h^*(\mathbf{M}_j, x') \right) \tag{4.13}$$

où la maximisation est faite sur l'ensemble des P, r tels que

$$\forall x, \forall a, \quad KL(\hat{P}_{t_j}(x,a;.), P(x,a;.)) \leq \frac{C_P}{N_{t_j}(x,a)} \ , \tag{4.14}$$

$$\forall x, \forall a, \quad |\hat{r}_{t_j}(x,a) - r(x,a)| \leq \frac{C_R}{\sqrt{N_{t_j}(x,a)}} \ , \tag{4.15}$$

où C_P et C_R sont des constantes qui contrôlent la taille des voisinages de confiance. Le vecteur de biais étant défini à une constante près, ce système d'équation n'est rien d'autre que l'équation du point fixe de l'opérateur suivant

$$\mathcal{L}_{ext,\mathcal{M}_{t_j}} : V \mapsto \max_{(\mathcal{X},\mathcal{A},P,r) \in \mathcal{M}_{t_j}} \max_{a \in \mathcal{A}} \left(r(.,a) + \sum_{x' \in \mathcal{X}} P(.,a;x') V(x') \right) \ . \tag{4.16}$$

Le système d'équations d'optimalité étendu a été étudié par [Nilim and El Ghaoui, 2005; Tewari and Bartlett, 2007; Auer et al., 2009b] pour différentes définitions des voisinages autour des probabilités de transition et des récompenses estimées. Ces études montrent qu'il existe un MDP $\mathbf{M}_j = (\mathcal{X}, \mathcal{A}, P_j, r_j) \in \mathcal{M}_{t_j}$ tel que, pour tout état x

$$\begin{aligned} V(x) &= \max_{(\mathcal{X},\mathcal{A},P,r) \in \mathcal{M}_{t_j}} \max_{a \in \mathcal{A}} \left(r(x,a) + \sum_{x' \in \mathcal{X}} P(x,a;x') V(x') \right) \\ &= \max_{a \in \mathcal{A}} \left(r_j(x,a) + \sum_{x' \in \mathcal{X}} P_j(x,a;x') V(x') \right) \ . \end{aligned}$$

Cela prouve donc l'existence du MDP optimiste.

Le calcul du modèle optimiste est décrit plus précisément dans le paragraphe suivant. En particulier, nous explicitons la maximisation d'une fonction linéaire sur un voisinage KL. La politique π_j est déduite de $h^*(\mathbf{M}_j)$ de la façon suivante : pour tout état $x \in \mathcal{X}$,

$$\pi^*(\mathbf{M}_j, x) \in \operatorname*{argmax}_{a \in \mathcal{A}} \left(r_j(x,a) + \sum_{x' \in \mathcal{X}} P_j(x,a;x') h^*(\mathbf{M}_j, x') \right) .$$

Cette politique est suivie tout au long de l'épisode j.

4.4.1 Recherche du modèle optimiste

Comme mentionné précédemment, le modèle optimiste est déterminé en résolvant l'équation du point fixe de l'opérateur $\mathcal{L}_{ext,\mathcal{M}}$. En pratique, l'algorithme d'*itération sur les valeurs étendu* peut être utilisé pour résoudre cette équation du point fixe. Il s'agit d'un algorithme similaire à l'algorithme d'itération sur les valeurs (voir algorithme 1.3). Cependant, contrairement à l'opérateur de Bellman $\mathcal{L}$, la maximisation dans l'opérateur $\mathcal{L}_{ext}$ ne se fait pas uniquement sur l'espace des actions mais aussi sur un ensemble de noyaux de transitions P et d'espérance de récompenses r. L'algorithme est décrit ci-dessous. [Auer et al., 2009b] ont

Algorithme 4.3 Algorithme d'itération sur les valeurs étendu

1: Pour tout $x \in \mathcal{X}$, $V_0(x) = 0$
2: **Pour tout** pour tout $n \geq 0$
3: On calcule $V_{n+1} = \mathcal{L}_{ext,\mathcal{M}} V_n$
4: **jusqu'à ce que**

$$osc(V_{n+1} - V_n) > \epsilon$$

prouvé que cet algorithme converge vers un vecteur V satisfaisant l'équation du point fixe de l'opérateur $\mathcal{L}_{ext,\mathcal{M}_{t_j}}$ dans le cas où les intervalles de confiance autour des probabilités de transition sont définis en utilisant une distance L^1. Il semble que l'algorithme converge également lorsque des voisinages KL sont utilisés mais cela n'a pas encore été prouvé. [Tewari and Bartlett, 2007] ont montré la convergence d'un autre algorithme permettant également de déterminer le modèle optimiste, pour une plus grande classe de modèles, appelés des *Processus de Décisions Markoviens à paramètres bornés*, incluant le cas où un voisinage KL est considéré. Ce deuxième algorithme, décrit algorithme 4.4, utilise le fait que la fonction de valeur optimale est également Blackwell optimale (voir section 1.2.2). Nous utiliserons néanmoins l'algorithme 4.3 dans la suite.

[Auer et al., 2009b; Bartlett and Tewari, 2009] ont prouvé que l'oscillation du vecteur V obtenu à chaque itération de l'algorithme 4.3, et donc également celle du vecteur de biais $h^*(\mathbf{M}_j)$, sont majorés par une constante $D(\mathbf{M})$ définie comme suit

$$D(\mathbf{M}) \stackrel{\text{def}}{=} \max_{x,x'} \min_{\pi} \mathbb{E}_{\mathbf{M},\pi}(\tau(x,x')) \tag{4.17}$$

où $\tau(x,x')$ désigne le temps d'atteinte de l'état x' partant de l'état x. Cette constante dépendant du modèle apparaît dans les bornes de regret et caractérise le temps minimum nécessaire pour atteindre tout état à partir de n'importe quel autre état sous une politique favorable. Pour tout MDP communiquant $\mathbf{M}$, $D(\mathbf{M})$ est fini.

A chaque étape de ces algorithmes, les problème de maximisation suivants doivent être résolus pour tout $x \in \mathcal{X}$ et tout $a \in \mathcal{A}$

$$\max_r r(x,a) \qquad \text{et} \qquad \max_P \sum_{x'} P(x,a;x) V(x')$$

Algorithme 4.4 Algorithme d'itération sur les valeurs pour des MDPs à paramètres bornés

1: Pour tout $x \in \mathcal{X}$, $V_0(x) = 0$
2: **Pour** tout $n \geq 0$ **faire**
3: $\alpha_n = \frac{n+1}{n+2}$
4: On calcule, pour tout $x \in \mathcal{X}$,

$$V_{n+1}(x) = \max_a \left\{ (1-\alpha_n) \max_r r(x,a) + \alpha_n \max_P \sum_{x'} P(x,a;x') V_n(x') \right\}$$

5: **fin Pour**

où r et P appartiennent à l'ensemble $\mathcal{M}_t$. Pour chaque état $x \in \mathcal{X}$ et chaque action $a \in \mathcal{A}$, la maximisation du terme $r(x,a)$ dans le voisinage L^1 autour de $\hat{r}_{t_j}(x,a)$ est obtenue de manière évidente en prenant

$$r_j(x,a) = \hat{r}_{t_j}(x,a) + \frac{C_R}{\sqrt{N_{t_j}(x,a)}} .$$

La principale difficulté réside donc dans la maximisation du produit scalaire entre le vecteur de probabilité $q = P(x,a;.)$ et le *vecteur de valeur* V sur le voisinage défini par la métrique KL autour du vecteur de probabilité connu $p = \hat{P}_{t_j}(x,a;.)$:

$$\max_{q \in \mathbb{S}^{|\mathcal{X}|}} V'q \quad \text{tel que} \quad KL(p,q) \leq \epsilon . \tag{4.18}$$

On rappelle que $V'q = \sum_{i=1}^N V_i q_i$ et que le rayon du voisinage $\epsilon = C_P / N_{t_j}(x,a)$ contrôle la taille du voisinage.

Optimisation linéaire sur un voisinage KL

Ce paragraphe est consacré à la résolution du problème de maximisation sous contrainte défini par l'équation (4.18). Dans [Nilim and El Ghaoui, 2005], un problème similaire est considéré dans un autre contexte et une solution quelque peu différente a été proposée pour le cas où toutes les composantes p_i sont strictement positives. Puisque il s'agit d'un problème de maximisation d'une fonction linéaire sous des contraintes convexes (voir [Boyd and Vandenberghe, 2004]), il suffit d'étudier le Lagrangien associé à cette maximisation défini par

$$\begin{aligned} &L(q, \lambda, \nu, \mu_1, \ldots, \mu_{|\mathcal{X}|}) \\ &\qquad = \sum_{i=1}^{|\mathcal{X}|} q_i V_i - \lambda \left(\sum_{i=1}^{|\mathcal{X}|} p_i \log \frac{p_i}{q_i} - \epsilon \right) - \nu \left(\sum_{i=1}^{|\mathcal{X}|} q_i - 1 \right) + \sum_{i=1}^{|\mathcal{X}|} \mu_i q_i . \end{aligned} \tag{4.19}$$

Si q maximise cette fonction alors il existe $\lambda \in \mathbb{R}$, $\nu \geq 0$, $\mu_i \geq 0$ $(i = 1 \ldots N)$ tels que les conditions suivantes sont simultanément satisfaites

$$\begin{cases} V_i + \lambda \dfrac{p_i}{q_i} - \nu + \mu_i = 0 & (4.20) \\ \lambda \left(\displaystyle\sum_{i=1}^{|\mathcal{X}|} p_i \log \frac{p_i}{q_i} - \epsilon \right) = 0 & (4.21) \\ \nu \left(\displaystyle\sum_{i=1}^{|\mathcal{X}|} q_i - 1 \right) = 0 & (4.22) \\ \mu_i q_i = 0 & (4.23) \end{cases}$$

Soit $Z = \{i, p_i = 0\}$. Les conditions (4.20) à (4.23) entraînent que $\lambda \neq 0$ et $\nu \neq 0$. Pour $i \in \bar{Z}$, l'équation (4.20) implique que $q_i = \lambda \frac{p_i}{\nu - \mu_i - V_i}$. Puisque $\lambda \neq 0$, $q_i > 0$ et ainsi, d'après (4.23), $\mu_i = 0$. Donc

$$\forall i \in \bar{Z} \ , \quad q_i = \lambda \frac{p_i}{\nu - V_i} \ . \tag{4.24}$$

Soit $s = \sum_{i \in Z} q_i$. En sommant sur les $i \in \bar{Z}$ et en utilisant les équations (4.24) et (4.22), on a

$$\lambda \sum_{i \in \bar{Z}} \frac{p_i}{\nu - V_i} = \sum_{i \in \bar{Z}} q_i = 1 - s \ . \tag{4.25}$$

En utilisant les équations (4.24) et (4.25), on peut écrire

$$\sum_{i \in \bar{Z}} p_i \log \frac{p_i}{q_i} = f(\nu) - \log(1 - s)$$

où f est définie pour tout $\nu \geq \max_{i \in \bar{Z}} V_i$ par

$$f(\nu) = \sum_{i \in \bar{Z}} p_i \log(\nu - V_i) + \log \left(\sum_{i \in \bar{Z}} \frac{p_i}{\nu - V_i} \right) \ . \tag{4.26}$$

Alors, q satisfait la condition (4.21) si et seulement si

$$f(\nu) = \epsilon + \log(1 - s) \ . \tag{4.27}$$

Considérons maintenant le cas où $i \in Z$. Soit $I^* = Z \cap \operatorname{argmax}_i V_i$. Pour tout $i \in Z \setminus I^*$, $q_i = 0$. En effet, sinon, μ_i devrait être égal à 0 et alors $\nu = V_i$ d'après l'équation (4.20), ce qui conduit à un dénominateur négatif dans l'équation (4.24). D'après l'équation (4.23), pour tout $i \in I^*$, soit $q_i = 0$, soit $\mu_i = 0$. Dans le second cas $\nu = V_i$ et $s > 0$ ce qui nécessite que $f(\nu) < \epsilon$ pour que (4.27) puisse être satisfaite avec $s > 0$. Donc,

- si $f(V_i) < \epsilon$ pour $i \in I^*$, alors $\nu = V_i$ et la constante s peut être calculée en résolvant l'équation $f(\nu) = \epsilon - \log(1 - s)$; les valeurs de q_i pour $i \in I^*$ peuvent être choisies de n'importe quelles façon tant que $\sum_{i \in I^*} q_i = s$;
- si pour tout $i \in I^*$ $f(V_i) \geq \epsilon$, alors $s = 0$, $q_i = 0$ pour tout $i \in Z$ et ν est la solution de l'équation $f(\nu) = \epsilon$.

Après avoir déterminé ν et s, les autres composantes de q sont calculées en utilisant l'équation (4.24) : on a pour $i \in \bar{Z}$,

$$q_i = \frac{(1 - s)\tilde{q}_i}{\sum_{i \in \bar{Z}} \tilde{q}_i}$$

où

$$\tilde{q}_i = \frac{p_i}{\nu - V_i} \ .$$

L'algorithme 4.5 ci-dessous résume les étapes à suivre pour maximiser une fonction linéaire $V'q$ sur un voisinage KL de la forme $KL(p; q) \leq \epsilon$.

Rappelons que le vecteur de probabilité $q = P(x, a; .)$ est le vecteur de transition à partir d'un couple état-action fixé (x, a) vers différents états et que V est le *vecteur de valeur* associé à ces états. Dans le cas particulier où, à l'instant t, l'état le plus prometteur, noté $x_M \stackrel{\text{def}}{=} \operatorname{argmax}_x V_x$, n'a jamais été atteint, c'est-à-dire si $\hat{P}_t(x, a; x_M) = 0$, l'algorithme permet de gérer un compromis entre autoriser une transition vers cet état avec probabilité positive $q_{x_M} > 0$ ou réaliser que cette transition est impossible et donc augmenter les probabilités de transition vers d'autres états qui sont atteignables et dont la valeur est également attractive. Ce compromis est fait en comparant la valeur relative de l'état le plus prometteur V_{x_M} avec les preuves statistiques accumulées concernant son accessibilité.

Algorithme 4.5 Maximisation d'une fonction linéaire $V'q$ sur un voisinage KL

Entrées: Une fonction de valeur V, un vecteur de probabilité p, une constante ϵ

1: $I^* = \{i : \ p_i = 0, \ V_i \geq V_j \ \forall j\}$
2: **si** $I^* \neq \emptyset$ et s'il existe $j \in I^*$ tel que $f(V_j) < \epsilon$ **alors**
3: Pour tout $i \neq j$ tel que $p_i = 0$, $q_i = 0$.
4: $s = 1 - \exp\{f(V_j) - \epsilon\}$ et $\nu = V_j$
5: $q_j = s$
6: **sinon**
7: Pour tout i tel que $p_i = 0$, $q_i = 0$ et $s = 0$.
8: Trouver ν tel que $f(\nu) = \epsilon$ en utilisant la méthode de Newton
9: **fin si**
10: Pour tout i tel que $p_i > 0$, $q_i = \frac{(1-s)\tilde{q}_i}{\sum_{i \in \bar{Z}} \tilde{q}_i}$ où $\tilde{q}_i = \frac{p_i}{\nu - V_i}$.

Propriétés de la fonction f

La fonction f définie par l'équation (4.26) joue un rôle clé dans la procédure de maximisation présentée ci-dessus. Nous analysons maintenant quelques propriétés de cette fonction.

