Mémoire

présenté

devant l'Université Paris 11

pour obtenir

l'habilitation à diriger des recherches
en Mathématiques

par

Aurélien Garivier

Laboratoire d'accueil : LTCI, UMR 5141, Telecom ParisTech
École Doctorale : Mathématiques
Composante universitaire : Laboratoire de Mathématiques d'Orsay

Titre du mémoire :

Analyses d'algorithmes pour l'estimation et l'optimisation stochastiques

À soutenir le 28 novembre 2011 devant la commission d'examen

M. :	Eric	Moulines	Président
MM. :	Olivier	Catoni	Rapporteurs
	Eva	Löcherbach	
	Pascal	Massart	
MM. :	Stéphane	Boucheron	Examinateurs
	Olivier	Cappé	
	Rémi	Munos	
	Elisabeth	Gassiat	

Remerciements

Je remercie Pascal Massart, qui me fait l'honneur de présider ce jury et qui a bien voulu rapporter mon dossier en un temps record. Je remercie également mes deux rapporteurs externes, Olivier Catoni et Eva Löcherbach, pour la qualité de leurs rapports et la vitesse avec laquelle ils les ont produits. Je dois à Elisabeth Gassiat, et à Eric Moulines, bien plus que des remerciements pour avoir bien voulu juger de ce travail : il ont été, l'un après l'autre, les deux tuteurs qui m'ont appris le métier de chercheur, et qui m'ont aidé à prendre ma place dans ce monde. Il en va de même pour Stéphane Boucheron et Olivier Cappé : le premier m'a aidé à débuter, le second m'encourage à prendre de nouvelles responsabilités. Je ne peux que souhaiter de continuer à apprendre à leur contact.

Grâce, avant tout, à l'énergie et au talent incomparables d'Eric, l'équipe STA du LTCI a été pour moi un environnement particulièment riche et épanouissant pendant ces quatre dernières années. J'en remercie tous les membres, et notamment Cédric Févotte, Céline Lévy-Leduc, François Roueff, Gersende Fort, Jamal Najim, Jérémie Jakubowicz, Pascal Bianchi et Stéphan Clémençon : les nombreux contacts que j'ai eus avec eux, que ce soit pour la recherche, pour l'enseignement, ou même au delà, m'auront été aussi utiles qu'agréables. Ces derniers temps, c'est tout particulièrement avec Olivier que j'ai travaillé : que ce soit pour la rédaction d'articles, les preuves, les calculs, les algorithmes, l'encadrement ou même la gestion d'équipe, il m'a aidé à progresser et à trouver ma voie. Je remercie bien sûr Sarah Filippi et Emilie Kaufmann, les deux doctorantes dont j'ai partagé la responsabilité et avec qui je continue d'apprendre : leurs qualités les promettent à un superbe avenir.

En dehors de l'équipe, c'est à la force de la communauté française d'apprentissage [1] que je dois d'avoir pu donner à mes travaux leur orientation récente. Je commence avec Rémi Munos et Sébastien Bubeck des collaborations que j'espère longues et fructueuses. Je suis particulièrement reconnaissant à Gilles Stoltz, camarade aussi agréable que compétent et dynamique, des occasions diverses et variées qui nous auront amenés à oeuvrer ensemble. J'ai eu la chance de voir travailler Randal Douc, inlassablement brillant, et Matthieu Lerasle qui jongle allègrement avec les pires difficultés techniques. Ma participation à un projet au long cours avec l'université de Sao Paulo m'a procuré le plaisir de rencontrer Florencia Leonardi et Antonio Galves. Ma collaboration avec Sylvain Arlot, bien qu'elle n'ait pas porté sur des projets de recherche, m'aura montré l'exemple exceptionnel d'une formidable puissance de travail associée à une sérénité inébranlable et à une gentillesse jamais démentie. Il en va de même pour Olivier Wintenberger : bien qu'ayant peu travaillé avec lui, j'ai à de nombreuses reprises apprécié autant sa finesse mathématique que ses diverses qualités humaines.

1. celle-ci a pu être appréciée lors des dernières éditions des conférences COLT et ALT.

Je remercie tout particulièrement ceux qui, par leur relecture, ont permis à ce mémoire d'exister malgré l'urgence dans laquelle il a été rédigé.

Au delà de ces quelques collègues, j'ai beaucoup profité de la grande richesse de la vie scientifique parisienne dans son ensemble : le contact permanent de tous ses membres constitue un foisonnement aussi sympathique que stimulant. Inutile de tous les citer : Adeline, Antoine, Arnak, Christian, Claudine, Damien, Dominique, Eric, Francis, François, Gérard, Guillaume, Ismaël, Jean-Yves, Johann, Judith, Karine, Nicolas, Sophie, Stéphanie, Vianney... Je suis heureux de pouvoir côtoyer dans mon métier des personnes aussi agréables et passionnantes, qui parfois deviennent des amis. Bien sûr, je n'aurais jamais pu travailler sereinement sans l'appui et la compréhension d'Elena et de nos deux fils Thomas et Raphaël : mon travail restera pour eux bien mystérieux pendant encore assez longtemps, mais cela ne les empêche pas d'excuser les retards et les absences qu'il occasionne parfois. Merci enfin à tous les membres de la famille, à côté desquels il est si plaisant de construire, petit à petit, notre route.

Table des matières

Chapitre 1

Introduction

1.1 Présentation

Ce mémoire expose de manière synthétique l'essentiel des travaux que j'ai effectués au cours des quatre dernières années comme chargé de recherche CNRS au sein de l'équipe STA du Laboratoire Traitement et Communication de l'Information. Ces travaux présentent une certaine diversité thématique et méthodologique ; l'occasion m'est donnée d'en esquisser une ligne directrice.

Je m'intéresse, de façon générale, à la conception et à l'analyse de stratégies et d'algorithmes pour la résolution de problèmes d'estimation ou d'optimisation en environnement stochastique. Dans les différents modèles que j'ai considérés (problèmes de bandits, processus de décision markoviens, modèles à mémoire variable ou à variables latentes), je me suis attaché à proposer des solutions numériquement efficaces à des problèmes motivés par des applications réelles, ainsi qu'à la preuve de garanties théoriques solides pour ces solutions. C'est donc dans l'articulation entre algorithmique et mathématiques qu'ont porté la plupart de mes efforts : il s'agit d'utiliser certaines techniques statistiques récentes à l'intérieur de procédures opératoires, et de savoir analyser des algorithmes d'inspiration heuristique à l'aide des bons outils probabilistes.