Proposition 4.2. *La fonction f est une fonction convexe décroissante de $]\max_{i \in \bar{Z}} V_i; \infty[$ dans $]0; \infty[$.*

Démonstration. En utilisant l'inégalité de Jensen, il est facile de montrer que la fonction f décroît de $+\infty$ dans 0. La dérivée seconde de f par rapport à ν est égale à

$$-\sum_i \frac{p_i}{(\nu - V_i)^2} + \frac{2\sum_i \frac{p_i}{(\nu-V_i)^3} \sum_i \frac{p_i}{\nu - V_i} - \left(\sum_i \frac{p_i}{(\nu-V_i)^2}\right)^2}{\left(\sum_i \frac{p_i}{\nu - V_i}\right)^2}.$$

Soit Z une variable aléatoire positive telle que $\mathbb{P}\left(Z = \frac{1}{\nu - V_i}\right) = p_i$. On a

$$f''(\nu) = \frac{2\mathbb{E}(Z^3)\mathbb{E}(Z) - \mathbb{E}(Z^2)\mathbb{E}(Z)^2 - \mathbb{E}(Z^2)^2}{\mathbb{E}(Z)^2} \ .$$

D'après l'inégalité de Cauchy-Schwarz, on a $\mathbb{E}(Z^2)^2 = \mathbb{E}(Z^{3/2}Z^{1/2})^2 \leq \mathbb{E}(Z^3)\mathbb{E}(Z)$. De plus $\mathbb{E}(Z^2)^2 \geq \mathbb{E}(Z^2)\mathbb{E}(Z)^2$. Ces deux inégalités montrent que $f''(\nu) \geq 0$ et donc que f est convexe. □

Comme mentionné précédemment, l'algorithme de Newton (voir [Boyd and Vandenberghe, 2004]) peut être utilisé pour résoudre l'équation $f(\nu) = \epsilon$ pour une valeur fixée de ϵ (dans l'étape 9 de l'algorithme), de manière à ce que la résolution de (4.18) ne soit qu'une question de quelques itérations. Quand ϵ est très proche de 0, la solution de l'équation est assez grande et une initialisation appropriée de l'algorithme permet d'accélérer de manière significative sa convergence. En utilisant un développement de Taylor du second ordre de la fonction f, on montre que, pour ν suffisamment grand,

$$f(\nu) = \frac{\sigma_{p,V}}{2\nu^2} + o(\frac{1}{\nu^2}) \ ,$$

où $\sigma_{p,V} = \sum_i p_i V_i^2 - (\sum_i p_i V_i)^2$. L'algorithme de Newton peut alors être initialisé en prenant $\nu_0 = \sqrt{\sigma_{p,V}/(2\epsilon)}$.

4.4.2 Résultats théoriques

Bornes de regret

Pour analyser les performances de l'algorithme KL-UCRL, nous comparons les récompenses accumulées lors de son utilisation avec celles qui seraient reçues, en moyenne, par un agent jouant une politique optimale. Le *regret* de l'algorithme après n instants est défini comme dans [Jaksch et al., 2010] :

$$\text{Regret}_n = \sum_{t=0}^{n} \left(\eta^*(\mathbf{M}) - R_t\right) \ . \tag{4.28}$$

L'analyse de la borne de regret que nous proposons est une adaptation de celle de [Jaksch et al., 2010] à l'utilisation d'un voisinage KL. Dans la suite du paragraphe, on suppose que la récompense $R_{\max}$ vaut 1 pour simplifier l'écriture des théorèmes ainsi que des démonstrations. Le théorème suivant établit une borne supérieure du regret de l'algorithme KL-UCRL lorsque C_P et C_R sont définis par

$$C_P = |\mathcal{X}| \left(B + \log\left(B + \frac{1}{\log(n)}\right) \left[1 + \frac{1}{B + \frac{1}{\log(n)}}\right]\right) \quad \text{où} \quad B = \log\left(\frac{2e|\mathcal{X}|^2|\mathcal{A}|\log(n)}{\delta}\right)$$

et

$$C_R = \sqrt{\frac{\log\left(4|\mathcal{X}||\mathcal{A}|\log(n)/\delta\right)}{1.99}} \ .$$

Théorème 4.4. *Avec probabilité $1-4\delta$, pour tout horizon $n > |\mathcal{A}|\log_2(8|\mathcal{A}|)^2$, le regret de KL-UCRL est majoré par*

$$\textit{Regret}_n \leq CD(\mathbf{M})|\mathcal{X}|\sqrt{|\mathcal{A}|n\log(\log(n)/\delta)} \ ,$$

pour une constante $C \leq 24$ indépendante du modèle.

Il est également possible de garantir un regret espéré logarithmique en l'horizon n. Cette majoration, présentée dans le théorème 4.5, fait apparaître la différence entre les politiques optimales et sous-optimales :

$$\Delta(\mathbf{M}) = \eta^*(\mathbf{M}) - \max_{\pi:\eta^\pi(\mathbf{M})<\eta^*(\mathbf{M})} \eta^\pi(\mathbf{M}) \ .$$

Théorème 4.5. *Pour tout horizon $n > |\mathcal{A}|\log_2(8|\mathcal{A}|)^2$, le regret espéré de KL-UCRL est majoré par*

$$\mathbb{E}(\textit{Regret}_n) \leq CD^2(\mathbf{M})\frac{|\mathcal{X}|^2|\mathcal{A}|\log(n)}{\Delta(\mathbf{M})} + C(\mathbf{M}) \ ,$$

où $C \leq 23^2$ est une constante indépendante du modèle, et $C(\mathbf{M})$ est une constante dépendant du modèle.

La constante $C(\mathbf{M})$ est liée à la perte pendant les épisodes trop courts pour que la récompense accumulée durant ceux-ci puisse être confondue avec la récompense moyenne de la politique jouée.

Nous avons remarqué dans le paragraphe 4.2.3 que des inégalités de concentration plus fines peuvent être obtenues autour des probabilités de transition si le vrai MDP est de faible connectivité et que l'agent connaît la structure du modèle (c'est-à-dire s'il sait quelles sont les probabilités de transitions égales à 0). Le théorème suivant présente des bornes supérieures du regret lorsque le MDP est de faible connectivité et que les récompenses sont déterministes.

Les récompenses étant déterministes, l'algorithme ne construit pas d'intervalle de confiance autour des récompenses espérées. Cela permet de se focaliser sur les inégalités de concentration autour des probabilités de transition. Notons que ce sont sous ces conditions que les premiers résultats en terme de regret ont été fournis pour des MDPs [Burnetas and Katehakis, 1997].

Théorème 4.6. *Si le MDP* $\mathbf{M}$ *est tel que les récompenses reçues sont une fonction déterministe de l'état courant et de l'action sélectionnée. Notons* $\kappa(\mathbf{M})$ *le degré de connectivité de* $\mathbf{M}$ *défini par l'équation* (4.6)*. Pour tout horizon* $n > |\mathcal{A}| \log_2(8|\mathcal{A}|)^2$*, le regret en suivant l'algorithme KL-UCRL avec*

$$C_P = \kappa \left(B + \log\left(B + \frac{1}{\log(n)} \right) \left[1 + \frac{1}{B + \frac{1}{\log(n)}} \right] \right) \quad \text{où} \quad B = \log\left(\frac{2e|\mathcal{X}|\kappa|\mathcal{A}| \log(n)}{\delta} \right)$$

est majoré par

$$Regret_n \leq CD(\mathbf{M}) \sqrt{\kappa(\mathbf{M})|\mathcal{X}||\mathcal{A}| n \log(\log(n)/\delta)} \quad \text{avec probabilité } 1 - 4\delta \ ,$$

où C *est une constante indépendante du modèle. De plus, pour tout horizon* $n > 5$*, le regret espéré est majoré par*

$$\mathbb{E}(Regret_n) \leq CD^2(\mathbf{M}) \frac{|\mathcal{X}|\kappa(\mathbf{M})|\mathcal{A}| \log(n)}{\Delta(\mathbf{M})} + C(\mathbf{M}) \ ,$$

où C *est une constante indépendante du modèle, et* $C(\mathbf{M})$ *est une constante dépendant du modèle.*

On remarque que la dépendance en le nombre d'état est de $\kappa(\mathbf{M})|\mathcal{X}|$ au lieu de $|\mathcal{X}|^2$ dans le cas le plus général. La borne du regret est donc beaucoup plus fine si la connectivité du modèle est connue à l'avance et est très faible.

Démonstrations

Comme mentionné précédemment, les preuves de ces théorèmes sont similaires à celles de [Auer et al., 2009a; Jaksch et al., 2010] (cette deuxième référence contient une version détaillée et corrigée des preuves) et de [Bartlett and Tewari, 2009]. Nous commençons par exposer un résultat important qui assure qu'avec grande probabilité le vrai modèle $\mathbf{M} = (\mathcal{X}, \mathcal{A}, P, r)$ appartient à l'ensemble $\mathcal{M}_t$ pour tout instant t.

Proposition 4.3. *Pour tout* n *et pour tout* $\delta > 0$

$$\mathbb{P}\left(\forall t \leq n \ , \ \mathbf{M} \in \mathcal{M}_t \right) \geq 1 - 2\delta \ .$$

Preuve de la proposition 4.3. Pour prouver que, à chaque instant $t \leq n$ et pour chaque couple état-action (x, a), la divergence de Kullback-Leibler entre la probabilité de transition empirique $\hat{P}_t(x, a; .)$ et la vraie probabilité de transition $P(x, a; .)$ est majorée par $C_P / N_t(x, a)$ avec grande probabilité, on utilise le théorème 4.2 avec $\epsilon = C_P$. On a :

$$\mathbb{P}\left(\forall t \leq n \ , \ KL(\hat{P}_t(x, a; .), P(x, a; .)) \leq \frac{C_P}{N_t(x, a)} \right) \geq 1 - \frac{\delta}{|\mathcal{X}||\mathcal{A}|} \ .$$

De plus, d'après le théorème A.1 de [Garivier and Moulines, 2008], on obtient que, pour tout $x \in \mathcal{X}$, $a \in \mathcal{A}$ et $\delta > 0$

$$\mathbb{P}\left(\forall t \leq n \ , \quad |\hat{r}(x, a) - r(x, a)| \leq \sqrt{\frac{\log\left(\frac{4|\mathcal{X}||\mathcal{A}| \log(n)}{\delta} \right)}{1.99 N_t(x, a)}} \right) \geq 1 - \frac{\delta}{|\mathcal{X}||\mathcal{A}|} \ .$$

En sommant sur tous les couples état-action, on obtient le résultat de la proposition. □

Preuve du théorème 4.4. Rappelons que le regret jusqu'à l'instant n est défini par (voir l'équation (4.28))

$$\mathrm{Regret}_n = n\,\eta^*(\mathbf{M}) - \sum_{t=0}^{n} R_t\ .$$

En remarquant que les variables aléatoires $\left(\sum_{(x,a)} \mathbb{1}_{\{X_t=x, A_t=a} r(x,a) - R_t\right)_{t\geq 0}$ sont des incréments de martingales, on peut utiliser l'inégalité d'Hoeffding-Azuma pour prouver que

$$\mathbb{P}\left(\sum_{(x,a)} N_{n+1}(x,a) r(x,a) - \sum_{k=0}^{n} R_t > \sqrt{\frac{\log(1/\delta)n}{2}}\right) \leq \delta\ .$$

Ainsi, avec probabilité $1-\delta$,

$$\begin{aligned}\mathrm{Regret}_n &\leq \sum_{(x,a)} N_{n+1}(x,a)(\eta^*(\mathbf{M}) - r(x,a)) + C_e(n,\delta)\\ &\leq \sum_{k=1}^{m(n)} \sum_{(x,a)} \tilde{N}_k(x,a)(\eta^*(\mathbf{M}) - r(x,a)) + C_e(n,\delta)\end{aligned}$$

où $C_e(n,\delta) = \sqrt{\frac{\log(1/\delta)n}{2}}$ et $m(n)$ est le nombre d'épisodes avant l'instant n. On rappelle que $\tilde{N}_k(x,a)$ est le nombre de visites au couple (x,a) durant l'épisode k. Durant cet épisode, la politique suivie étant π_k, pour tout x et $a \neq \pi_k(x)$, on a $\tilde{N}_k(x,a) = 0$. Soit $\mathrm{RegretEp}_k$ le regret dans l'épisode k défini par

$$\mathrm{RegretEp}_k \stackrel{\mathrm{def}}{=} \sum_{(x,a)} \tilde{N}_k(x,a)(\eta^*(\mathbf{M}) - r(x,a)) = \sum_{x} \tilde{N}_k(x,\pi_k(x))(\eta^*(\mathbf{M}) - r(x,\pi_k(x)))\ .$$

Soit $\mathbf{M}_k = (\mathcal{X}, \mathcal{A}, P_k, r_k)$ le modèle optimiste appartenant à l'ensemble $\mathcal{M}_{t_k}$. Notons $\eta_k = \eta^*(\mathbf{M}_k)$ la récompense moyenne optimale de ce MDP et h_k une fonction de biais associée. L'équation d'optimalité de Bellman pour ce MDP $\mathbf{M}_k$ est

$$h_k + \eta_k\,\mathbf{e} = r_k^{\pi_k} + P_k^{\pi_k} h_k\ , \tag{4.29}$$

où π_k est une politique optimale pour $\mathbf{M}_k$ et le vecteur $\mathbf{e}$ est le vecteur tel que pour tout i $\mathbf{e}(i) = 1$. Le vecteur $r_k^{\pi_k}$ désigne les récompenses moyennes reçues sous le modèle optimiste $\mathbf{M}_k$ sous la politique π_k

$$\forall x \in \mathcal{X}\ ,\quad r_k^{\pi_k}(x) = r_k(x,\pi_k(x))$$

et la matrice $P_k^{\pi_k}$ est la matrice de transition sous le modèle optimiste $\mathbf{M}_k$ et sous la politique π_k. Dans la suite, on notera r^{π_k} (resp. P^{π_k}) le vecteur des récompenses moyennes (resp. la matrice de transition entre les états) sous la politiques π_k sous le vrai MDP. On a, pour tout $x, x' \in \mathcal{X}$,

$$r^{\pi_k}(x) = r(x,\pi_k(x)) \quad \text{et} \quad P^{\pi_k}(x,x') = P(x,\pi_k(x);x')\ .$$

De même, pour tout $x, x' \in \mathcal{X}$, et tout $t \geq 0$

$$\hat{r}_t^{\pi_k}(x) = \hat{r}_t(x,\pi_k(x)) \quad \text{et} \quad \hat{P}_t^{\pi_k}(x,x') = \hat{P}_t(x,\pi_k(x);x')\ .$$

De plus $\tilde{N}_k^{\pi_k}$ désigne le vecteur de compte des visites aux différents couples $(x,\pi_k(x))$ durant l'épisode k. Pour simplifier les écritures suivantes, le vecteur $\tilde{N}_k^{\pi_k}$ est un vecteur ligne.

Si le vrai modèle $\mathbf{M}$ appartient à l'ensemble $\mathcal{M}_{t_j}$, alors, par définition du modèle optimiste, $\eta^*(\mathbf{M}) \leq \eta_k$ et alors

$$\begin{aligned}
\text{RegretEp}_k &\leq \sum_{(x,a)} \tilde{N}_k(x,a)(\eta_k - r(x,a)) \\
&= \sum_{(x,a)} \tilde{N}_k(x,a)(\eta_k - r_k(x,a)) + \sum_{(x,a)} \tilde{N}_k(x,a)(r_k(x,a) - r(x,a)) \\
&= \tilde{N}_k^{\pi_k}(\eta_k \mathbf{e} - r_k^{\pi_k}) + \tilde{N}_k^{\pi_k}(r_k^{\pi_k} - r^{\pi_k}) \\
&= \tilde{N}_k^{\pi_k}(P_k^{\pi_k} - I)h_k + \tilde{N}_k^{\pi_k}(r_k^{\pi_k} - r^{\pi_k}) \\
&= \tilde{N}_k^{\pi_k}(P_k^{\pi_k} - P^{\pi_k})h_k + \tilde{N}_k^{\pi_k}(P^{\pi_k} - I)h_k + \tilde{N}_k^{\pi_k}(r_k^{\pi_k} - r^{\pi_k}) \; .
\end{aligned}$$

La quatrième égalité découle de l'équation (4.29). D'après la proposition 4.3, le vrai modèle appartient à $\mathcal{M}_{t_j}$ avec probabilité $1-\delta$. Ainsi, avec probabilité $1-2\delta$, on a

$$\text{Regret}_n \leq \sum_{k=1}^{m(n)} \left(\tilde{N}_k^{\pi_k}(P_k^{\pi_k} - P^{\pi_k})h_k + \tilde{N}_k^{\pi_k}(P^{\pi_k} - I)h_k + \tilde{N}_k^{\pi_k}(r_k^{\pi_k} - r^{\pi_k}) \right) + C_e(n,\delta) \; .$$

La suite de la preuve consiste à majorer chacun des trois termes de la somme apparaissant dans l'inégalité précédente. On supposera à chaque fois que le modèle $\mathbf{M} \in \mathcal{M}_{t_k}$ pour tout $x \in \mathcal{X}$, et pour tout $a \in \mathcal{A}$,

1. Pour majorer le premier terme, nous commençons par soustraire le vecteur constant $\min_x(h_k(x))\mathbf{e}$

$$\begin{aligned}
&\tilde{N}_k^{\pi_k}(P_k^{\pi_k} - P^{\pi_k})h_k \\
&\qquad = \tilde{N}_k^{\pi_k}(P_k^{\pi_k} - P^{\pi_k})(h_k - \min_x(h_k(x))\mathbf{e}) \\
&\qquad \leq \sum_x \tilde{N}_k(x,\pi_k(x)) \| P_k(x,\pi_k(x);.) - P(x,\pi_k(x);.) \|_1 \left\| h_k - \min_x(h_k(x))\mathbf{e} \right\|_\infty
\end{aligned}$$

   Comme mentionné dans le paragraphe 4.4, [Auer et al., 2009b] ont montré que l'oscillation du vecteur h_k est majorée par $D(\mathbf{M})$ ce qui implique que $\|h_k - \min_x(h_k(x))\mathbf{e}\|_\infty \leq D(\mathbf{M})$. D'où

$$\begin{aligned}
&\tilde{N}_k^{\pi_k}(P_k^{\pi_k} - P^{\pi_k})h_k \qquad (4.30) \\
&\leq D(\mathbf{M}) \sum_x \tilde{N}_k(x,\pi_k(x)) \left(\left\| \hat{P}_{t_k}^{\pi_k}(x,.) - P^{\pi_k}(x,.) \right\|_1 + \left\| \hat{P}_{t_k}^{\pi_k}(x,.) - P_k^{\pi_k}(x,.) \right\|_1 \right) \\
&\leq D(\mathbf{M}) \sum_x \tilde{N}_k(x,\pi_k(x)) \left\{ \sqrt{2KL(\hat{P}_{t_k}^{\pi_k}(x,.); P^{\pi_k}(x,.))} + \sqrt{2KL(\hat{P}_{t_k}^{\pi_k}(x,.); P_k^{\pi_k}(x,.))} \right\} \\
&\leq 2D(\mathbf{M})\sqrt{2} \sum_x \tilde{N}_k(x,\pi_k(x)) \sqrt{\frac{C_P}{N_{t_k}(x,\pi_k(x))}} \; . \qquad (4.31)
\end{aligned}$$

2. Majorons le deuxième terme. Soit $\mathbf{e}_x$ le vecteur tel que $\mathbf{e}_x(x) = 1$ et, pour tout $x' \neq x$, $\mathbf{e}_x(x') = 0$. On a

$$\begin{aligned}
\tilde{N}_k^{\pi_k}(P^{\pi_k} - I)h_k &= \sum_{t=t_k}^{t_{k+1}-1} (P(X_t, A_t;.) - \mathbf{e}_{X_t})h_k \\
&= \sum_{t=t_k}^{t_{k+1}-1} (P(X_t, A_t;.) - \mathbf{e}_{X_{t+1}})h_k + h_k(X_{t+1}) - h_k(X_t) \; .
\end{aligned}$$

Pour tout $t \in [t_k, t_{k+1}-1]$, posons $\xi_t = (P(X_t, A_t; .) - \mathbf{e}_{X_{t+1}})h_k$. On remarque que ξ_t est un incrément de martingale majoré par $D(\mathbf{M})$. En appliquant l'inégalité de Hoeffding-Azuma, et en utilisant le fait que pour tout $x, x' \in \mathcal{X}$, $h_k(x) - h_k(x') \leq D(\mathbf{M})$, on obtient que

$$\sum_{k=1}^{m(n)} \tilde{N}_k^{\pi_k}(P^{\pi_k} - I)h_k = \sum_{t=0}^{n} \xi_t + m(n)D(\mathbf{M}) \\ \leq D(\mathbf{M})\sqrt{\frac{n\log(1/\delta)}{2}} + m(n)D(\mathbf{M}) \quad \text{avec probabilité } 1-\delta \, . \tag{4.32}$$

3. Pour finir, on a

$$\tilde{N}_k^{\pi_k}(r_k^{\pi_k} - r^{\pi_k}) \leq \tilde{N}_k^{\pi_k}|r_k^{\pi_k} - \hat{r}_{t_k}^{\pi_k}| + \tilde{N}_k^{\pi_k}|\hat{r}_{t_k}^{\pi_k} - r^{\pi_k}| \leq 2\sum_x \tilde{N}_k(x, \pi_k(x))\frac{C_R}{\sqrt{N_{t_k}(x, \pi_k(x))}} \, . \tag{4.33}$$

D'après les équations (4.33), (4.31) et (4.32) et en utilisant la proposition 4.3, le regret jusqu'à l'instant n est majoré, avec probabilité $1-4\delta$, par

$$\text{Regret}_n \leq D(\mathbf{M})\sqrt{\frac{n\log(1/\delta)}{2}} + m(n)D(\mathbf{M}) + C_e(n,\delta) \\ + \sum_{k=1}^{m(n)} \sum_{(x,a)} \tilde{N}_k(x,a)\left(2\frac{C_R}{\sqrt{N_{t_k}(x,a)}} + 2D(\mathbf{M})\sqrt{2\frac{C_P}{N_{t_k}(x,a)}}\right) \, .$$

On peut montrer par récurrence que, pour toute séquence d'entiers $w_1, \ldots, w_n$ tels que, pour tout i, $w_i \leq W_{i-1} \stackrel{\text{def}}{=} \max\{1, \sum_{k=1}^{j-1} w_j\}$, (voir lemme 14 de [Auer et al., 2009a])

$$\sum_{k=1}^{n} \frac{w_k}{\sqrt{W_{k-1}}} \leq (\sqrt{2}+1)\sqrt{W_n} \, .$$

Ainsi, on en déduit que

$$\sum_{k=1}^{m(n)} \sum_{x,a} \frac{\tilde{N}_k(x,a)}{\sqrt{N_{t_k}(x,a)}} \leq (\sqrt{2}+1)\sum_{(x,a)} \sqrt{N_n(x,a)} \leq (\sqrt{2}+1)\sqrt{|\mathcal{X}||\mathcal{A}|n} \, , \tag{4.34}$$

où l'inégalité de Jensen à été utilisé pour majorer $\sum_{(x,a)} \sqrt{N_n(x,a)}$ par $\sqrt{|\mathcal{X}||\mathcal{A}|n}$. De plus, le nombre d'épisodes $m(n)$ peut être aisément majoré en se souvenant que pour tout épisode k, il existe un couple état-action (x,a) tel que $\tilde{N}_k(x,a) = N_{t_k}(x,a)$. Soit $K_{(x,a)}$ l'ensemble des épisodes au cours desquels le nombre de visites au couple (x,a) est doublé. Il existe au plus $1 + \log_2(n)$ épisodes dans $K_{(x,a)}$. En sommant sur les couples états-actions et en ajoutant les épisodes initiaux, on a

$$m(n) \leq |\mathcal{X}||\mathcal{A}|\log_2\left(\frac{8n}{|\mathcal{X}||\mathcal{A}|}\right) \, .$$

En combinant tous les termes, on a

$$\begin{aligned}
\text{Regret}_n &\leq D(\mathbf{M})\sqrt{\frac{n\log(1/\delta)}{2}} + D(\mathbf{M})|\mathcal{X}||\mathcal{A}|\log_2\left(\frac{8n}{|\mathcal{X}||\mathcal{A}|}\right) + \sqrt{\frac{n\log(1/\delta)}{2}} \\
&\quad + 2(\sqrt{2}+1)\sqrt{|\mathcal{X}||\mathcal{A}|n}\left(\sqrt{\frac{\log\left(4|\mathcal{X}||\mathcal{A}|\log(n)/\delta\right)}{1.99}}\right. \\
&\quad \left. + D(\mathbf{M})\sqrt{2|\mathcal{X}|\left(B + \log\left(B + \frac{1}{\log(n)}\right)\left[1 + \frac{1}{B + \frac{1}{\log(n)}}\right]\right)}\right) \\
&\leq 2D(\mathbf{M})|\mathcal{X}|\sqrt{|\mathcal{A}|n}(\sqrt{2}+1)\left[\frac{\sqrt{\log(1/\delta)}}{|\mathcal{X}|\sqrt{|\mathcal{A}|}\sqrt{2}(\sqrt{2}+1)} + \sqrt{\frac{\log(4|\mathcal{X}||\mathcal{A}|\log(n)/\delta)}{1.99D(\mathbf{M})^2|\mathcal{X}|}}\right. \\
&\quad \left. + \sqrt{2\left(B + \log\left(B + \frac{1}{\log(n)}\right)\left[1 + \frac{1}{B + \frac{1}{\log(n)}}\right]\right)}\right] \\
&\quad + D(\mathbf{M})|\mathcal{X}||\mathcal{A}|\log_2\left(\frac{8n}{|\mathcal{X}||\mathcal{A}|}\right)
\end{aligned}$$

En remarquant que

$$B + \log(B + 1/\log(n))\left[1 + 1/(B + 1/\log(n))\right] \leq 4B$$

et que $B = \log\left(\frac{2e|\mathcal{X}|^2|\mathcal{A}|\log(n)}{\delta}\right)$, on a

$$\begin{aligned}
\text{Regret}_n &\leq 2D(\mathbf{M})|\mathcal{X}|\sqrt{|\mathcal{A}|n}(\sqrt{2}+1)\left[\frac{1}{3}\sqrt{\log(1/\delta)} + \sqrt{\frac{\log(4|\mathcal{X}||\mathcal{A}|\log(n)/\delta)}{1.99}}\right. \\
&\quad \left. + 2\sqrt{2\log\left(\frac{2e|\mathcal{X}|^2|\mathcal{A}|\log(n)}{\delta}\right)}\right] + D(\mathbf{M})|\mathcal{X}||\mathcal{A}|\log_2\left(\frac{8n}{|\mathcal{X}||\mathcal{A}|}\right) \\
&\leq 2D(\mathbf{M})|\mathcal{X}|\sqrt{|\mathcal{A}|n\log(\log(n)/\delta)}(\sqrt{2}+1)(1/3 + 1/\sqrt{1.99} + 2\sqrt{2}) \\
&\quad + 2D(\mathbf{M})|\mathcal{X}|\sqrt{|\mathcal{A}|n}(\sqrt{2}+1)\left(\sqrt{\frac{\log(4|\mathcal{X}||\mathcal{A}|)}{1.99}} + 2\sqrt{2\log(2e|\mathcal{X}|^2|\mathcal{A}|)}\right) \\
&\quad + D(\mathbf{M})|\mathcal{X}||\mathcal{A}|\log_2\left(\frac{8n}{|\mathcal{X}||\mathcal{A}|}\right)
\end{aligned}$$

Si $n \geq |\mathcal{A}|\log_2(8|\mathcal{A}|)$ alors

$$D(\mathbf{M})|\mathcal{X}||\mathcal{A}|\log_2\left(\frac{8n}{|\mathcal{X}||\mathcal{A}|}\right) \leq\leq D(\mathbf{M})|\mathcal{X}|\sqrt{|\mathcal{A}|\log(\log(n)/\delta)}(\sqrt{2}+1)\sqrt{n}$$

et ainsi

$$\text{Regret}_n \leq\leq 22D(\mathbf{M})|\mathcal{X}|\sqrt{|\mathcal{A}|n\log(\log(n)/\delta)} + 23D(\mathbf{M})|\mathcal{X}|\sqrt{|\mathcal{A}|n\log(2e|\mathcal{X}||\mathcal{A}|)}\,.$$

□

Preuve du théorème 4.5. L'idée de la preuve de ce second théorème est de considérer séparément les épisodes selon deux caractéristiques : leur longueur et le type de politique jouée. Pour commencer, introduisons quelques notations qui seront utiles par la suite. Pour toute

politique π notons T_π l'entier tel que pour tout $n \geq T_\pi$, la moyenne des récompenses espérée sur les n premiers instants est $\Delta(\mathbf{M})/2$-proche de la récompense moyenne sous π, c'est-à-dire

$$\forall n \geq T_\pi \ , \quad \eta^\pi - \frac{1}{n}\mathbb{E}^\pi \left[\sum_{t=0}^{n-1} R_t\right] \leq \frac{\Delta(\mathbf{M})}{2} \ . \tag{4.35}$$

On dit qu'un épisode k est *ϵ-mauvais* si l'espérance de la moyennes des récompenses reçues pendant l'épisode n'est pas ϵ-proche de la récompense moyenne optimale :

$$\mathbb{E}\left(\sum_{t=t_k}^{t_{k+1}-1} \eta^* - R_t\right) \geq \epsilon \, \mathbb{E}\left(t_{k+1} - t_k\right) \ .$$

Il est dit *ϵ-bon* si

$$\mathbb{E}\left(\sum_{t=t_k}^{t_{k+1}-1} \eta^* - R_t\right) < \epsilon \, \mathbb{E}\left(t_{k+1} - t_k\right) \ .$$

Notons K_ϵ l'ensemble des indices des épisodes ϵ-mauvais. Remarquons que, si une politique π est jouée dans un épisode $\Delta(\mathbf{M})/2$-bon de longueur supérieure à T_π, alors la politique π est une politique optimale. En effet :

$$\eta^* - \eta^\pi = \left(\eta^* - \frac{\mathbb{E}\left(\sum_{t=t_k}^{t_{k+1}-1} R_t\right)}{\mathbb{E}\left(t_{k+1} - t_k\right)}\right) + \left(\frac{\mathbb{E}\left(\sum_{t=t_k}^{t_{k+1}-1} R_t\right)}{\mathbb{E}\left(t_{k+1} - t_k\right)} - \eta^\pi\right) < \frac{\Delta(\mathbf{M})}{2} + \frac{\Delta(\mathbf{M})}{2} = \Delta(\mathbf{M})$$

Pour majorer le regret, on séparera alors les épisodes selon qu'ils sont $\Delta(\mathbf{M})/2$-bons ou mauvais et selon qu'ils sont d'une longueur plus petite ou plus grande que T_{π_k}. Notons K_c l'ensemble des indices des épisodes « courts » c'est-à-dire les épisodes k de longueur inférieure à T_{π_k}.

Dans cette démonstration, notons

$$\mathrm{RegretEp}'_k \stackrel{\text{def}}{=} \sum_{t=t_k}^{t_{k+1}-1} \eta^* - R_t$$

le regret durant l'épisode k. Notons que $\mathrm{RegretEp}'_k$ est différent de $\mathrm{RegretEp}_k$, utilisé dans la démonstration précédente. L'espérance du regret est décomposée comme suit

$$\begin{aligned}
\mathbb{E}\left[\mathrm{Regret}_n\right] &= \mathbb{E}\left[\sum_{k=1}^{m(n)} \mathrm{RegretEp}'_k\right] \\
&= \mathbb{E}\left[\sum_{k \in K_{\Delta(\mathbf{M})/2}} \mathrm{RegretEp}'_k\right] + \mathbb{E}\left[\sum_{k \notin K_{\Delta(\mathbf{M})/2}} \mathrm{RegretEp}'_k)\right] \\
&= \mathbb{E}\left[\sum_{k \in K_{\Delta(\mathbf{M})/2}} \mathrm{RegretEp}'_k\right] + \mathbb{E}\left[\sum_{\substack{k \notin K_{\Delta(\mathbf{M})/2} \\ k \in K_c}} \mathrm{RegretEp}'_k\right] + \mathbb{E}\left[\sum_{\substack{k \notin K_{\Delta(\mathbf{M})/2} \\ k \notin K_c}} \mathrm{RegretEp}'_k\right] \\
&\leq \mathbb{E}\left[\sum_{k \in K_{\Delta(\mathbf{M})/2}} \mathrm{RegretEp}'_k\right] + \mathbb{E}\left[\sum_{k \in K_c} \mathrm{RegretEp}'_k\right] + \mathbb{E}\left[\sum_{k, \pi_k \text{ est optimale}} \mathrm{RegretEp}'_k\right]
\end{aligned}$$

Dans la suite, nous majorons chacun de ces trois termes séparément.