Une illustration simple et significative de cette approche d'interface se trouve dans l'étude que je propose avec Olivier Cappé du modèle le plus élémentaire envisagé en apprentissage par renforcement, celui des bandits multi-bras. En termes statistiques, le travail semble trivial : il s'agit uniquemement d'estimer l'espérance de variables aléatoires indépendantes, identiquement distribuées et bornées. Pour simple qu'il paraisse, encore faut-il accomplir correctement cet exercice, en prenant en compte les objectifs et les contraintes spécifiques au problème. Une étude fine de l'heuristique dite "optimiste" nous a conduits à proposer un algorithme universellement plus efficace que la référence antérieure, et à montrer qu'il satisfait une propriété d'optimalité. Nos autres travaux sur les modèles de bandits,

avec Sarah Filippi, contribuent à généraliser cette approche à des modèles plus riches, prenant en compte la complexité des situations réalistes où ils sont utilisés (absence de borne sur les récompenses, non stationnarité, paramétrage des bras). Nous avons enfin étendu notre analyse des méthodes optimistes au champ beaucoup plus large des processus de décision markoviens, fournissant en contrepartie une analyse beaucoup moins précise : ces travaux sont présentés dans le chapitre 2, qui se termine par l'étude d'un problème d'exploration ne relevant pas réellement de l'apprentissage par renforcement, mais pour lequel je propose avec Sébastien Bubeck une solution originale directement inspirée de l'heuristique optimiste.

Grâce à l'expertise d'Eric Moulines et de Randal Douc, j'ai également été initié à la mise en oeuvre et à l'étude des méthodes de filtres particulaires pour les modèles de Markov cachés. Outre l'obtention, pour une famille assez générale de filtres, de bornes théoriques quantifiant la consistance des méthodes de Monte-Carlo séquentielles, nous avons travaillé à la mise en oeuvre de méthodes particulaires spécifiques au problème étudié par Steffen Barembruch de déconvolution aveugle pour les canaux à modulation linéaire : le chapitre 3 expose ces travaux.

Le troisième axe de mes travaux est l'étude des modèles probabilistes de "à arbres de contextes", qui étendent les potentialités modélisatrices des chaînes de Markov tout en conservant une certaine simplicité opératoire. Deux aspects distincts m'ont particulièrement intéressé : le problème de la simulation parfaite, pour lequel je propose un algorithme généralisant le couplage par le passé de Propp et Wilson ; et le problème de l'estimation, sur lequel j'ai travaillé dans le cadre d'une collaboration franco-brésilienne avec l'université de Sao Paulo impliquant également Elisabeth Gassiat. Les résultats auxquels nous sommes parvenus sont évoqués dans le chapitre 4.

Ce mémoire ne reprend pas de façon exhaustive toutes les pistes que j'ai peu ou prou explorées ces dernières années, grâce aux formidables environnements scientifiques que constituent l'équipe STA, le LTCI, et plus généralement le grand campus qu'est la région parisienne. Il ne fait pas non plus référence aux travaux qui se rattachent à la queue de la comète de ma thèse [Gar06b]. Mais il rassemble des résultats qui donneront, je l'espère, une bonne idée de mes principaux centres d'intérêt. Pour lui conserver une certaine unité de ton, et pour ne pas obscurcir inutilement cet exposé de trop lourdes notations, j'ai choisi de ne pas présenter ces résultats de façon trop formelle, préférant plutôt les mettre en perspective, en insistant sur mes motivations et en privilégiant l'esprit à la lettre. Les détails sont disponibles dans les articles, auxquels il est fréquemment fait référence de façon précise (avec le numéro de la section ou du théorème concerné). Avec le même objectif, j'ai conservé autant que possible une cohérence de notations qui m'a parfois m'éloigné de celles qui sont utilisées dans les articles cités.

Mais avant d'entamer cette description, j'ai saisi l'occasion qui m'est offerte ici pour rassembler quelques éléments d'utilisation récurrente dans mes travaux, portant sur l'étude des déviations auto-normalisées de certaines martingales à incréments contrôlés, et sur son application en estimation. Ces éléments n'ont d'autre prétention que de constituer une porte d'entrée aux travaux qui les suivent, en isolant parmi ceux-ci certains aspects statistiques et probabilistes suffisamment simples pour être détaillés ici.

1.2 Déviations auto-normalisées et estimation

1.2.1 Motivation

Plusieurs applications présentées dans ce mémoire font apparaître le besoin de construire séquentiellement, pour l'espérance commune μ d'une suite de variables aléatoires réelles $(X_t)_t$, une suite d'intervalles de confiance $[a_t, b_t]$ dont on puisse garantir *conjointement* la validité sur toute une plage de temps $t \in \{1, \dots, n\}$: on cherche donc à assurer une bonne probabilité à l'événement $\bigcap_{t \leqslant n} \{\mu \in [a_t, b_t]\}$. Si les bornes de cet intervalle de confiance sont (très classiquement [1]) choisies telles que $\bar{X}_t - a_t = b_t - \bar{X}_t = c/\sqrt{t}$ pour une certaine constante c, il apparaît donc nécessaire de contrôler

$$\sup_{t \leqslant n} \sqrt{t} \left| \bar{X}_t - \mu \right| \ .$$

Ce choix n'étant toutefois pas toujours optimal, j'ai suivi une approche légèrement différente, consistant à mesurer les déviations de $\bar{X}_t$ non pas en valeur absolue, mais dans une métrique informationnelle adaptée au problème, et à en déduire a posteriori les bornes des intervalles de confiance. L'idée est la suivante : supposons que l'on puisse écrire, pour toutes les valeurs des paramètres (et en particulier pour toute valeur de μ) une borne de Cramer, avec une fonction de taux $I(\cdot, \mu)$, du type [2]

$$\forall x_t \geqslant \mu, P(\bar{X}_t \geqslant x_t) \leqslant \exp(-tI(x_t; \mu)) \ .$$

La fonction $I(\cdot; \mu)$ étant croissante sur $[\mu, +\infty[$, cette borne peut être relue $P(I(\bar{X}_t; \mu) \geqslant I(x_t; \mu), \bar{X}_t \geqslant \mu) \leqslant \exp(-tI(x_t; \mu))$ soit, en posant $\delta = tI(x_t; \mu)$, $P(tI(\bar{X}_t; \mu) \geqslant \delta, \bar{X}_t \geqslant \mu) \leqslant \exp(-\delta)$; en procédant de même pour les déviations à gauche, on obtient

$$P\left(tI(\bar{X}_t; \mu) \geqslant \delta\right) \leqslant 2\exp(-\delta) \ .$$

1. Des travaux récents d'Olivier Catoni [Cat10] proposent et étudient des estimateurs alternatifs basés sur des bornes PAC-bayésiennes pouvant s'avérer meilleurs que l'espérance et la variance empiriques.

2. Pour avoir un exemple en tête, on pourra penser à l'estimation du paramètre μ d'une suite de variables aléatoires i.i.d. de Bernoulli, pour laquelle on peut prendre $I(x; \mu) = x \log(x/\mu) + (1-x)\log((1-x)/(1-\mu))$.

Cela conduit à choisir pour intervalle de confiance de risque α des voisinages de $\bar{X}_t$ au sens de la pseudo-métrique définie par I :

$$[a_t, b_t] = \left\{ \mu : tI(\bar{X}_t; \mu) \leqslant \log \frac{2}{\alpha} \right\} .$$

Si l'on souhaite utiliser ce type d'intervalles de confiance séquentiellement, en contrôlant $P\left(\bigcap_{t \leqslant n} \{\mu \in [a_t, b_t]\}\right)$, on est donc amené à étudier

$$\sup_{t \leqslant n} t\, I\big(\bar{X}_t; \mu\big) \ .$$

1.2.2 Inégalités de déviations auto-normalisées

Pour une filtration croissante $(\mathcal{F}_t)_{t \geqslant 0}$ donnée sur un espace probabilisé, considérons un processus réel à temps discret $(S_t)_{t \geqslant 0}$ adapté tel que $S_0 = 0$ et dont les incréments $X_t = S_t - S_{t-1}$ sont dominés de la façon suivante : il existe $\lambda_1 \in [-\infty, 0[$ et $\lambda_2 \in]0, +\infty]$ ainsi qu'une fonction $\phi :]\lambda_1, \lambda_2[\to \mathbb{R}$ tels que pour tout $\lambda \in]\lambda_1, \lambda_2[$ et pour $t \geqslant 1$,

$$\mathbb{E}\left[\exp(\lambda X_t) \,|\mathcal{F}_{t-1}\right] \leqslant \exp\left(\phi(\lambda)\right) \ .$$

En d'autres termes, la fonction ϕ domine les fonctions logarithmiques génératrices de moments (flgm) de tous les incréments $(X_t)_t$, qui doivent donc partager la même espérance conditionnelle finie μ. Si les incréments X_t sont identiquement distribués, on peut choisir pour ϕ la flgm commune, mais nous verrons qu'il est utile de pouvoir considérer des cas plus généraux. On supposera toutefois que ϕ satisfait toutes les propriétés d'une flgm (voir [DZ10], chapitre 2) : ϕ est une fonction convexe, nulle en μ, de classe C^∞ sur $]\lambda_1, \lambda_2[$; sa transformée de Fenchel-Legendre, que nous notons $I(\cdot; \mu)$ pour des raisons qui apparaîtront par la suite, est définie sur $\mathbb{R}$ par la relation

$$I(x; \mu) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \phi(\lambda)\} \ ;$$

c'est une fonction de taux convexe à valeur dans $\mathbb{R}^+ \cup \{+\infty\}$, finie et de classe C^∞ sur un intervalle ouvert $\mathcal{D}_I$ de $\mathbb{R}$ contenant 0, telle que $I(\mu, \mu) = 0$; pour tout x tel que $I(x) < \infty$ il existe un unique réel $\lambda(x) \in]\lambda_1, \lambda_2[$ pour lequel

$$\phi'(\lambda(x)) = x \quad \text{et } I(x; \mu) = \lambda(x)x - \phi(\lambda(x)) \ .$$

$I(x; \mu)$ tend vers l'infini quand x tend vers l'infini, et elle peut être égale à $+\infty$ à l'extérieur d'un intervalle (x_-, x_+) où elle est finie : on notera I_- et I_+ les limites respectives de $I(\cdot, \mu)$ quand x tend vers x_- et x_+, et on a $P(X_t \in [x_-, x_+]) = 1$. On montre alors le résultat suivant :

Théorème 1. *Soit $\delta > 0$. Pour tout réel $\eta > 0$,*

$$P\left(\exists t \in \{1,\dots,n\} : tI(\bar{X}_t;\mu) \geqslant \delta\right) \leqslant 2\left\lceil \frac{\log n}{\log(1+\eta)} \right\rceil \exp\left(-\frac{\delta}{1+\eta}\right) .$$

En particulier, pour $\eta = \delta/(\delta-1)$, on obtient :

$$P\left(\exists t \in \{1,\dots,n\} : tI(\bar{X}_t;\mu) \geqslant \delta\right) \leqslant 2e\left\lceil \delta\log(n)\right\rceil \exp(-\delta) .$$

Preuve du théorème 1

Donnons la courte preuve de ce résultat, inspirée de celle présentée par Neveu de la loi du Logarithme itéré pour les martingales [Nev72]. On découpe la plage d'indices $\{1,\dots,n\}$ en "tranches" $\{t_{k-1}+1,\dots,t_k\}$ de tailles exponentiellement croissantes : soit $t_0 = 0$ et, pour tout entier k strictement positif, $t_k = \lfloor (1+\eta)^k \rfloor$. Si l'on appelle $D = \lceil \log(n)/\log(1+\eta)\rceil$ le premier entier tel que $t_D \geqslant n$, on a donc :

$$P\left(\bigcup_{t=1}^{n}\left\{tI\left(\bar{X}_t;\mu\right) \geqslant \delta\right\}\right) \leqslant \sum_{k=1}^{D} P(A_k) ,$$

où $A_k = \bigcup_{t=t_{k-1}+1}^{t_k}\left\{tI\left(\bar{X}_t;\mu\right) \geqslant \delta\right\}$. Notons s le plus petit entier tel que $\delta/(s+1) \leqslant I_+$: pour $t \leqslant s$, il est clair que $P(tI(\bar{X}_t;\mu) \geqslant \delta, \bar{X}_t > \mu) = 0$ et donc $P(A_k) = 0$ si $t_k \leqslant s$.

Soit donc k tel que $t_k > s$: on note $\tilde{t}_{k-1} = \max\{t_{k-1}, s\}$. Pour tout $t \in \{\tilde{t}_{k-1}+1,\dots,t_k\}$, il existe un réel $x_t \in [\mu, x_+]$ tel que $tI(x_t;\mu) = \delta$. Posons $\lambda_k = \lambda(x_{t_k})$, de sorte que $I(x_{t_k};\mu) = \lambda_k x_{t_k} - \phi(\lambda_k)$, et considérons la sur-martingale positive $(W_t^k)_t$ définie par $W_0^k = 1$ et, pour tout $t \geqslant 1$,

$$W_t^k = \exp\left(\lambda_k S_t - t\phi(\lambda_k)\right) .$$

L'inégalité maximale assure que, pour tout réel $c > 0$,

$$P\left(\bigcup_{t=t_{k-1}+1}^{t_k}\left\{W_t^k \geqslant c\right\}\right) \leqslant P\left(\sup_{t\geqslant 0} W_t^k \geqslant c\right) \leqslant \frac{1}{c} .$$

Montrons comment en déduire une borne pour $P(A_k)$. Comme $tI(x_t;\mu) = \delta$, on a

$$I(x_{t_k};\mu) \leqslant I(x_t;\mu) < I(x_{t_k};\mu)\,(1+\eta) ,$$

et, comme $I(\cdot;\mu)$ est croissante à droite de μ, $x_t \geqslant x_{t_k}$ donc

$$\lambda_k x_t - \phi(\lambda_k) \geqslant \lambda_k x_{t_k} - \phi(\lambda_k) = I(x_{t_k};\mu) \geqslant \frac{I(x_t;\mu)}{1+\eta} .$$