Regret dans les épisodes $\Delta(\mathbf{M})/2$-mauvais

Notons $L_\epsilon = \sum_{k\in K_\epsilon}(t_{k+1}-t_k)$ la longueur cumulée de tous les épisodes ϵ-mauvais et $\text{Regret}_{\epsilon\text{-bad}}$ le regret associé :

$$\text{Regret}_{\epsilon\text{-bad}} \stackrel{\text{def}}{=} \sum_{k\in K_\epsilon}\sum_{t=t_k}^{t_{k+1}-1}(\eta^*(\mathbf{M})-R_t)\ .$$

En suivant la même décomposition que dans la preuve du théorème précédent, le regret $\text{Regret}_{\epsilon\text{-bad}}$ est majoré, avec probabilité $1-\delta$, par

$$\begin{aligned}\text{Regret}_{\epsilon\text{-bad}} &\le L_\epsilon\eta^*(\mathbf{M}) - \sum_{k\in K_\epsilon}\sum_{(x,a)}\tilde{N}_k(x,a)r(x,a) + C_e(L_\epsilon,\delta)\\ &\le \sum_{k\in K_\epsilon}\Big(\tilde{N}_k^{\pi_k}(P_k^{\pi_k}-P^{\pi_k})h_k + \tilde{N}_k^{\pi_k}(P^{\pi_k}-I)h_k + \tilde{N}_k^{\pi_k}(r_k^{\pi_k}-r^{\pi_k})\Big) + C_e(L_\epsilon,\delta)\ ,\end{aligned}$$

où $C_e(L_\epsilon,\delta)=\sqrt{\log(1/\delta)L_\epsilon/2}$. Pour tout $t\in[t_k,t_{k+1}-1]$, posons

$$\xi_t = (P(X_t,A_t;.)-\mathbf{e}_{X_{t+1}})h_k\mathbb{1}_{\{k\in K_\epsilon\}}\ .$$

En utilisant le théorème A.6, on a

$$\sum_{k\in K_\epsilon}\tilde{N}_k^{\pi_k}(P^{\pi_k}-I)h_k = \sum_{t=0}^{n}\xi_t \le D(\mathbf{M})\sqrt{\frac{\log(4\log(n)/\delta)}{1.99}L_\epsilon} \quad \text{a.p. } 1-\delta.$$

En utilisant les équations (4.33), (4.31) et d'après la proposition 4.3, le regret jusqu'à l'instant n dans les épisodes ϵ-mauvais est majoré, avec probabilité $1-4\delta$, par

$$\begin{aligned}\text{Regret}_{\epsilon\text{-bad}} \le\ & D(\mathbf{M})\sqrt{\frac{\log(4\log(n)/\delta)}{1.99}L_\epsilon} + m(n)D(\mathbf{M}) + C_e(L_\epsilon,\delta)\\ &+ \sum_{k\in K_\epsilon}\sum_{(x,a)}\tilde{N}_k(x,a)(2C_R(x,a,t_k)+2D(\mathbf{M})\sqrt{2C_P(x,a,t_k)})\ .\end{aligned} \tag{4.36}$$

De manière similaire à l'équation (4.34), on a

$$\sum_{k\in K_\epsilon}\sum_{x,a}\frac{\tilde{N}_k(x,a)}{\sqrt{N_{t_k}(x,a)}} \le (\sqrt{2}+1)\sqrt{|\mathcal{X}||\mathcal{A}|L_\epsilon}\ .$$

Cette inégalité diffère de l'équation (4.34) car la longueur totale L_ϵ des épisodes dans K_ϵ est aléatoire. Ce résultat est prouvé dans [Auer et al., 2009a]. En simplifiant l'équation (4.36) et en utilisant l'inégalité ci-dessus, on a

$$\begin{aligned}\text{Regret}_{\epsilon\text{-bad}} \le\ & D(\mathbf{M})|\mathcal{X}|\sqrt{|\mathcal{A}|}\\ &\Big(C_1\sqrt{L_\epsilon\log(\log(n)/\delta)} + C_2\sqrt{L_\epsilon\log(2e|\mathcal{X}|^2|\mathcal{A}|)} + \sqrt{|\mathcal{A}|}(2+\log_2(n))\Big)\end{aligned}$$

où C_1 et C_2 sont des constantes telles que $C_1\le 20,2$ et $C_2\le 18,5$. Pour pouvoir simplifier cette écriture, étudions séparemment le cas où le terme dominant est $\sqrt{|\mathcal{A}|}\log_2(n)$ ou $\sqrt{L_\epsilon\log(\log(n)/\delta)}$. Si la longueur des épisodes ϵ-mauvais est petite, plus précisément si

$$L_\epsilon \le \frac{|\mathcal{A}|(\log_2(n))^2}{\log(\log(n)/\delta)}$$

alors

$$\text{Regret}_{\epsilon\text{-bad}} \leq D(\mathbf{M})|\mathcal{X}|\sqrt{|\mathcal{A}|} \left(C_1 \log_2(n)\sqrt{|\mathcal{A}|} + C_2\sqrt{\frac{|\mathcal{A}|(\log_2(n))^2 \log(2e|\mathcal{X}|^2|\mathcal{A}|)}{\log(\log(n)/\delta)}} + \sqrt{|\mathcal{A}|}(2+\log_2(n)) \right)$$

avec probabilité $1-4\delta$. En prenant $\delta = \log(n)/n$, on déduit de l'inégalité précédente qu'il existe $C \leq 23$ et $C' \leq 27$ telles que

$$\mathbb{E}\left[\text{Regret}_{\epsilon\text{-bad}}\right] \leq D(\mathbf{M})|\mathcal{X}||\mathcal{A}| \left(C\log_2(n) + C'\sqrt{\log(n)\log(2e|\mathcal{X}|^2|\mathcal{A}|)} \right) .$$

Par ailleurs, si la longueur des épisodes ϵ-mauvais est suffisamment grande, i.e. si

$$L_\epsilon \geq \frac{|\mathcal{A}|(\log_2(n))^2}{\log(\log(n)/\delta)} ,$$

il existe des constantes $C_1 \leq 23$ et $C_2 \leq 19$ telles que

$$\text{Regret}_{\epsilon\text{-bad}} \leq D(\mathbf{M})|\mathcal{X}|\sqrt{|\mathcal{A}|L_\epsilon} \left(C_1\sqrt{\log(\log(n)/\delta)} + C_2\sqrt{\log(2e|\mathcal{X}|^2|\mathcal{A}|)} \right) ,$$

avec probabilité $1-4\delta$. Par définition de $\text{Regret}_{\epsilon\text{-bad}}$,

$$\epsilon L_\epsilon \leq \text{Regret}_{\epsilon\text{-bad}} .$$

On en déduit que

$$\sqrt{L_\epsilon} \leq \frac{D(\mathbf{M})|\mathcal{X}|\sqrt{|\mathcal{A}|}}{\epsilon} \left(C_1\sqrt{\log(\log(n)/\delta)} + C_2\sqrt{\log(2e|\mathcal{X}|^2|\mathcal{A}|)} \right) ,$$

et donc,

$$\text{Regret}_{\epsilon\text{-bad}} \leq \frac{D(\mathbf{M})^2|\mathcal{X}|^2|\mathcal{A}|}{\epsilon} \left(C_1\sqrt{\log(\log(n)/\delta)} + C_2\sqrt{\log(2e|\mathcal{X}|^2|\mathcal{A}|)} \right)^2 ,$$

avec probabilité $1-4\delta$. En posant $\delta = \log(n)/n$, on en déduit que

$$\mathbb{E}\left[\text{Regret}_{\epsilon\text{-bad}}\right] \leq \frac{D(\mathbf{M})^2|\mathcal{X}|^2|\mathcal{A}|}{\epsilon} \left(C_1^2\log(n) + C_2^2\log(2e|\mathcal{X}|^2|\mathcal{A}|) + 2C_1C_2\log(n)\log(2e|\mathcal{X}|^2|\mathcal{A}|) \right)$$

Regret dans les épisodes « courts »

Considérons l'ensemble des épisodes de longueur plus petite que T_{π_k}. La courte durée de ces épisodes ne permet pas de considérer que la récompense moyenne accumulée le long de l'épisode est proche de η^{π_k}. Le regret d'un tel épisode est alors majoré par T_{π_k}. Il reste à majorer le nombre d'épisodes « courts ». Pour cela, on fixe un couple (x,a) et on considère l'ensemble $K_{c,(x,a)}$ des épisodes « courts » tels que le nombre d'occurrences de ce couple a été doublé pendant l'épisode. On rappelle qu'un épisode s'arrête dès que un couple état-action a été vu autant de fois pendant l'épisode qu'avant l'épisode. On s'intéresse donc à l'ensemble des épisodes k tels que $\tilde{N}_k(x,a) = N_{t_k}(x,a)$. Le regret de chacun de ces épisodes est majoré par

$$T_{(x,a)} = \max_{\pi,\pi(x)=a} T_\pi .$$

La suite $\left(\tilde{N}_k(x,a)\right)_{k\in K_{c,(x,a)}}$ des nombres de visites au couple (x,a) dans les épisodes $k \in K_{c,(x,a)}$ est une suite croissante dont les éléments sont au moins égaux au double de l'élément

précédent. Plus précisément, si on pose u_j le nombres de visites $\tilde{N}_k(x,a)$ au couple (x,a) pendant l'épisode k qui est le j-ième épisode de $K_{c,(x,a)}$ depuis le début de l'interaction, alors pour tout j, $u_{j+1} \geq 2u_j$. Il y a donc au plus $\lceil 1 + \log_2(T_{(x,a)}) \rceil$ épisodes de longueur inférieure à $T_{(x,a)}$ tels que $\tilde{N}_k(x,a) = N_{t_k}(x,a)$. En sommant sur tous les couples état-action, on obtient que :

$$\mathbb{E}\left(\sum_{k \in K_c} \text{RegretEp}'_k\right) \leq \sum_{(x,a)} \lceil 1 + \log_2(T_{(x,a)}) \rceil T_{(x,a)} \ .$$

Regret dans les épisodes où une politique optimale est suivie

On rappelle que $D(\mathbf{M}) = \max_{x,x'} \min_\pi \mathbb{E}^\pi_{\mathbf{M}}[\tau(x,x')]$ où $\tau(x,x')$ est le plus petit temps d'atteinte de x' en partant de x. Dans tous les épisodes k tels que π_k est une politique optimale, le regret $\mathbb{E}\left[\text{RegretEp}_k\right]$ peut alors être majoré par $D(\mathbf{M})$. D'où

$$\mathbb{E}\left[\sum_{k, \pi_k \text{ est optimale}} \text{RegretEp}'_k\right] \leq D(\mathbf{M}) m(n) \leq D(\mathbf{M}) |\mathcal{X}||\mathcal{A}| \log_2(8n) \ .$$

Addition de tous les épisodes

En utilisant les majorations des regrets dans les trois cas étudiés ci-dessus, on obtient que

$$\begin{aligned}
\mathbb{E}(\text{Regret}_n) &\leq \mathbb{E}\left[\sum_{k \in K_{\Delta(\mathbf{M})/2}} \text{RegretEp}'_k\right] + \mathbb{E}\left[\sum_{k \in K_c} \text{RegretEp}'_k\right] + \mathbb{E}\left[\sum_{k, \pi_k \text{ est optimale}} \text{RegretEp}'_k\right] \\
&\leq \frac{2D(\mathbf{M})^2 |\mathcal{X}|^2 |\mathcal{A}|}{\Delta(\mathbf{M})} \left(C \log(n) + C' \log(2e|\mathcal{X}|^2|\mathcal{A}|)\right) \\
&\qquad + \sum_{(x,a)} \lceil 1 + \log_2(T_{(x,a)}) \rceil T_{(x,a)} + C_4 D(\mathbf{M}) |\mathcal{X}||\mathcal{A}| \log_2(8n) \ .
\end{aligned}$$

□

Preuve du théorème 4.6. La preuve de ce théorème est similaire aux deux preuves des théorèmes précédents. Les récompenses étant déterministes, pour tout horizon le regret n, et en utilisant l'inégalité de concentration présentée dans le théorème 4.1, le regret est majoré, avec probabilité $1 - \delta$ par

$$\begin{aligned}
\text{Regret}_n &\leq D(\mathbf{M}) \sqrt{\frac{n \log(1/\delta)}{2}} + m(n) D(\mathbf{M}) + C_e(n,\delta) + 2D(\mathbf{M}) \sum_{k=1}^{m(n)} \sum_{(x,a)} \tilde{N}_k(x,a) \sqrt{2 \frac{C_P}{N_{t_k}(x,a)}} \\
&\leq C D(\mathbf{M}) \sqrt{|\mathcal{X}||\mathcal{A}| n C_P} \\
&\leq C' D(\mathbf{M}) \sqrt{|\mathcal{X}||\mathcal{A}| \kappa n \log(\log(n)/\delta)}
\end{aligned}$$

pour des constantes C et C' indépendantes du modèle.

Le calcul de la borne supérieure du regret espéré à partir de cette inégalité est identique au raisonnement suivi dans la preuve du théorème 4.5. □

4.4.3 Résultats numériques

Dans cette section, nous comparons l'algorithme KL-UCRL à l'algorithme ϵ-glouton et à l'algorithme UCRL2. On rappelle que l'algorithme ϵ-glouton consiste à jouer l'action optimale

pour le modèle estimé avec probabilité $1-\epsilon$ et une action au hasard avec probabilité ϵ. On fixe ϵ à $1/\sqrt{t}$ qui semble donner de bons résultats empiriques. Pour les algorithmes KL-UCRL et UCRL2, les constantes C_P et C_R sont fixées de manière à garantir que les bornes de majoration du regret du théorème 4.4 ci-dessus et du théorème 2 de [Jaksch et al., 2010] soient satisfaites avec probabilité 0.95.

Dans un premier temps, un générateur d'environnement aléatoire a été utilisé pour créer des environnement avec faible connectivité à 10 états et 5 actions. Les récompenses sont aléatoires et comprises entre 0 et 1. Sous ces environnements aléatoires, chaque état est connecté, en moyenne, à cinq autres états (avec des probabilités de transition tirées selon une loi de Dirichlet). La figure 4.4 représente le regret cumulé au cours de 100 réplications Monte-Carlo en générant à chaque fois un nouvel environnement et en suivant de manière indépendante les trois algorithmes Glouton, UCRL2 et KL-UCRL. On observe que l'algorithme Glouton permet d'obtenir un regret très faible et que le regret accumulé en suivant l'algorithme Kl-UCRL est plus petit que celui en suivant l'algorithme UCRL2. Sur un grand nombre d'environnements, les performances de l'algorithme Glouton sont en effet remarquables. Cependant, cet algorithme est centré sur l'exploitation au détriment de l'exploration ce qui peut engendrer de très grandes pertes sous certains MDP.

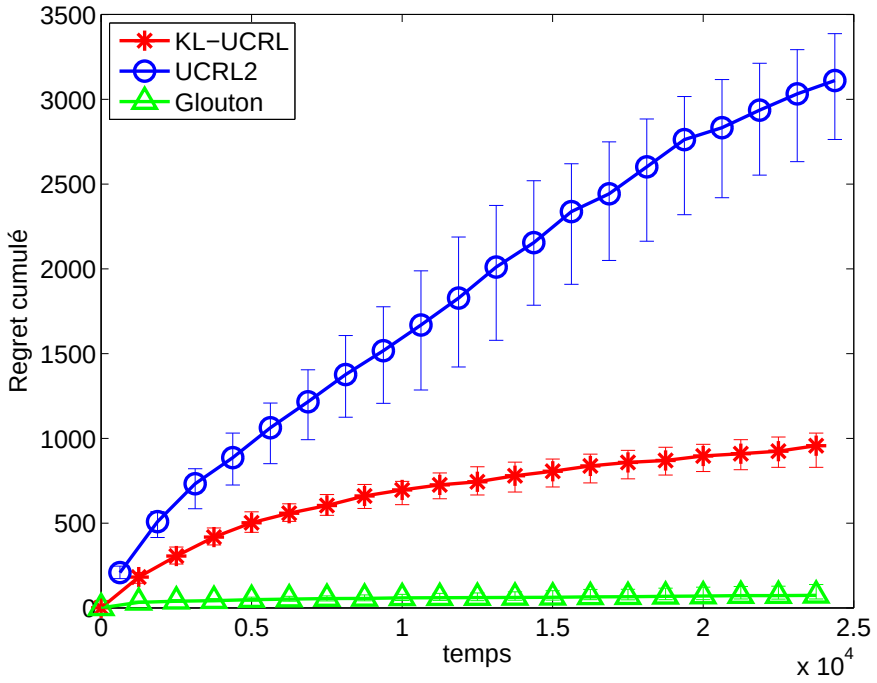

FIGURE 4.4 – Comparaison du regret des algorithmes Glouton, UCRL2 et KL-UCRL dans des modèles à faible connectivité générés de manière aléatoire.