Par conséquent,

$$\begin{aligned} tI\left(\bar{X}_t;\mu\right) \geqslant \delta \text{ et } \bar{X}_t \geqslant \mu &\implies \lambda_k \bar{X}_t - \phi(\lambda_k) \geqslant \lambda_k x_t - \phi(\lambda_k) \geqslant \frac{\delta}{t(1+\eta)} \\ &\implies \lambda_k S_t - t\phi(\lambda_k) \geqslant \frac{\delta}{1+\eta} \\ &\implies W_t^k \geqslant \exp\left(\frac{\delta}{1+\eta}\right) , \end{aligned}$$

ce qui entraîne que

$$\begin{aligned} P\left(\bigcup_{t=t_{k-1}+1}^{t_k} \left\{tI\left(\bar{X}_t;\mu\right) \geqslant \delta\right\} \cap \left\{\bar{X}_t > \mu\right\}\right) \\ \leqslant P\left(\bigcup_{t=t_{k-1}+1}^{t_k} \left\{W_t^k \geqslant \exp\left(\frac{\delta}{1+\eta}\right)\right\}\right) \leqslant \exp\left(-\frac{\delta}{1+\eta}\right) . \end{aligned}$$

On prouve de la même façon que

$$P\left(\bigcup_{t=t_{k-1}+1}^{t_k} \left\{tI\left(\bar{X}_t;\mu\right) \geqslant \delta\right\} \cap \left\{\bar{X}_t < \mu\right\}\right) \leqslant \exp\left(-\frac{\delta}{1+\eta}\right) ,$$

ce qui finit la preuve. Le choix particulier $\eta = \delta/(\delta-1)$ est à peu près optimal, et permet, en remarquant que $\log(1+1/(\delta-1)) \geqslant 1/\delta$, d'obtenir une borne simple mettant bien en exergue le coût de l'uniformité temporelle dans le facteur $e\lceil \delta \log(n) \rceil$. Il apparaît (tout particulièrement dans [GC11, GM11, GL11]) que la modicité de ce coût, par rapport à un facteur n que donnerait une borne de l'union, est un élément décisif pour l'analyse de certains algorithmes.

Améliorations et variantes

Ce résultat peut être amélioré significativement si la fonction $I(\cdot;\mu)$ satisfait certaines hypothèses :

Théorème 2. *Soit $\delta > 0$. Si la fonction $I(\cdot;\mu)$ est log-concave, alors pour tout réel $\eta > 0$,*

$$P\left(\exists t \in \{1,\dots,n\} : tI(\bar{X}_t;\mu) \geqslant \delta\right) \leqslant 2\left\lceil \frac{\log n}{\log(1+\eta)} \right\rceil \exp\left(-\left(1-\frac{\eta^2}{8}\right)\delta\right) .$$

En particulier, pour $\eta = 2/\sqrt{\delta}$, on obtient :

$$P\left(\exists t \in \{1,\dots,n\} : tI(\bar{X}_t;\mu) \geqslant \delta\right) \leqslant 2\sqrt{e}\left\lceil \frac{\sqrt{\delta}}{2}\log(n) \right\rceil \exp(-\delta) .$$

La loi du logarithme itéré suggère que cette amélioration n'est pas très loin d'être optimale : dans le cadre sous-gaussien, où l'on possède la majoration $I(x;\mu) \geqslant (x-\mu)^2/(2\sigma^2)$, on obtient en effet que pour toute constante réelle $c > 1$:

$$P\left(\sup_{t\leqslant n} \frac{S_t - t\mu}{\sqrt{2\sigma^2 t \log\log(n)}} > c\right) \leqslant P\left(\sup_{t\leqslant n} tI\left(\bar{X}_t;\mu\right) > c^2 \log\log(n)\right) \to 0$$

quand n tend vers l'infini. Notons d'ailleurs que la log-concavité de $I(\cdot,\mu)$, bien qu'elle ne soit pas toujours vérifiée (en particulier pour les variables bornées), est au moins localement autour de μ une hypothèse assez raisonnable si l'on pense au régime gaussien.

Pour prouver le théorème 2, il suffit d'améliorer le contrôle de ce que l'on perd en utilisant une unique valeur λ_k commune pour tous les indices $t \in \{\tilde{t}_{k-1}+1,\ldots,t_k\}$ au lieu des $\lambda(x_t)$. En choisisant, à la place de λ_k, la valeur plus "centrale" $\tilde{\lambda}_k = \lambda(z_k)$, où $z_k \in]x_{t_k}, x_{t_{k-1}+1}[$ est pris tel que $I(z_k;\mu) = \delta/(1+\eta)^{k-1/2}$, on obtient que pour tout $t \in \{\tilde{t}_{k-1}+1,\ldots,t_k\}$:

$$r_t = \frac{I(x_t;\mu)}{I(z_k;\mu)} \in \left[\frac{1}{\sqrt{1+\eta}}; \sqrt{1+\eta}\right[\ .$$

Par ailleurs, en désignant par $I'(\cdot,\mu)$ la dérivée de $I(\cdot,\mu)$, et en utilisant le fait que la transformation de Fenchel-Legendre est ici involutive, on peut écrire

$$\begin{aligned}
I(x_t;\mu) - \left(\tilde{\lambda}_k x_t - \phi(\tilde{\lambda}_k)\right) &= I(x_t;\mu) - \left(\tilde{\lambda}_k x_t - (\tilde{\lambda}_k z_k - I(z_k;\mu))\right) \\
&= \tilde{\lambda}_k(z_k - x_t) + I(x_t;\mu) - I(z_k;\mu) \\
&= I'(z_k;\mu)(z_k - x_t) + I(x_t;\mu) - I(z_k;\mu) \\
&= I(z_k;\mu)\left(r_t - 1 - \frac{I'(z_k;\mu)}{I(z_k;\mu)}(x_t - z_k)\right) \\
&\leqslant \frac{I(x_t;\mu)}{r_t}\left(r_t - 1 - \log(r_t)\right) \qquad (\star) \\
&= g(r_t)\, I(x_t;\mu)
\end{aligned}$$

en posant $g(r) = (r - 1 - \log(r))/r$. C'est dans l'inégalité $(\star)$ que l'on utilise l'hypothèse de concavité de $\log I(\cdot;\mu)$:

$$\log(I(x_t;\mu)) - \log(I(z_k;\mu)) \leqslant I'(z_k;\mu)/I(z_k;\mu)(x_t - z_k) \ .$$

On termine en remarquant que g, qui est la primitive de la fonction $r \mapsto \log(r)/r^2$ s'annulant en $r = 1$, est décroissante sur $]0,1[$, puis croissante sur $]1,+\infty[$, et que pour $r \geqslant 1$ on a $g(r) \leqslant g(1/r) \leqslant (r-1)^2/2$. Ainsi, pour tout $t \in \{\tilde{t}_{k-1}+1,\ldots,t_k\}$,

$$g(r_t) \leqslant \max\left\{g\left(\sqrt{1+\eta}\right), g\left(\frac{1}{\sqrt{1+\eta}}\right)\right\} \leqslant \frac{\left(\sqrt{1+\eta}-1\right)^2}{2} \leqslant \frac{\eta^2}{8} \ ,$$

et donc

$$\tilde{\lambda}_k x_t - \phi(\lambda_k) \geqslant \left(1 - \frac{\eta^2}{8}\right) I(x_t, \mu) \; .$$