Illustrons maintenant les performances de notre algorithme dans un contexte où l'optimisme est très important, c'est-à-dire où il est nécessaire de suivre une bonne politique d'exploration pour déterminer les actions optimales. Pour cela, nous considérons les environnements de référence *rivière* et *six bras* proposés par [Strehl and Littman, 2008]. L'environnement *rivière* a déjà été présenté dans le paragraphe 1.1.1 mais on considère une généralisation consistant en une rivière de longueur $|\mathcal{X}|$ dans laquelle l'agent peut nager soit vers la gauche soit vers la droite. Nager vers la droite (contre le courant de la rivière) permet d'atteindre l'état de droite avec probabilité 0.35 ; cela laisse l'agent dans le même état avec une grande probabilité égale à 0.6, et dévie l'agent vers la gauche avec une probabilité 0.05 (voir figure 4.5). Au contraire, nager vers la gauche (avec le courant de la rivière) permet toujours d'atteindre

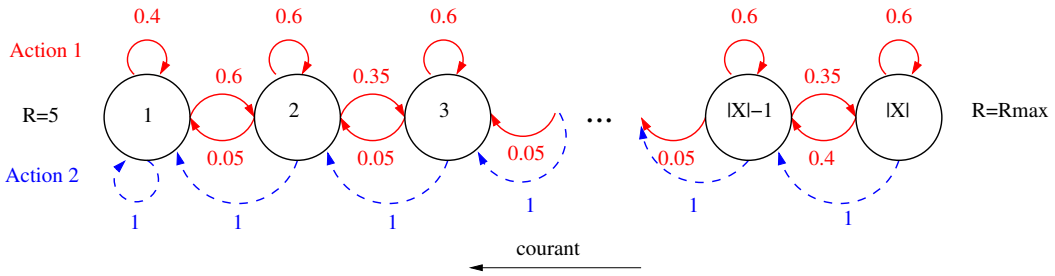

FIGURE 4.5 – Modèle de transition dans la *Rivière* : les flèches continues (resp. en pointillés) représentent les transitions si l'action 1 (resp. 2) a été choisie.

l'état de gauche. L'agent reçoit une petite récompense égale à 5 quand il atteint la rive de gauche et une récompense beaucoup plus grande égale à $R_{\max}$ quand il atteint la rive de droite – les autres états ne permettent pas de recevoir de récompense. De manière évidente, si la récompense reçue sur la rive de droite est assez grande comparée à celle de gauche, la politique optimale consiste à nager en permanence vers la droite. Ce MDP nécessite une procédure d'exploration suffisante, puisque l'agent, n'ayant aucune idée a priori des récompenses reçues dans chaque état, doit atteindre au moins une fois l'état le plus à droite pour découvrir que c'est l'état le plus prometteur. On considère tout d'abord le modèle standard tel que présenté dans le paragraphe 1.1.1 où la récompense maximale $R_{\max}$ vaut 10000 et la longueur de la rivière est $|\mathcal{X}| = 6$. Nous observons dans la figure 4.6 que, comme prévu, l'algorithme Glouton a de très faibles performances et que l'algorithme KL-UCRL a un plus petit regret que l'algorithme UCRL2 sur cet environnement de référence.

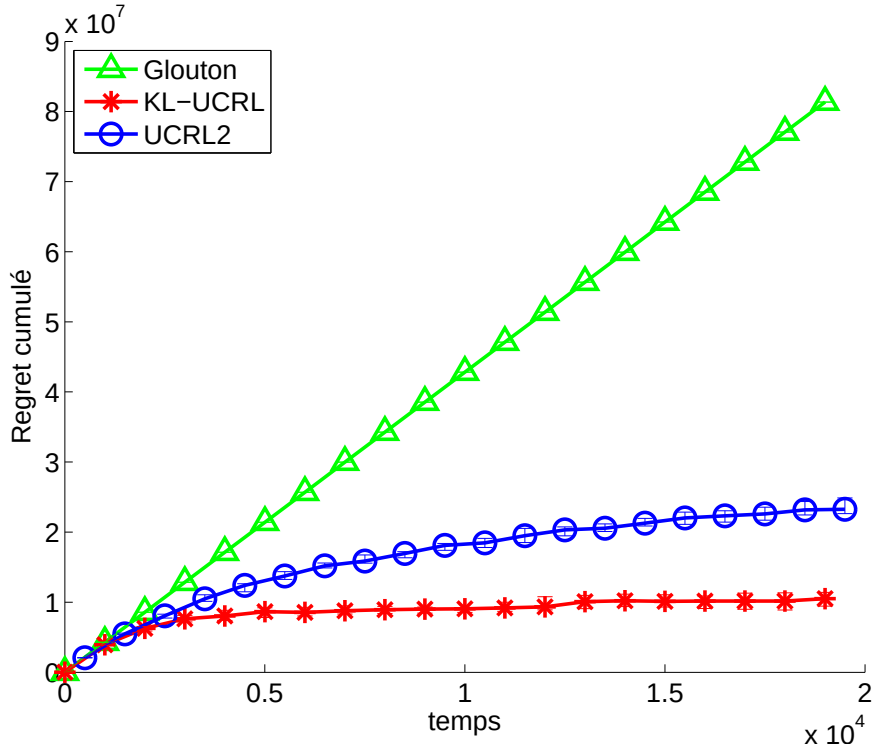

FIGURE 4.6 – Comparaison du regret des algorithmes Glouton, UCRL2 et KL-UCRL dans le modèle de rivière.

L'exploration est d'autant plus importante dans ce modèle que la récompense maximale

est très grande. De plus, la longueur de la rivière étant assez petite, un algorithme avec une politique d'exploration pertinente finit forcément par atteindre l'état le plus prometteur et à déterminer la politique optimale. La figure 4.7 représente le regret moyen empirique calculé sur les 50000 premiers instants en suivant les algorithmes UCRL2 et KLUCRL au cours de 10 réplications Monte-Carlo pour différentes valeurs de $R_{\max}$ et de $|\mathcal{X}|$. On observe que les regrets varient considérablement d'un cas à l'autre et augmentent avec la longueur de la rivière et la valeur de la récompense maximale. On remarque également que plus la récompense maximale est grande plus la différence de regret entre les deux algorithmes est marquante. Celle-ci diminue lorsque la longueur de la rivière augmente.

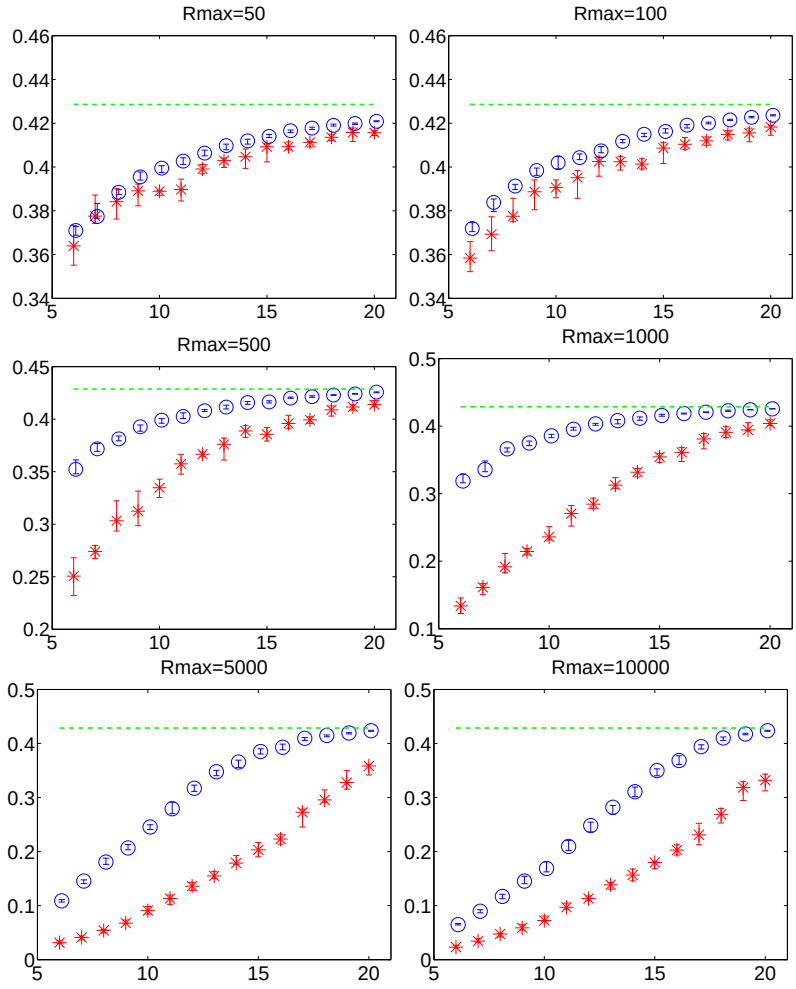

FIGURE 4.7 – Regret moyenné sur 50000 instants en suivant les algorithmes UCRL2 (rond) et KL-UCRL (étoile) en fonction de la longueur de la rivière $|\mathcal{X}| \in \{5, \ldots, 20\}$ pour différentes valeurs de $R_{\max}$. La ligne en pointillé désigne la récompense moyenne optimale $\eta^*(\mathbf{M})$ pour le modèle.

Le deuxième environnement proposé par [Strehl and Littman, 2008], appelé *six bras*, s'apparente à un problème de bandit sans vraiment en être un. Il consiste en 7 états dont un (l'état 0) est l'état initial. A partir de ce dernier, l'agent peut choisir une parmi les 6 actions : l'action $a \in \{1, \ldots, 6\}$ mène à l'état $x = a$ avec probabilité p_a (voir figure 4.8) et laisse l'agent dans l'état initial avec probabilité $1 - p_a$. A partir des autres états, la conséquence des actions est déterministe : certaines actions mènent à l'état initial tandis que d'autres permettent de rester dans l'état courant . En restant dans un des états de 1 à 6, l'agent reçoit une récompense égale à R_x (voir figure 4.8), sinon, aucune récompense n'est reçue. La récompense maximale

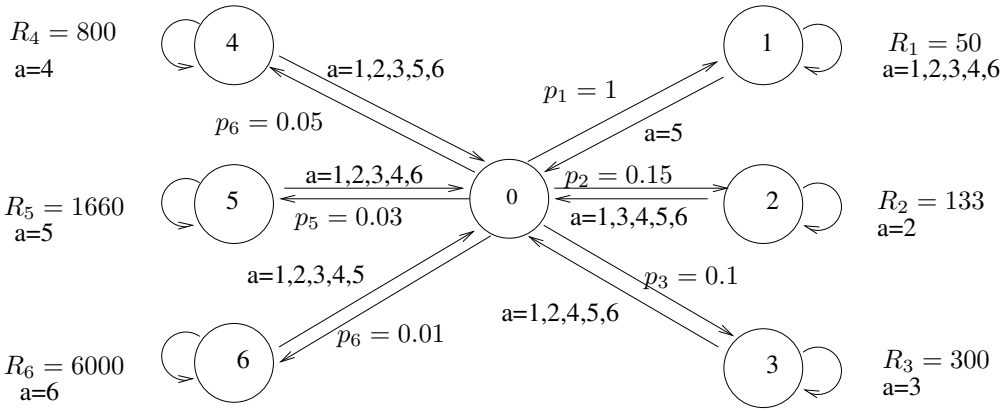

FIGURE 4.8 – Modèle de transition de l'environnement *six bras*.

est reçue en choisissant l'action 6 dans l'état 6 ; de plus cette action permet de rester dans l'état 6. Ainsi, la politique optimale consiste à jouer l'action 6 en permanence, même si, au début, cela aboutit à ne recevoir aucune récompense puisque la probabilité de transition entre l'état 0 et l'état 6 est très petite. Remarquons que ce MDP a une connectivité assez faible : excepté l'état initial, tous les états ne sont connectés qu'à deux états. Dans cet environnement, les récompenses étant déterministes, on modifie légèrement les deux algorithmes en considérant que l'agent les connaît à l'avance. Nous observons dans la figure 4.9 que l'algorithme KL-UCRL a un plus petit regret que l'algorithme UCRL2 sur ce deuxième environnement de référence. Tout comme dans l'environnement *rivière*, il est crucial pour l'agent d'apprendre qu'aucune action ne mène certains états à l'état le plus prometteur.

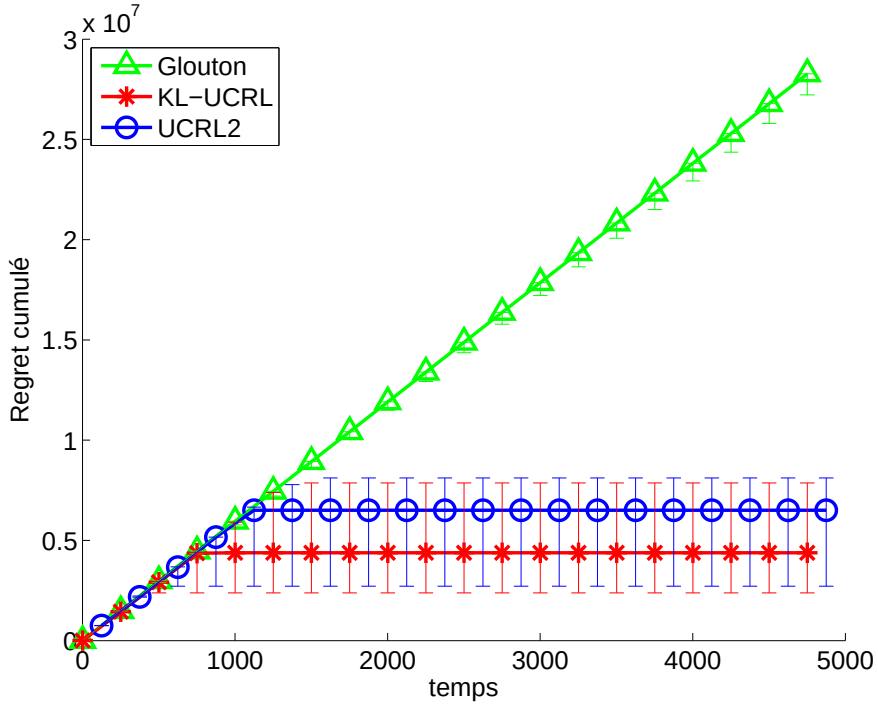

FIGURE 4.9 – Comparaison du regret des algorithmes Glouton, UCRL2 et KL-UCRL dans le modèle à six bras.

4.4.4 Discussion

Dans cette section, nous exposons les avantages de l'utilisation d'une région de confiance basée sur la divergence de Kullback-Leibler plutôt qu'une boule L^1, comme proposé par exemple dans [Jaksch et al., 2010; Tewari and Bartlett, 2008], dans le calcul de la politique optimiste. Ceci nous permettra d'interpréter les différences de performance observées lors des simulations. Dans l'algorithme KL-UCRL, l'optimisme revient à maximiser la fonction linéaire $V'q$ dans un voisinage de $\mathbb{S}^{|\mathcal{X}|}$ défini par la pseudo-métrique de Kullback-Leibler (voir l'équation (4.18)), tandis que l'algorithme UCRL2 utilise une norme L^1 :

$$\max_{q \in \mathbb{S}^{|\mathcal{X}|}} V'q \quad \text{tel que} \quad \|p - q\|_1 \leq \epsilon' \,. \tag{4.37}$$

Continuité

Soit un vecteur de probabilité p ; notons q^{KL} (resp. q^1) le vecteur de probabilité qui maximise l'équation (4.18) (resp. équation (4.37)). De manière évidente, q^{KL} et q^1 sont respectivement situés sur le bord de l'espace convexe $\{q \in \mathbb{S}^{|X|} : KL(p;q) \leq \epsilon\}$ et à l'un des sommets du polytope $\{q \in \mathbb{S}^{|X|} : \|p - q\|_1 \leq \epsilon'\}$. Une première différence notable entre ces voisinages est que, grâce à la régularité du voisinage KL, q^{KL} varie continûment par rapport au vecteur V, ce qui n'est pas le cas de q^1.