Signalons pour l'anecdote que le cas quadratique où $I(x;\mu) = 2(x-\mu)^2/K^2$, on obtient la très légèrement meilleure borne suivante :

$$P\left(\exists t \in \{1,\dots,n\} : tI(\bar{X}_t;\mu) \geqslant \delta\right) \leqslant 2\left\lceil \frac{\log n}{\log(1+\eta)} \right\rceil \exp\left(-\left(1 - \frac{\eta^2}{16}\right)\delta\right) \; .$$

Si l'on souhaite une borne non asymptotique plus conforme à l'esprit de la loi du logarithme itéré, valable pour tout $t \geqslant 1$, on peut montrer par exemple de manière très semblable le résultat suivant :

Théorème 3. *Pour tout $\delta > 1$ et $c > 1$,*

$$P\left(\exists t \geqslant 1 : tI(\bar{X}_t;\mu) \geqslant \frac{\delta c}{\delta - 1} \log\log t + \delta\right) \leqslant \frac{2\,\mathrm{e}\, c\delta^c}{c-1} \exp(-\delta) \; .$$

En particulier, pour $c = 1 + 1/\log(\delta)$, on obtient

$$P\left(\exists t \geqslant 1 : tI(\bar{X}_t;\mu) \geqslant \frac{\delta(1+\log\delta)}{(\delta-1)\log\delta} \log\log t + \delta\right) \leqslant 2e^2\delta \exp(-\delta) \; .$$

Forme auto-normalisée

Dans les applications présentées dans ce mémoire, il arrive que le besoin d'assurer la validité conjointe de tous les intervalles de confiance provienne du fait que l'on n'observe les variables $(X_t)_t$ qu'épisodiquement, de façon prévisible au sens où il existe, pour tout $t \in \{1,\dots,n\}$, une variable aléatoire $\varepsilon_t \in \{0,1\}$ qui soit $\mathcal{F}_{t-1}$-mesurables et telle que l'estimée courante au temps n de la moyenne soit

$$\bar{X}(n) = S(n)/N(n) \; , \quad \text{où } S(n) = \sum_{t=1}^{n} \varepsilon_t X_t \quad \text{et } N(n) = \sum_{t=1}^{n} \varepsilon_t \; . \tag{1.1}$$

On obtient :

$$P\left(I\left(\bar{X}(n);\mu\right) \geqslant \frac{\delta}{N(n)}\right) \leqslant 2e\left\lceil \delta\log(n)\right\rceil \exp(-\delta) \; .$$

C'est d'ailleurs plutôt sous cette forme que le théorème 1 apparaît dans les articles présentés par ce mémoire : cela explique la dénomination de déviations *auto-normalisées*. Des quantités un peu analogues ont fait l'objet, ces dernières années, de travaux très intéressants (voir [DLPKL04, BT08], et aussi [BGR02]), dont

je n'ai pris connaissance qu'assez tard. La définition (1.1) fait apparaître $S(n)$ comme une transformée de martingale ou, si l'on veut, une intégrale stochastique discrète. Des variantes à temps continu du Théorème 1, peuvent être obtenues par une technique semblable pour des intégrales stochastiques.

Cette approche peut en outre être adaptée à un contexte non stationnaire : supposons pour faire simple [3] que les variables $(X_t)_t$ sont indépendantes et bornées par B, d'espérances μ_t. Si μ_t ne varie pas trop vite (ou pas trop souvent) en fonction de t, on peut s'intéresser pour $\gamma \in]0,1[$ à un estimateur escompté $\bar{X}_\gamma(n)$ de μ_n défini par

$$\bar{X}_\gamma(n) = S_\gamma(n)/N_\gamma(n) \ , \quad \text{où } S_\gamma(n) = \sum_{t=1}^{n} \gamma^{n-t}\varepsilon_t X_t \quad \text{et } N_\gamma(n) = \sum_{t=1}^{n} \gamma^{n-t}\varepsilon_t \ .$$

L'écart entre $\bar{X}_\gamma(n)$ et μ_n se décompose en un terme de biais (dont nous ne parlerons pas ici) et un terme de fluctuation $\bar{X}_\gamma(n) - M_\gamma(n)/N_\gamma(n)$, où $M_\gamma(n) = \sum_{t=1}^{n} \gamma^{n-t}\varepsilon_t\mu_t$. Ce terme de fluctuation peut être contrôlé par les mêmes techniques de martingales que précédemment : on obtient par exemple que

$$P\left(\frac{S_\gamma(n) - M_\gamma(n)}{\sqrt{N_{\gamma^2}(n)}} \geqslant \delta\right) \leqslant \left\lceil \frac{\log \nu_\gamma(n)}{\log(1+\eta)} \right\rceil \exp\left(-\frac{2\delta^2}{B^2}\left(1 - \frac{\eta^2}{16}\right)\right) \ ,$$

avec $\nu_\gamma(n) = \sum_{t=1}^{n} \gamma^{n-t} = (1-\gamma^n)/(1-\gamma) < \min\{(1-\gamma)^{-1}, n\}$.

1.2.3 Application à l'estimation

Montrons maintenant brièvement comment sont utilisées ces inégalités dans les travaux que je présente dans ce mémoire. L'idée fondamentale est que l'on peut, grâce au théorème 1, construire une suite $([a_t, b_t])_{1\leqslant t\leqslant n}$ d'intervalles de confiance pour μ simultanément valides avec grande probabilité en conservant toutes les valeurs qui sont dans un voisinage de $\bar{X}_t$ pour la pseudo-distance définie par I : si

$$[a_t, b_t] = \left\{\mu : tI\left(\bar{X}_t; \mu\right) \leqslant \delta\right\} \ ,$$

alors

$$P\left(\bigcap_{t=1}^{n} \{\mu \in [a_t, b_t]\}\right) \geqslant 1 - 2e\left\lceil \delta \log(n)\right\rceil \exp(-\delta) \ .$$

De même, on obtient des intervalles de confiance pour le cas présenté dans l'équation (1.1). Ce cadre concerne autant les problèmes de bandits, où l'on observe seulement la récompense du bras choisi, que l'estimation dans les modèles markoviens où seules sont mises à jour, à chaque instant, les estimées des

3. Ce cadre est emprunté à l'article [GM11].

lois conditionnelles au passé observé à cet instant. Bien entendu, dans ces deux exemples, le fait d'observer ou non n'est pas du tout indépendant des valeurs précédemment observées. En choisissant δ tel que $2e\lceil\delta\log(n)\rceil\exp(-\delta) \leqslant \alpha$, on obtient avec $\left\{\mu : I\left(\bar{X}(n);\mu\right) \leqslant \delta/N(n)\right\}$ un intervalle de confiance de risque au plus α.

Modèle exponentiel

Plaçons-nous par exemple dans le cas où les variables $(X_t)_t$ sont indépendantes et identiquement distribuées, avec une loi P_{θ_0} appartenant à un modèle exponentiel canonique $\{P_\theta : \theta \in \Theta\}$, où Θ est un intervalle réel et où P_θ admet par rapport à une mesure de référence la densité $p_\theta : \mathbb{R} \to \mathbb{R}$ définie par :

$$p_\theta(x) = \exp\left(x\theta - b(\theta) + c(x)\right) \ .$$

Ici, c est une fonction réelle et la log-fonction de partition b est supposée deux fois différentiable. On sait alors qu'en notant $\mu(\theta) = \dot{b}(\theta)$ l'espérance de P_θ on définit une fonction μ différentiable bijective. Dans ce cas, un petit calcul (probablement bien connu) assure que la fonction de taux I est en correspondance directe avec la divergence de Kullback-Leibler (qui est en l'occurrence une divergence de Bregman pour b) : pour $\beta, \theta \in \Theta$,

$$\mathrm{KL}(P_\beta; P_\theta) = I(\mu(\beta);\mu(\theta)) = b(\theta) - b(\beta) - \dot{b}(\beta)(\theta - \beta) \ . \tag{1.2}$$

Ainsi, on construit une séquence $(R_t)_{t\geqslant 1}$ d'intervalles de confiance pour le paramètre θ_0 qui sont tous valides avec probabilité $1 - 2e\lceil\delta\log(n)\rceil\exp(-\delta)$ en choisissant

$$R_t = \left\{\theta : \mathrm{KL}\left(P_{\mu^{-1}(\bar{X}_t)}; P_\theta\right) \leqslant \frac{\delta}{t}\right\} = \left\{\theta : I\left(\bar{X}_t;\mu(\theta)\right) \leqslant \frac{\delta}{t}\right\} \ .$$

Cela s'applique en particulier aux variables de Poisson, exponentielles, de loi Gamma avec paramètre de forme fixé, etc. L'article [GC11] contient un exemple d'application au cas des variables exponentielles, pour lequel $I(x,y) = x/y - 1 - \log(x/y)$. Mais le cas des variables de Bernoulli mérite d'être souligné, car il permet en fait de traiter plus généralement les variables bornées de façon bien plus satisfaisante que les inégalités de type Hoeffding ou même, dans certains cas (voir [GC11]), que les inégalités de Bennett et Bernstein. En effet, comme l'avait remarqué Hoeffding [Hoe63], pour les moments expontiels aussi [4] les variables de Bernoulli sont les "moins concentrées" à espérance fixée : si X est une variable aléatoire à valeur dans l'ensemble $[0,1]$ d'espérance μ, alors pour tout $\lambda \in \mathbb{R}$ on a

$$E\left[\exp(\lambda X)\right] \leqslant 1 - \mu + \mu\exp(\lambda) \ ,$$

4. Cela est bien connu en ce qui concerne la variance.

avec égalité si et seulement si $X \sim \mathcal{B}(\mu)$. Si l'on note kl la fonction d'entropie binaire relative

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

qui – c'est une conséquence de l'équation (1.2) – est aussi la fonction de taux pour les variables de Bernoulli, le théorème 1 permet de conclure que pour des variables X_t bornées dans $[0,1]$,

$$P\left(\sup_{t\leqslant n}\mathrm{kl}\left(\bar{X}_t,\mu\right)\geqslant\frac{\delta}{t}\right)\leqslant 2e\left\lceil\delta\log(n)\right\rceil\exp(-\delta)\ .$$

Ce résultat permet bien entendu, après translation/dilatation et après application du lemme de Pinsker $\mathrm{kl}(p,q)\geqslant 2(p-q)^2$, de retrouver une contrepartie uniforme sur la plage $t\in\{1,\ldots,n\}$ à l'inégalité de Hoeffding :

$$P\left(\sup_{t\leqslant n}\left|\bar{X}_t-\mu\right|\geqslant\frac{\delta}{\sqrt{t}}\right)\leqslant 4e\left\lceil\delta^2\log(n)\right\rceil\exp(-2\delta^2)\ .$$

L'exemple des problèmes de bandits, détaillé dans [GC11], illustre toutefois l'intérêt de ne pas procéder à de telles majorations : même celles de Bernstein, en faisant apparaître un terme de reste le plus souvent négligeable ailleurs mais qui ne l'est pas ici, conduit à des algorithmes nettement moins efficaces.

Lois multinomiales

Comme le suggère le principe des grandes déviations, ce type d'inégalités ne se limite pas strictement au cas des variables réelles présentées ci-dessus. On peut également construire, par exemple, des régions de confiance informationnelles pour les lois multinomiales (dont on a besoin, par exemple, pour estimer des matrices de transition de chaînes de Markov discrètes, cf. [GL11, FCG10]). Soient en effet P et Q deux éléments de l'ensemble $\mathcal{S}$ des lois de probabilités sur l'ensemble fini A. On montre aisément que

$$\mathrm{KL}(P;Q)\leqslant\sum_{x\in A}\mathrm{kl}\left(P(x);Q(x)\right)\ ,$$

en remarquant que

$$\begin{aligned}-\mathrm{KL}(P;Q)+\sum_{x\in A}\mathrm{kl}\left(P(x);Q(x)\right)&=\sum_{x\in A}\left(1-P(x)\right)\log\frac{1-P(x)}{1-Q(x)}\\&=(|A|-1)\sum_{x\in A}\frac{1-P(x)}{|A|-1}\log\left(\frac{(1-P(x))/(|A|-1)}{(1-Q(x))/(|A|-1)}\right)\geqslant 0\ .\end{aligned}$$

On en déduit que si $X_1, \dots, X_n$ sont des variables indépendantes identiquement distribuées sous la loi $P_0 \in \mathcal{S}$, et si $\hat{P}_t(k) = \sum_{s=1}^{t} \mathbb{1}\{X_s = k\}/t$, on a

$$\begin{aligned} P\Big(\exists t \in \{1,\dots,n\} : \ \mathrm{KL}\left(\hat{P}_t; P_0\right) \geqslant \frac{\delta}{t}\Big) & \\ \leqslant P\left(\exists t \in \{1,\dots,n\} : \sum_{a\in A} \mathrm{kl}\left(\hat{P}_t(a); P_0(a)\right) \geqslant \frac{\delta}{t}\right) & \\ \leqslant \sum_{a\in A} P\left(\exists t \in \{1,\dots,n\} : \mathrm{kl}\left(\hat{P}_t(a); P_0(a)\right) \geqslant \frac{\delta}{|A|t}\right) & \\ \leqslant 2e\left(\delta \log(n) + |A|\right) \exp\left(-\frac{\delta}{|A|}\right) \ . & \qquad (1.3) \end{aligned}$$

Le fait que cette borne porte directement sur la divergence de Kullback-Leibler entre la mesure empirique et la vraie loi permet, pour l'estimation des modèles de mémoire markoviens (voir [GL11]), de s'affranchir d'hypothèses inutiles qui résultaient de l'application des majorations de Bernstein.