Ceci est illustré sur la figure 4.10 pour des voisinages L^1 et KL autour d'un vecteur de probabilité à 3 dimensions. L'ensemble $\mathbb{S}^3$ des vecteurs de probabilité à 3 dimension est représenté par un triangle dont les sommets sont les vecteurs $(1,0,0)'$, $(0,1,0)'$ et $(0,0,1)'$ en projection sur le plan défini par ces trois points. Le vecteur de probabilité p est représenté par une étoile blanche, et les vecteurs q^{KL} et q^1 par des points blancs. La flèche représente la direction de la projection du vecteur V sur le simplexe et indique le gradient de la fonction linéaire à maximiser. On observe que le vecteur de probabilité q^1 qui maximise (4.37) peut varier de manière significative pour des petits changements de la fonction de valeur, tandis que q^{KL} varie de manière continue.

Transitions invraisemblables

Soit $i_m = \text{argmin}_j V(j)$ et $i_M = \text{argmax}_j V(j)$. Le vecteur q^1 qui maximise l'équation (4.37) est tel que $q^1_{i_m} = \max(p_{i_m} - \epsilon'/2, 0)$ et $q^1_{i_M} = \min(p^1_{i_M} + \epsilon'/2, 1)$. Ceci a deux conséquences

1. si p est tel que $0 < p_{i_m} < \epsilon'/2$, alors le vecteur $q^1_{i_m} = 0$; donc le modèle optimiste peut mettre à 0 une transition qui a en fait déjà été observée, ce qui va à l'encontre du principe même de l'optimisme. En effet, un MDP optimiste ne devrait pas interdire des transitions qui existent réellement, même si elles mènent à des états ayant des faibles valeurs.
2. si p est tel que $p_{i_M} = 0$, alors $q^1_{i_M}$ n'est jamais égal à 0 ; donc, un algorithme optimiste utilisant des voisinages L^1 attribue toujours des probabilités de transition positives à i_M même si la transition est impossible sous le vrai MDP même si suffisamment d'observations ont été accumulées. Le bonus d'exploration est donc gaspillé alors qu'il pourrait être utilisé pour favoriser d'autres transitions.

Ceci explique une grande partie des avantages de l'algorithme KL-UCRL par rapport à l'algorithme UCRL2 observées lors des expériences. En effet, q^{KL} attribue toujours une probabilité strictement positive à une transition qui a été observée. L'algorithme renonce éventuellement à des transitions non observées si suffisamment de preuves statistiques ont été accumulées allant à l'encontre de l'existence de telles transitions, même si celles-ci mènent à un état prometteur. L'algorithme 4.5 fonctionne de la façon suivante :

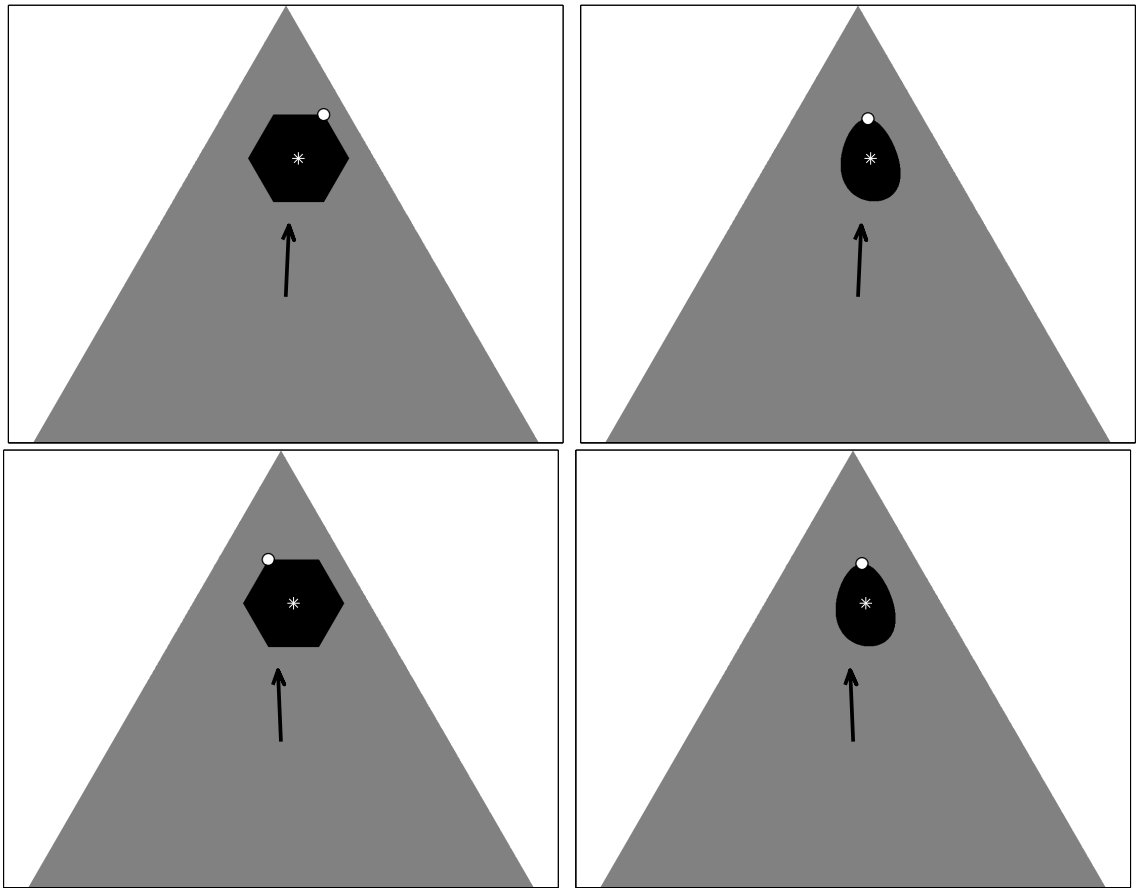

FIGURE 4.10 – Les voisinages L^1 $\{q \in \mathbb{S}^3 : \|p-q\|_1 \leq 0.2\}$ (gauche) et KL $\{q \in \mathbb{S}^3 : KL(p,q) \leq 0.02\}$ (droite) autour du vecteur de probabilité $p = (0.15, 0.2, 0.65)'$ (étoile blanche). Les points blancs désignent la probabilité qui maximise les équations (4.18) et (4.37) avec $V = (0, 0.05, 1)'$ (haut) et $V = (0, -0.05, 1)'$ (bas). Le paramètre utilisé est $\epsilon = 0.05$.

- pour tout i tel que $p_i \neq 0$, on a $q^{KL}(i) \neq 0$;
- pour tout $i \neq i_M$ tel que $p_i = 0$, on a $q^{KL}(i) = 0$;
- si $p_{i_M} = 0$ et si $f(V_{i_M}) < \epsilon$, alors $q^{KL}_{i_M} > 0$, sinon $q^{KL}_{i_M} = 0$.

Dans l'algorithme optimiste KL-UCRL, le rayon du voisinage ϵ décroît lorsque le nombre de visites aux différents couples état-action augmente. Ainsi, après un certain nombre de visites, la valeur de ϵ est suffisamment petite pour que la condition $f(V_{i_M}) < \epsilon$ ne soit plus vérifiée et que la transition $q^{KL}_{i_M}$ vers l'état le plus prometteur soit mise à 0. Nous illustrons ces deux différences importantes dans la figure 4.11, en représentant les voisinages L^1 et KL ainsi que les vecteurs de probabilités q^{KL} et q^1, tout d'abord si p_{i_m} est positif et très petit, et ensuite si p_{i_M} est égal à 0. La figure 4.12 illustre également le dernier cas, en représentant l'évolution du vecteur de probabilité q qui maximise (4.37) et (4.18) pour un exemple avec $p = (0.3, 0.7, 0)'$, $V = (1, 2, 3)'$ et ϵ qui décroît de $1/2$ à $1/500$.

4.5 Conclusion

Les algorithmes KL-UCB et KL-UCRL que nous avons proposés sont des algorithmes d'apprentissage par renforcement *« model-based »* optimistes, respectivement dans un modèle de bandit à récompenses binaires et dans un MDP à espaces d'états et d'actions finis, qui utilisent la divergence de Kullback-Leibler pour calculer le modèle optimiste. Ils reposent sur

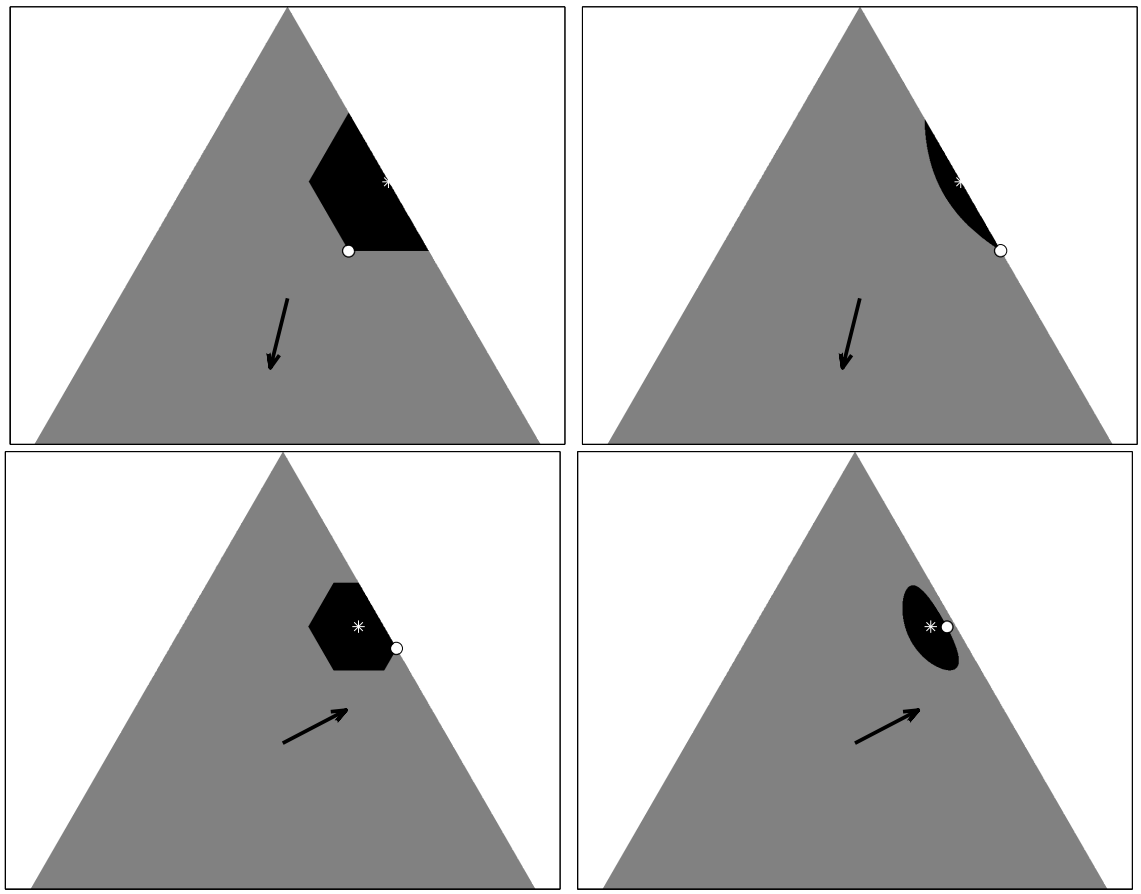

FIGURE 4.11 – Les voisinages L^1 (gauche) et KL (droite) autour des vecteurs de probabilité $p = (0, 0.4, 0.6)'$ (haut) et $p = (0.05, 0.35, 0.6)'$ (bas). Le point blanc désigne la probabilité qui maximise les équations (4.18) et (4.37) avec $V = (-1, -2, -5)'$ (haut) et $V = (-1, 0.05, 0)'$ (bas). Les paramètres utilisés sont $\epsilon = 0.05$ (haut), $\epsilon = 0.02$(bas) et $\epsilon' = \sqrt{2\epsilon}$.

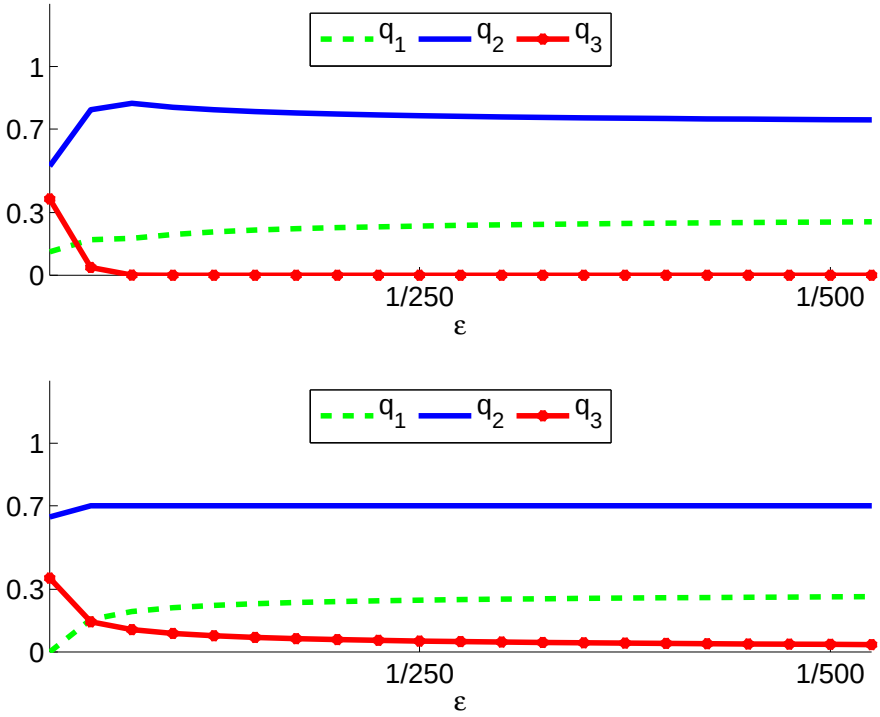

FIGURE 4.12 – Evolution du vecteur de probabilité q qui respectivement maximise l'équation (4.18) (haut) et l'équation (4.37) (bas) avec $p = (0.3, 0.7, 0)'$, $V = (1, 2, 3)'$ en fonction des rayons des voisinages ϵ et ϵ' qui décroissent de $1/2$ à $1/500$

des inégalités de concentration récentes permettant de majorer l'écart entre une distribution et la distribution empirique associée mesuré à l'aide de la divergence de Kullback-Leibler. Il est important de noter que, contrairement à l'algorithme optimiste de [Burnetas and Katehakis, 1997] qui utilise également la divergence de KL, l'algorithme KL-UCRL ne nécessite aucune connaissance sur la structure du MDP. En particulier, l'agent ne connaît pas à l'avance les transitions entre états qui sont impossibles sous le MDP considéré.

Une analyse théorique garantit des bornes supérieures de regret de ces algorithmes qui sont logarithmiques en l'horizon temporel n et polynomiales en le nombre d'états et d'actions. Ces bornes sont similaires à celles du regret accumulé en suivant les algorithmes optimistes de référence que sont UCB (et UCB-V) pour les modèles de bandit et UCRL2 pour les MDP. Nous avons également fourni des bornes de regret qui dépendent de la connectivité du modèle, dans le cas où les récompenses sont déterministes. Si la connectivité est faible, ces bornes sont plus fines que dans le cas général. Les performances pratiques, quant à elles, sont nettement plus grandes en utilisant nos algorithmes que ceux de la littérature. Elles s'expliquent par le fait que la divergence de Kullback-Leibler est plus adaptée à la géométrie des simplexes de probabilité que la norme L^1. La maximisation sur un voisinage KL pour calculer le modèle optimiste permet notamment de gérer de manière pertinente les situations où les probabilités de transition, réelles ou estimées, ont une composante égale à 0. Les performances des algorithmes KL-UCB et KL-UCRL sont en outre dues à la qualité des bornes de concentrations pour la divergence de KL que nous avons utilisées.