Par ailleurs, la majoration (1.3) permet de construire une suite $(R_t)_{t\leqslant n}$ de régions de confiance "de type Sanov" pour P_0 simultanément valides avec probabilité $1-\alpha$ en choisissant des voisinages de Kullback-Leibler du maximum de vraisemblance :

$$R_t = \left\{ Q \in \mathcal{S} : \mathrm{KL}(\hat{P}_t; Q) \leqslant \frac{\delta}{t} \right\} \ ,$$

avec δ tel que $2e\left(\delta \log(n) + |A|\right) \exp\left(-\delta/|A|\right) = \alpha$. Comme cela est illustré à la figure 1.1, ces régions R_t du simplexe, tenant compte en particulier des différences de variabilité entre composantes, ont des propriétés géométriques agréables qui sont exploitées dans l'article [FCG10].

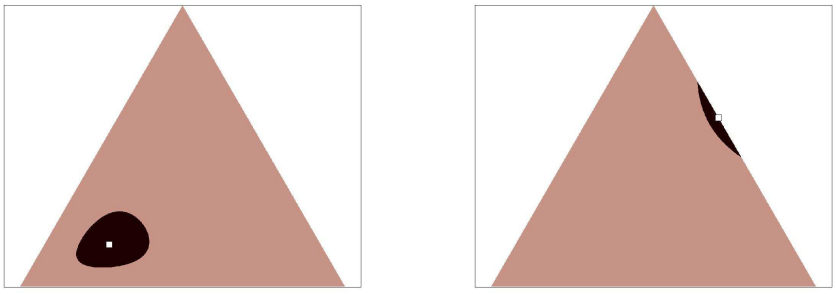


FIGURE 1.1 – Deux exemples de régions de confiance de Kullback-Leibler (en noir) autour d'un estimateur (carré blanc) dans le simplexe $\mathcal{S}$ de dimension 2.

Chapitre 2

Apprentissage par renforcement

J'ai commencé à travailler sur des questions d'apprentissage par renforcement avec Eric Moulines, Olivier Cappé et Sarah Filippi, qui rédigeait une thèse pour la société Orange sur différents problèmes posés par l'évolution technologique des réseaux de télécommunication, et notamment sur la gestion des réseaux sans fil nomades ainsi que sur la publicité en ligne. Comme souvent à la découverte d'un nouveau sujet, nous avons commencé par réfléchir aux questions les plus difficiles, avant de comprendre quels étaient les vrais enjeux sur lesquels il convenait de réfléchir en premier lieu. Ce chapitre remet les choses en ordre : je commence par présenter les modèles les plus élémentaires, auxquels une réponse très fine peut être donnée, puis je reviens sur la possibilité d'étendre notre approche à des modèles plus ambitieux et réalistes.

2.1 Le problème des bandits stochastiques

Le problème des bandits multi-bras (voir [CBL06]) est un modèle paradigmatique de l'apprentissage par renforcement où un agent, qui fait face à une machine à sous possèdant plusieurs bras, essaye de maximiser son profit par un choix judicieux de tirage des bras. Dans la version stochastique [1] la plus simple du problème, l'agent choisit à chaque étape $t = 1, 2, \ldots$ un bras $A_t \in \{1, \ldots, K\}$, et il reçoit une récompense X_t telle que, conditionnellement au choix des bras $A_1, A_2, \ldots$, les récompenses soient indépendantes et identiquement distribuése, d'espérances $\mu_{A_1}, \mu_{A_2}, \ldots$. On appelle sa *politique* la règle de décision (potentiellement randomisée) qui, aux observations passées $(A_1, X_1, \ldots, A_{t-1}, X_{t-1})$, associe

1. Il existe une variante dite *adversariale*, relevant de la théorie de jeux, où les récompenses sont choisies non pas au hasard mais par un adversaire ; il n'en sera pas question ici : outre [CBL06], le lecteur intéressé pourra se réfèrer avec profit aux travaux de Gilles Stoltz pour en saisir les subtilités.

son prochain choix A_t. Le meilleur choix est le bras a^* qui correspond à la récompense moyenne maximale μ_{a^*}. La performance de sa politique est mesurée par le *regret* R_n, qui est défini comme la différence entre les récompenses qu'elle accumule jusqu'au temps $t = n$ et ce qu'elle aurait pu accumuler pendant la même période si elle avait su depuis le début quel bras offre la meilleure récompense moyenne.

Notre agent fait face à ce que l'on appelle en apprentissage par renforcement un "dilemme exploration-exploitation" : au temps t, il doit tirer profit de l'information qu'il a accumulée au cours des tours précédents, en choisissant le bras qui s'est jusqu'alors avéré le plus intéressant, mais il ne doit pas négliger non plus la possibilité que les autres bras soient en fait sous-évalués et il doit donc les jouer suffisamment souvent. Depuis les travaux de [Git79] dans les années 1970, ce problème a soulevé beaucoup d'intérêt et de nombreuses variantes et extensions ont été proposées (voir [EDMM02] et les références citées).

Deux familles de modèles de bandits peuvent être distinguées : dans la première, on suppose que la loi de la récompense X_t conditionnellement à l'événement $A_t = a$ appartient à une famille $\{p_\theta, \theta \in \Theta_a\}$ de lois de probabilités. Dans un contexte paramétrique particulier, [LR85] a montré pour ce cas une borne inférieure asymptotique à la performance de n'importe quelle politique. Ces résultats ont été généralisés dans l'article [BK97], où il est montré que le nombre $N_a(n)$ de tirages jusqu'au temps n d'un bras sous-optimal a est borné inférieurement de la façon suivante :

$$N_a(n) \geqslant \left(\frac{1}{\inf_{\theta \in \Theta_a : E[p_\theta] > \mu_{a^*}} \mathrm{KL}\left(p_{\theta_a}, p_\theta\right)} + o(1) \right) \log(n), \tag{2.1}$$

KL désignant la divergence de Kullback-Leibler et $E[p_\theta]$ désignant l'espérance sous p_θ ; par conséquent, le regret est borné inférieurement :

$$\liminf_{n \to \infty} \frac{\mathbb{E}[R_n]}{\log(n)} \geqslant \sum_{a : \mu_a < \mu_{a^*}} \frac{\mu_{a^*} - \mu_a}{\inf_{\theta \in \Theta_a : E[p_\theta] > \mu_{a^*}} \mathrm{KL}\left(p_{\theta_a}, p_\theta\right)} \,. \tag{2.2}$$

Dans un cadre bayésien particulier, une politique optimale peut être trouvée pour ce problème : Gittins [Git79] a montré qu'elle pouvait même être choisie sous la forme très simple d'une *politique d'indice*, consistant à calculer pour chaque bras un indice d'allocation dynamique (qui dépend des tirages passés de ce bras, et qui est légèrement plus grand que la moyenne a posteriori), puis à choisir le bras qui possède l'indice le plus grand. Récemment, Honda et Takemura [HT10] ont proposé un algorithme appelé *Deterministic Minimum Empirical Divergence (DMED)* dont ils ont montré l'optimalité au premier ordre. Cet algorithme, qui tient à jour une liste des bras suffisamment proches du meilleur (et qui donc doivent être joués), est inspiré par des considérations de grandes déviations et

s'appuie sur la connaissance de la fonction de taux associée à la distribution des récompenses.