Au vu des nettes différences de performances entre les algorithmes optimistes utilisant une distance L^1 et ceux que nous proposons, il semble possible d'obtenir des bornes de regret théoriques plus petites pour nos algorithmes que pour ceux de la littérature. L'analyse théorique que nous avons exposée est similaire à celles proposées pour les algorithmes UCB et UCRL2,

la principale différence étant due à l'utilisation de l'inégalité de Pinsker pour relier la divergence de KL à la distance L^1. Une analyse plus fine utilisant les propriétés de la divergence de KL pourrait vraisemblablement permettre d'obtenir des bornes de regret dépendant non pas de l'écart minimal entre l'espérance de la récompense reçue en jouant la politique optimale et celle reçue en jouant une politique sous-optimale mais d'une constante liée à la divergence de KL. Pour les modèles de bandits, la borne de regret asymptotique d'un algorithme optimiste utilisant la divergence de KL proposée par [Lai and Robbins, 1985] dépend de l'inverse de $\min_a kl(r(a); r(a^*))$. Tandis que celle démontrée par [Burnetas and Katehakis, 1997] dans le cadre des MDP fait apparaître la quantité suivante

$$\sum_{a \in \mathcal{A}:\ a \neq \pi^*(x)} \min_{q_{(x,a)}} KL(P(x,a;.); q_{(x,a)})$$

où $q_{(x,a)}$ est un vecteur de probabilité qui rendrait l'action a optimale : si on remplaçait le vecteur $P(x,a;.)$ par $q_{(x,a)}$ alors l'action a serait l'unique action optimale lorsque l'état de l'environnement est x. Ces deux résultats laissent à penser que les bornes de regret de nos algorithmes pourraient dépendre d'une constante plus petite construite à partir de la divergence de KL plutôt qu'avec la distance L^1 entre les récompenses optimales et sous-optimales. L'analyse asymptotique du regret de l'algorithme proposé par [Honda and Takemura, 2010] permet d'obtenir une telle constante.

ANNEXE A

Inégalités de concentration et théorie de l'information

A.1 Inégalités de concentration utilisant un voisinage L^1

Nous exposons ici toutes les inégalités de concentration que nous utilisons dans cette thèse. Tout d'abord, commençons par présenter l'inégalité d'Hoeffding [Hoeffding, 1963].

Théorème A.1. *Soient $X_1, \ldots, X_n$ des variables aléatoires réelles indépendantes telles que pour tout $i = 1, \ldots, n$ il existe un couple (a_i, b_i) vérifiant $\mathbb{P}(a_i \leq X_i \leq b_i) = 1$. Alors, pour tout $\epsilon > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) > \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

et

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) < -\epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) .$$

Ce théorème est basé sur le lemme suivant

Lemme A.1. *Soit X une variable aléatoire telle que $a \leq X \leq b$. Alors, pour tout $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[\exp\{\lambda X\}] \leq \exp\{\lambda \mathbb{E}[X]\} \exp\left\{\frac{\lambda^2 (b-a)^2}{8}\right\} .$$

Le deuxième résultat est une borne de concentration sur des incréments de martingale. Il s'agit de l'inégalité d'Hoeffding-Azuma [Azuma, 1967]. Une séquence de variables aléatoires $V_1, \ldots, V_n$ est appelée séquence d'incréments de martingales par rapport à une séquence de variables aléatoires $X_1, \ldots, X_n$ si, pour tout $i = 1, \ldots, n$, V_i est une fonction de $X_1, \ldots, X_i$ et

$$\mathbb{E}[V_i \,|\, X_1, \ldots, X_i] = 0 \text{ p.s. .}$$

Théorème A.2. *Soit $V_1, \ldots, V_n$ une séquence d'incréments de martingales par rapport à une séquence de variables aléatoires $X_1, \ldots, X_n$ telle que pour tout $i = 1, \ldots, n$, il existe une variable aléatoire A_i mesurable par rapport à $X_1, \ldots, X_{i-1}$ et une constante positive c_i telle que $V_i \in [A_i, A_i + c_i]$. Alors, pour tout $\epsilon > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n} V_i > \epsilon\right) \leq \exp\left\{-\frac{2\epsilon^2}{\sum_{i=1}^{n} c_i^2}\right)$$

et

$$\mathbb{P}\left(\sum_{i=1}^{n} V_i < -\epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n} c_i^2}\right) .$$

Le résultat suivant est un raffinement de l'inégalité de concentration de Hoeffding-Azuma qui prend en compte la variance des variables aléatoires. Cette inégalité est souvent appelée inégalité de Bernstein pour les incréments de martingales puisque elle découle de l'inégalité de Bernstein [Bernstein, 1945] (voir [Freedman, 1975])

Théorème A.3. *Soit $V_1, \ldots, V_n$ une séquence d'incréments de martingales par rapport à une séquence de variables aléatoires $X_1, \ldots, X_n$ telle que pour tout $i = 1, \ldots, n$, il existe deux constantes positives c et v telles que $|V_i| \leq b$ et $\mathbb{E}\left[V_i^2 \,\middle|\, \mathcal{F}_{i-1}\right] \leq v$. Alors, pour tout $\epsilon > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n} V_i > \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2nv + 2c\epsilon/3}\right)$$

et pour tout $\delta > 0$, avec probabilité $1 - \delta$, on a

$$\sum_{i=1}^{n} V_i \leq \sqrt{2nv \log(1/\delta)} + \frac{c \log(1/\delta)}{3} .$$

[Aud, 2008] ont prouvé qu'il était possible de remplacer la vraie variance par la variance empirique dans le théorème ci-dessus.

Théorème A.4. *Soit $X_1, \ldots, X_n$ une séquence de variables aléatoires indépendantes, identiquement distribuées et centrée telles que pour tout $i = 1, \ldots, n$, il existe une constante positive c telle que $o \leq X_i \leq b$. Alors, pour tout $\delta > 0$ et tout entier $t \leq n$*

$$\sum_{i=1}^{t} W_i \leq \sqrt{2nW_t \log(3/\delta)} + 3 \log(3/\delta)$$

où

$$W_t = \frac{1}{t} \sum_{i=1}^{t} (X_i - \frac{1}{t} \sum_{j=1}^{t} X_i)^2 .$$

L'inégalité suivante permet de majorer la norme L^1 de la différence entre une probabilité estimée et la vraie probabilité engendrant les observations. Ce théorème a été prouvé par [Weissman et al., 2003].

Théorème A.5. *Soit $X_1, \ldots, X_n$ une séquence de variables aléatoires indépendantes, identiquement distribuées selon une probabilité p sur $\{1, \cdots, N\}$. Soit $\hat{p}$ la probabilité empirique définie pour tout $i \in \{1, \cdots, N\}$ par*

$$\hat{p}_i = \frac{\sum_{k=1}^{n} \mathbb{1}_{\{X_k = i\}}}{n} .$$

Pour tout $\epsilon > 0$,

$$\mathbb{P}\left(\|p - \hat{p}\|_1 > \epsilon\right) < (2^N - 2) \exp\left\{-\frac{n\phi(C_p))\epsilon^2}{4}\right\} ,$$

où $\phi : p \to \frac{1}{1-2p} \log\left(\frac{1-p}{p}\right)$ et $c_p = \max_i \min\{p_i, 1 - p_i\}$.

La sixième inégalité de concentration que nous présentons concerne des moyennes auto-normalisées par un entier aléatoire. Soit $(X_t)_t$ une séquence de variables aléatoires indépendantes et identiquement distribuées, positives majorées par $X_{\max}$. Pour tout $s > t$, la variable X_s est supposée indépendante de la sigma-algèbre $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$. Soit une séquence de variables de Bernoulli $(\xi_t)_t$ telle que pour tout $t > 0$, ξ_t est $\mathcal{F}_{t-1}$ mesurable.

Théorème A.6. *Pour tout entier t et $\epsilon > 0$,*

$$\mathbb{P}\left[\frac{\sum_{k=1}^{t}(X_k - \mathbb{E}[X_k])\xi_k}{\sqrt{\sum_{k=1}^{t}\xi_k}} > \epsilon\right] \leq \left\lceil \frac{\log(t)}{\log(1+\eta)} \right\rceil \exp\left(-\frac{2\epsilon^2}{X_{\max}^2}\left(1 - \frac{\eta^2}{16}\right)\right)$$

pour tout $\eta > 0$. Pour $\eta = 0.3$, on a

$$P\left[\frac{\sum_{k=1}^{t}(X_k - \mathbb{E}[X_k])\xi_k}{\sqrt{\sum_{k=1}^{t}\xi_k}} > \epsilon\right] \leq 4\log(t)\exp\left(-\frac{1.99\epsilon^2}{X_{\max}^2}\right) . \tag{A.1}$$

Ce théorème a été prouvé dans [Garivier and Moulines, 2008] pour un cas plus compliqué où la somme des variables aléatoire peut être pondérée par un facteur $0 < \gamma \leq 1$. Nous avons présenté ici la version qui nous intéresse dans cette thèse. Les auteurs prouvent également une version plus forte du théorème précédent majorant la probabilité que, à chaque instant t, la valeur moyenne des variables s'éloignent de leur espérance. C'est ce que l'on appelle une borne uniforme en temps. Ils montrent que, pour tout entier positif n et pour tout $\delta > 0$,

$$\mathbb{P}\left[\sup_{1\leq t\leq n}\frac{\sum_{k=1}^{t}(X_k - \mathbb{E}[X_k])\xi_k}{\sqrt{\sum_{k=1}^{t}\xi_k}} > \epsilon\right] \leq 4\log(t)\exp\left(-\frac{1.99\epsilon^2}{X_{\max}^2}\right) .$$

L'inégalité de concentration proposée par [Glynn and Ormoneit, 2002] concerne des chaînes de Markov ergodiques.

Théorème A.7. *Soit $(X_n)_{n\geq 0}$ une chaîne de Markov à valeur dans $\mathcal{X}$. On suppose qu'il existe une mesure Φ sur $\mathcal{X}$, un réel strictement positif λ et un entier m tels que $\mathbb{P}(X_m \in .) \geq \lambda\Phi(.)$. Soit f une fonction de $\mathcal{X}$ dans $\mathbb{R}$ et S_n la variable aléatoire définie par $S_n = \sum_{t=0}^{n-1} f(X_t)$. Supposons que $\|f\|_\infty \stackrel{\text{def}}{=} \sup\{f(x),\ x \in \mathcal{X}\} < \infty$. On a pour tout $\epsilon > 0$ et pour tout $n > 2\|f\|_\infty m/(\lambda\epsilon)$*

$$\mathbb{P}(S_n - \mathbb{E}[S_n] > \epsilon n) \leq \exp\left\{-\frac{\lambda^2(n\epsilon - 2\|f\|_\infty m/\lambda)^2}{2n\|f\|_\infty^2 m^2}\right\} .$$

Une autre inégalité exponentielle que nous utilisons dans cette thèse mais qui n'est pas une inégalité de concentration est celle démontrée par [De La Pena et al., 2004] dans un article traitant de processus auto-normalisés. Cette inégalité (corollaire 2.2 [De La Pena et al., 2004]) ne fait pas directement apparaître une somme de termes d'un processus.

Théorème A.8. *Soient A et B deux variables aléatoires telles que $B \geq 0$ et que*

$$E\left[\exp\left\{\lambda A - \frac{\lambda}{2}B^2\right\}\right] \leq 1 \quad \textit{pour tout } \lambda \in \mathbb{R} . \tag{A.2}$$

Alors, pour tout $c \geq \sqrt{2}$, et tout $y > 0$

$$\mathbb{P}\left(|A| \geq c\sqrt{(B^2+y)\left(1 + \frac{1}{2}\log\left(\frac{B^2}{y}+1\right)\right)}\right) \leq \exp\{-\frac{c^2}{2}\} . \tag{A.3}$$

A.2 Propriété de la divergence de Kullback-Leibler

Nous rappelons quelques définitions élémentaires de la théorie de l'information (voir [Cover and Thomas]). La divergence de Kullback-Leibler (KL) entre deux vecteurs de probabilités p, q de dimension N est définie par

$$KL(p;q) \stackrel{\text{def}}{=} \sum_{i=1}^{N} p_i \log\left(\frac{p_i}{q_i}\right) .$$

Dans le cas de variables de Bernoulli de paramètres p_0 et q_0, on note kl la divergence de KL entre les deux vecteurs de probabilité $p = (p_0, 1-p_0)'$ et $q = (q_0, 1-q_0)'$:

$$kl(p_0;q_0) \stackrel{\text{def}}{=} KL(p;q) = p_0 \log\left(\frac{p_0}{q_0}\right) + (1-p_0)\log\left(\frac{1-p_0}{1-q_0}\right) .$$

Pour tout vecteur de probabilité p fixé, la fonction $q \to KL(p;q)$ est convexe.

La proposition suivante relie la divergence de KL et la distance L^1 entre deux vecteurs de probabilité. Il s'agit de l'inégalité de Pinsker [Cover and Thomas].

Proposition A.1. *Pour tous vecteurs de probabilité p et q,*

$$\|p-q\|_1 \leq \sqrt{2KL(p;q)} . \tag{A.4}$$

Elle peut également être liée à la distance du Chi-2 [Csiszár and Talata, 2006] :

Proposition A.2. *Pour tous vecteurs de probabilité p et q,*

$$KL(p;q) \leq \sum_{i=1}^{n} \frac{(p_i-q_i)^2}{q_i} . \tag{A.5}$$

Il est intéressant de remarquer que cette inégalité est large lorsque p est dans un voisinage de q. En effet, dans ce cas, en faisant un développement de Taylor, on obtient que

$$KL(p;q) = \frac{1}{2}\sum_{i=1}^{n} \frac{(p_i-q_i)^2}{q_i} + o\left(\|p-q\|_2^2\right) .$$

Nous nous intéressons dans cette thèse à des inégalités de déviation de la divergence de Kullback-Leibler entre deux probabilités. Soit p une probabilité sur un espace fini $\mathcal{X}$ et soit $\hat{p}_t$ un estimateur de cette probabilité à l'instant t défini par

$$\hat{p}_t = \frac{\sum_{k=1}^{t} X_t \epsilon_t}{\sum_{k=1}^{t} \epsilon_t} ,$$

où ϵ_t est une variable aléatoire appartenant à $\{0,1\}$. On notera dans la suite $N_t = \sum_{k=1}^{t} \epsilon_t$. Le théorème suivant a été prouvé par [Garivier and Leonardi, 2010]

Théorème A.9. *Pour tout entier t et pour tout $\delta > 0$,*

$$\mathbb{P}\left(N_t KL(\hat{p}_t;p) > \delta\right) \leq 2e\,(\delta\log(t) + |\mathcal{X}|)\exp\left(-\frac{\delta}{|\mathcal{X}|}\right) .$$

Bibliographie

Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 2008.

N. Abe and A. Nakamura. Learning to Optimally Schedule Internet Banner Advertisements. *International Conference on Machine Learning*, pages 12–21, 1999.

D. Aberdeen. *Policy-gradient algorithms for partially observable Markov decision processes.* PhD thesis, Australian National University, 2003a.

D. Aberdeen. A (revised) survey of approximate methods for solving partially observable Markov decision processes. *National ICT Australia, Canberra, Australia, Tech. Rep*, 2003b.

D. Aberdeen and J. Baxter. Scaling internal-state policy-gradient methods for POMDPs. *International Conference on Machine Learning*, pages 3–10, 2002.

R. Agrawal. Sample mean based index policies with O (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4) :1054–1078, 1995.

I. F. Akyildiz, L. Won-Yeol, M. C. Vuran, and S. Mohanty. A survey on spectrum management in cognitive radio networks. *IEEE Communications Magazine*, 46(4) :40–48, 2008.

E. Altman. *Constrained Markov decision processes.* Chapman & Hall, 1999.

V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays. I : i.i.d rewards. *IEEE Transaction on Automatic and Control*, 32 :977–982, 1987a.

V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays. II : Markovian rewards. *IEEE Transaction on Automatic and Control*, 32 :977–982, 1987b.

A. Antos, C. Szepesvári, and R. Munos. Value-iteration based fitted policy iteration : learning with a single trajectory. *IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning, 2007. ADPRL 2007*, pages 330–337, 2007.

K. Astrom. Optimal control of Markov decision processes with incomplete state estimation. *Journal of Mathematical Analysis and Applications*, 10 :174–205, 1965.

J. Audibert. PAC-Bayesian aggregation and multi-armed bandits. 2010.

J. Audibert, R. Munos, and C. Szepesvári. Tuning bandit algorithms in stochastic environments. *Lecture Notes in Computer Science*, 4754 :150, 2007.

P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3 :397–422, 2002.

P. Auer and R. Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in Neural Information Processing Systems*, page 49, 2007.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2) :235–256, 2002.

P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in Neural Information Processing Systems*, 21, 2009a.

P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning (full version). Technical report, URL : http ://institute.unileoben.ac.at/infotech/publications/ ucrl2.pdf., 2009b.

K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19(3) :357–367, 1967.

L. Baird. Residual algorithms : Reinforcement learning with function approximation. *International Machine Learning Conference*, pages 30–37, 1995.

P. Bartlett and A. Tewari. REGAL : A Regularization based Algorithm for Reinforcement Learning in Weakly Communicating MDPs. *Annual Conference on Uncertainty in Artificial Intelligence*, 2009.

J. Baxter and P. L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15 :319–350, 2001.

J. Baxter, P. L. Bartlett, and L. Weaver. Experiments with infinite-horizon, policy-gradient estimation. *Journal of Artificial Intelligence Research*, 2001.

R. Bellman. A problem in the sequential design of experiments. *Sankhyā : The Indian Journal of Statistics*, pages 221–229, 1956.

S. Bernstein. *Theory of probability.* 1945.

D. Bertsekas. *Dynamic Programming and Optimal Control, Two Volume Set.* Athena Scientific, 1995.

D. Bertsekas and J. Tsitsiklis. *Neuro-dynamic programming.* 1996.

D. Bertsimas and J. Niño-Mora. Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research*, 48(1) :80–90, 2000.

P. Billingsley. *Probability and measure.* John Wiley & Sons, 1979.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Advances in Neural Information Processing Systems*, 14 :601–608, 2002.

B. Bonet. An e-optimal grid-based algorithm for partially obserable Markov decision processes. *International Conference on Machine Learning*, 2002.

J. Boyan. Technical update : Least-squares temporal difference learning. *Machine Learning*, 49(2) :233–246, 2002.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge Univ Pr, 2004.

S. Bradtke and A. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22(1) :33–57, 1996.

R. Brafman and M. Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3 :213–231, 2003.

R. Brafman, G. Shani, and S. Shimony. Partial observability under noisy sensors ?from model-free to model-based. *Workshop on Rich Representations for Reinforcement Learning*, 2005.

S. Bubeck. *Jeux de bandits et fondations du clustering*. PhD thesis, 2010.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. *International conference on Algorithmic learning theory*, pages 23–37, 2009a.

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvári. Online optimization in X-armed bandits. *Advances in Neural Information Processing Systems*, 21 :201–208, 2009b.

A. Burnetas and M. Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2) :122–142, 1996. ISSN 0196-8858.

A. Burnetas and M. Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, pages 222–255, 1997.

O. Cappé, E. Moulines, and T. Rydén. *Inference in hidden Markov models*. Springer Verlag, 2005.

G. Casella and R. Berger. *Statistical inference*. Thomson Brooks/Cole, 1990.

A. Cassandra, M. Littman, and N. Zhang. Incremental pruning : A simple, fast, exact method for partially observable Markov decision processes. *Conference on Uncertainty in Artificial Intelligence*, pages 54–61, 1997.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge Univ Pr, 2006.

K. Chen, I. Hu, and Z. Ying. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *Annals of Statistics*, 27(4) : 1155–1163, 1999.

Y. Chen, Q. Zhao, and A. Swami. Joint design and separation principle for opportunistic spectrum access in the presence of sensing errors. *IEEE Transactions on Information Theory*, 54(5) :2053–2071, 2008.

L. Chrisman. Reinforcement learning with perceptual aliasing : the perceptual distinctions approach. *Conference on Artificial Intelligence*, pages 183–188, 1992.

T. Cover and J. Thomas. *Elements of information theory*. Wiley-Interscience, New York.

I. Csiszár and Z. Talata. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Transactions on Information Theory*, 52(3) :1007, 2006.

V. Dani, T. Hayes, and S. Kakade. Stochastic linear optimization under bandit feedback. *Conference on Learning Theory*, 2008.

V. De La Pena, M. Klass, and T. Lai. Self-normalized processes : exponential inequalities, moment bounds and iterated logarithm laws. *Annals of Probability*, 32(3) :1902–1933, 2004.

L. Dorard, D. Glowacka, and J. Shawe-Taylor. Gaussian process modelling of dependencies in multi-armed bandit problems. *International Symposium on Operation Research*, 2009.

A. Dutech and M. Samuelides. *Revue d'Intelligence Artificielle, RIA*, 17(4), 2003.

D. Ernst, P. Geurts, and L. Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(1) :503, 2006.

E. Even-Dar, S. Mannor, and Y. Mansour. PAC bounds for multi-armed bandit and Markov decision processes. *Computational Learning Theory*, pages 193–209, 2002.

E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7 :1079–1105, 2006.

E. Feinberg and A. Shwartz. *Handbook of Markov decision processes : methods and applications*. Kluwer Academic, 2002.

F. Force. Report of the spectrum efficiency working group. *Federal Communications Commision, Technical Report*, pages 02–155, 2002.

D. Freedman. On tail probabilities for martingales. *Annals of Probability*, 3(1) :100–118, 1975.

E. Frostig and G. Weiss. Four proofs of gittins multiarmed bandit theorem. *Applied Probability Trust*, pages 1–20, 1999.

A. Garivier and F. Leonardi. Context tree selection : A unifying view. preprint, 2010.

A. Garivier and E. Moulines. On upper-confidence bound policies for non-stationary bandit problems. *Arxiv preprint arXiv :0805.3415*, 2008.

J. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, 41(2) :148–177, 1979.

K. Glazebrook, J. Niño-Mora, and P. Ansell. Index policies for a class of discounted restless bandits. *Advances on Applied Probability*, 34 :754–774, 2002.

P. Glynn and D. Ormoneit. Hoeffding's inequality for uniformly ergodic Markov chains. *Statistics and Probability Letters*, 56(2) :143–146, 2002.

E. Greensmith, P. L. Bartlett, and J. Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Jounral of Machine Learning Research*, 5, 2004.

S. Guha and K. Munagala. Approximation algorithms for partial-information based stochastic control with Markovian rewards. *IEEE Symposium on Foundations of Computer Science*, pages 483–493, 2007.

S. Guha, K. Munagala, and P. Shi. On index policies for restless bandit problems. *Arxiv preprint arXiv :0711.3861*, 2008.

E. Hansen. Solving POMDPs by searching in policy space. *Conference on Uncertainty in Artificial Intelligence*, pages 211–219, 1998.

M. Hauskrecht. Value-function approximations for partially observable Markov decision processes. *Journal of Artificial Intelligence Research*, 13 :33–94, 2000.

S. Haykin. Cognitive radio : Brain-empowered wireless communications. *IEEE Journal of Selected Areas in Communications*, 23(2) :201–220, 2005.

A. O. Hero, D. A. Castanon, D. Cochran, and K. Kastella, editors. *Foundations and Applications of Sensor Management*. Springer, 2008.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301) :13–30, 1963.

J. Honda and A. Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *The Twenty-Third Annual Conference on Learning Theory, COLT*, 2010.

R. Horn and C. Johnson. *Matrix analysis*. Cambridge Univ Pr, 1990.

D. Hsu, W. Lee, and N. Rong. What makes some POMDP problems easy to approximate. *Advances in Neural Information Processing Systems*, 2007.

T. Jaakkola, S. Singh, and M. Jordan. Reinforcement learning algorithm for partially observable Markov decision problems. *Advances in Neural Information Processing Systems*, 7 : 345–352, 1995.

T. Jaksch, R. Ortner, and P. Auer. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11 :1563–1600, 2010.

W. Jank and G. Shmueli. *Statistical Methods in E-commerce Research*. Wiley-Interscience, 2008.

L. Kaelbling, M. Littman, and A. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1) :99–134, 1996.

S. Kakade, S. Shalev-Shwartz, and A. Tewari. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th International Conference on Machine learning*, pages 440–447. ACM, 2008.

S. Kalyanakrishnan and P. Stone. Batch reinforcement learning in a complex domain. *International joint conference on Autonomous agents and multiagent systems*, pages 1–8, 2007.

M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. *Journal of machine learning*, 49(2-3) :209–232, 2002.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces. *ACM symposium on Theory of computing*, pages 681–690, 2008.

L. Lai, H. El Gamal, H. Jiang, and H. Vicent Poor. Optimal medium access protocols for cognitive radio networks. *International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks and Workshops*, 2008.

T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1) :4–22, 1985.

J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in Neural Information Processing Systems*, pages 817–824, 2008.

J. Le Ny, M. Dahleh, and E. Feron. Multi-UAV dynamic routing with partial observations using restless bandit allocation indices. *American Control Conference*, pages 4220–4225, 2008.

L. Li, W. Chu, J. Langford, and R. Schapire. A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

K. Liu and Q. Zhao. A restless bandit formulation of opportunistic access : Indexablity and index policy. *Annual Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks Workshops*, pages 1–5, 2008.

K. Liu and Q. Zhao. Decentralized multi-armed bandit with multiple distributed players. *Information Theory and Applications*, 2010a.

K. Liu and Q. Zhao. Indexability of restless bandit problems and optimality of whittle's index for dynamic multichannel access. *to appear in IEEE Transactions on Information Theory*, 2010b.

X. Long, X. Gan, Y. Xu, J. Liu, and M. Tao. An estimation algorithm of channel state transition probabilities for cognitive radio systems. In *Cognitive Radio Oriented Wireless Networks and Communications*, 2008.

S. Mannor and J. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *The Journal of Machine Learning Research*, 5 :648, 2004.

A. McCallum. *Reinforcement Learning with Selective Perception and Hidden State*. PhD thesis, University of Rochester, 1996.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, 1989.

N. Meuleau, K. Kim, L. Kaelbling, and A. Cassandra. Solving POMDP s by searching the space of finite policies. *Conference on Uncertainty in Artificial Intelligence*, pages 417–426, 1999.

J. Mitola. *Cognitive Radio - An Integrated Agent Architecture for Software Defined Radio*. PhD thesis, Royal Institute of Technology, Kista, Sweden, May 8 2000.

J. Mitola III and G. Maguire Jr. Cognitive radio : making software radios more personal. *IEEE personal communications*, 6(4) :13–18, 1999.

V. Mnih, C. Szepesvári, and J. Audibert. Empirical Bernstein stopping. *International Conference on Machine learning*, pages 672–679, 2008.

R. Munos. Error bounds for approximate policy iteration. *International Conference in Machine Learning*, 20(2) :560, 2003.

R. Munos and C. Szepesvári. Finite-time bounds for fitted value iteration. *The Journal of Machine Learning Research*, 9 :815–857, 2008.

A. Nilim and L. El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5) :780–798, 2005. ISSN 0030364X.

J. Niño-Mora. Restless bandits, partial conservation laws and indexability. *Advances in Applied Probability*, 33(1) :76–98, 2001.

R. Ortner. Exploiting Similarity Information in Reinforcement Learning. Similarity Models for Multi-Armed Bandits and MDPs. 2010.

S. Pandey, D. Agarwal, D. Chakrabarti, and V. Josifovski. Bandits for taxonomies : A model-based approach. *International Conference on Data Mining*, 2007a.

S. Pandey, D. Chakrabarti, and D. Agarwal. Multi-armed bandit problems with dependent arms. *International Conference on Machine learning*, pages 721–728, 2007b.

C. Papadimitriou. *Computational complexity.* John Wiley and Sons Ltd., 2003.

C. Papadimitriou and J. Tsitsiklis. The complexity of optimal queueing network control. *Structure in Complexity Theory Conference*, pages 318–322, 1994.

J. Pineau, G. Gordon, and S. Thrun. Anytime point-based approximations for large POMDP s. *Journal of Artificial Intelligence Research*, 27 :335–380, 2006.

P. Poupart and C. Boutilier. Bounded finite state controllers. *Advances in Neural Information Processing Systems*, 16, 2004a.

P. Poupart and C. Boutilier. Vdcbpi : an approximate scalable algorithm for large POMDP s. *Advances in Neural Information Processing Systems*, 17 :1081–1088, 2004b.

M. Puterman. *Markov Decision Processes : Discrete Stochastic Dynamic Programming.* John Wiley & Sons, Inc. New York, NY, USA, 1994.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of American Mathematical Society*, 58(5) :527–535, 1952.

P. Rusmevichientong and J. Tsitsiklis. Linearly parameterized bandits. Arxiv preprint arXiv :0812.3465, 2008.

P. Rusmevichientong, A. Mersereau, and J. Tsitsiklis. A Structured Multiarmed Bandit Problem and the Greedy Policy. *IEEE Transactions on Automatic Control*, 54 :2787–2802, 2009.

A. Schein, L. Saul, and L. Ungar. A generalized linear model for principal component analysis of binary data. *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, pages 14–21, 2003.

G. Shani. A survey of model-based and model-free methods for resolving perceptual aliasing. Technical report, Department of Computer Science at the Ben-Gurion University in the Negev, November 2004.

G. Shani, R. Brafman, and S. Shimony. Model-based online learning of POMDPs. *European Conference on Machine Learning*, 2005.

O. Sigaud and O. Buffet. *Processus décisionnels de Markov en intelligence artificielle.* Lavoisier - Hermes Science Publications, 2008.

S. Singh, T. Jaakkola, M. Littman, and C. Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38(3) :287–308, 2000.

S. P. Singh, T. Jaakkola, and M. I. Jordan. Learning without state-estimation in partially observable Markovian decision processes. *International Conference on Machine Learning*, pages 284–292, 1994.

A. Slivkins. Contextual bandits with similarity information. *Arxiv preprint arXiv*, 907, 2009.

R. Smallwood and E. Sondik. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research*, 21(5) :1071–1088, 1973.

E. Sondik. The optimal control of partially observable Markov processes over the infinite horizon : Discounted costs. *Operations Research*, 26(2) :282–304, 1978.

A. Strehl and M. Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8) :1309–1331, 2008.

R. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3 (1) :9–44, 1988.

R. Sutton and A. Barto. *Reinforcement learning : An introduction*. The MIT press, 1998.

R. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12 :1057–1063, 2000.

C. Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan Claypool Publisher, 2010.

C. Szepesvári and W. Smart. Interpolation-based q-learning. In *ICML '04 : Proceedings of the twenty-first international conference on Machine learning*, page 100, New York, NY, USA, 2004. ACM. ISBN 1-58113-828-5. doi : http://doi.acm.org/10.1145/1015330.1015445.

I. Szita and A. Lőrincz. The many faces of optimism : a unifying approach. *International conference on Machine learning*, pages 1048–1055, 2008.

A. Tewari and P. Bartlett. Bounded parameter Markov decision processes with average reward criterion. *Lecture Notes in Computer Science*, 4539 :263, 2007.

A. Tewari and P. Bartlett. Optimistic linear programming gives logarithmic regret for irreducible MDPs. *Advances in Neural Information Processing Systems*, 20 :1505–1512, 2008.

S. Thrun. Monte carlo POMDP s. *Advances in Neural Information Processing Systems*, 12 : 1064–1070, 2000.

M. Toussaint and A. Storkey. Probabilistic inference for solving discrete and continuous state Markov Decision Processes. *international conference on Machine learning*, page 952, 2006.

B. Van Roy. Performance loss bounds for approximate value iteration with state aggregation. *Mathematics of Operations Research*, 31(2) :234–244, 2006.

C. Wang, S. Kulkarni, and H. Poor. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3) :338–355, 2005.

C. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8(3) :279–292, 1992.

R. Weber and G. Weiss. On an index policy for restless bandits. *Journal of Applied Probability*, 27(3) :637–648, 1990.

T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. Weinberger. Inequalities for the L1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

P. Whittle. Restless bandits : Activity allocation in a changing world. *Journal of Applied Probability*, 25 :287–298, 1988.

Q. Zhao and B. Sadler. A survey of dynamic spectrum access. *IEEE Signal Processing Magazine*, 24(3) :79–89, 2007.

Q. Zhao, L. Tong, and A. Swami. Decentralized cognitive MAC for dynamic spectrum access. *IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks*, pages 224–232, 2007a.

Q. Zhao, L. Tong, A. Swami, and Y. Chen. Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks : A POMDP framework. *IEEE Journal on Selected Areas in Communications*, 25(3) :589–600, 2007b.

Q. Zhao, B. Krishnamachari, and K. Liu. On myopic sensing for multi-channel opportunistic access : Structure, optimality, and performance. *IEEE Transactions on Wireless Communications*, 7(12) :5431–5440, 2008.

Table des figures

Liste des algorithmes