La seconde famille de bandits ne possède pas de modèle paramétrique : les récompenses sont seulement supposées bornées (sans perte de généralité, entre 0 et 1). Dans ce cadre, l'attention s'est vite portée sur des politiques d'indices qui, plutôt que d'utiliser directement un estimateur de la récompense moyenne associée à un bras, s'appuient plutôt sur une *borne supérieure de confiance* (upper-confidence bound, UCB). Cette approche est parfois appelée "optimiste" : l'agent fait à chaque instant comme si, parmi toutes les distributions de récompenses possibles statistiquement compatibles avec ses observations passées, il avait face à lui la plus favorable. La référence sur ce thème est l'article [ACBF02] (à la suite de [Agr95]), où sont proposées et analysées deux variantes : UCB1 (souvent appelé simplement UCB par la suite) et UCB2. UCB est une procédure séquentielle qui ne nécessite pas de connaître l'horizon n, et pour laquelle est prouvée l'existence d'une constante C telle que

$$\mathbb{E}[R_n] \leqslant \sum_{a:\mu_a<\mu_{a^*}} \frac{8\log(n)}{(\mu_{a^*}-\mu_a)} + C \; . \tag{2.3}$$

La variante UCB2 requiert un paramètre α qui doit être ajusté en fonction notamment de l'horizon, et qui satisfait la borne améliorée

$$\mathbb{E}[R_n] \leqslant \sum_{a:\mu_a<\mu_{a^*}} \frac{(1+\epsilon(\alpha))\log(n)}{2\,(\mu_{a^*}-\mu_a)} + C(\alpha) \; ,$$

où $\epsilon(\alpha) > 0$ est une constante qui devient arbitrairement petite quand α tend vers 0, au prix d'une augmentation de $C(\alpha)$. La constante $1/2$ devant le facteur $\log(n)/\left(\mu_{a^*}-\mu_a\right)$ ne peut pas être améliorée. Dans l'article [GC11], nous obtenons comme corollaire de notre analyse (proposition 4) qu'une version correctement paramétrée d'UCB satisfait en fait la même borne, de sorte qu'UCB2 n'offre pas de garantie supérieure. Il apparaît en revanche dans les simulations présentées dans [ACBF02] qu'une autre variante, appelée UCB-Tuned, peut être significativement meilleure en intégrant une estimation de la variance. Cette variante n'a jamais été analysée, mais dans l'article ultérieur [AMS09] a été proposée une politique de même inspiration, appelée UCB-V, qui utilise une version empirique de la borne de Bernstein pour obtenir des bornes plus fines dans certains cas.

L'algorithme KL-UCB (pour Kullback-Leibler UCB), que j'ai étudié avec Olivier Cappé dans l'article [GC11], réconcilie les deux familles de bandits : il donne une solution séquentielle générique particulièrement efficace dans la seconde, tout en formant une solution optimale dans la première. Il dépend d'un seul paramètre

fonctionnel, lié à la famille des distributions possibles pour chaque bras, qui définit une pseudo-métrique sur l'enveloppe convexe de l'ensemble des récompenses possibles. Il est décrit dans l'algorithme 1.

Algorithme 1 KL-UCB

Entrées: n (horizon), K (nombre de bras), REWARD (fonction de récompense), d (pseudo-métrique)

Pour $t = 1$ **to** K **Faire**
  $N[t] \leftarrow 1$
  $S[t] \leftarrow \text{REWARD}(\text{arm} = t)$
Fin Pour
Pour $t = K+1$ **to** n **Faire**
  $a \leftarrow \arg\max_{1 \leqslant a \leqslant K} \max \left\{ q \in \Theta : N[a]\, d\left(\frac{S[a]}{N[a]}, q\right) \leqslant \log(t) + c\log(\log(t)) \right\}$
  $r \leftarrow \text{REWARD}(\text{arm} = a)$
  $N[a] \leftarrow N[a] + 1$
  $S[a] \leftarrow S[a] + r$
Fin Pour

Dans le cadre paramétrique d'une famille de récompenses qui appartient au modèle exponentiel canonique, nous montrons des bornes de regret à horizon fini pour la politique KL-UCB qui atteignent la borne asymptotique de [LR85]. Dans le cadre non paramétrique des bandits bornés, nous montrons qu'un choix particulier de pseudo-métrique permet à KL-UCB d'être la première politique d'indice à la fois universellement plus efficace qu'UCB et optimale dans le cas binaire (théorèmes 1 et 2, voir aussi la section 4) : pour le choix du paramètre $c = 3$ et en prenant pour pseudo-distance d la divergence entre deux lois du modèle d'espérances respectives μ et ν , il apparaît que pour tout entier n le nombre de tirages du bras sous-optimal a satisfait l'inégalité :

$$\mathbb{E}[N_n(a)] \leqslant \frac{\log(n)}{d(\mu_a, \mu_{a^*})}(1+\epsilon) + C_1 \log(\log(n)) + \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}} ,$$

où C_1 est une constante positive et où $C_2(\epsilon)$ et $\beta(\epsilon)$ sont des fonctions de ϵ. Il s'ensuit que

$$\limsup_{n\to\infty} \frac{\mathbb{E}[N_n(a)]}{\log(n)} \leqslant \frac{1}{d(\mu_a, \mu_{a^*})} .$$

et donc que

$$\limsup_{n\to\infty} \frac{\mathbb{E}[R_n]}{\log(n)} \leqslant \sum_{a:\mu_a<\mu_{a^*}} \frac{\mu_{a^*} - \mu_a}{d(\mu_a, \mu_{a^*})} .$$

Pour les bandits bornés, nous montrons que l'on peut utiliser pour d la divergence de Kullback-Leibler $d(p,q) = p\log(p/q) + (1-p)\log((1-p/(1-q))$ entre

deux variables de Bernoulli de paramètres p et q : les résultats ci-dessus restent alors valides. Malgré la présence de d, cette borne n'est absolument pas spécifique au cas binaire et s'applique à toutes les distributions bornées entre 0 et 1 (et donc, en remettant le cas échéant à l'échelle, à toutes les distributions bornées). Par l'inégalité de Pinsker $d(\mu_a, \mu_{a^*}) > 2(\mu_a - \mu_{a^*})^2$, on voit que KL-UCB possède des garanties théoriques strictement meilleures que celles d'UCB, alors qu'il a le même cadre d'application. Ce choix de pseudo-distance permet donc d'avoir à la fois une procédure générale plus efficace que tous ses concurrents pour les bandits bornés, et une solution asymptotiquement optimale pour le cas binaire. La principale clé, dans la preuve de ces résultats, réside dans l'utilisation (au sein d'un schéma de preuve original) des inégalités de déviations auto-normalisées mesurées dans la bonne métrique (théorèmes 10 et 11), que nous avons présentées en introduction.

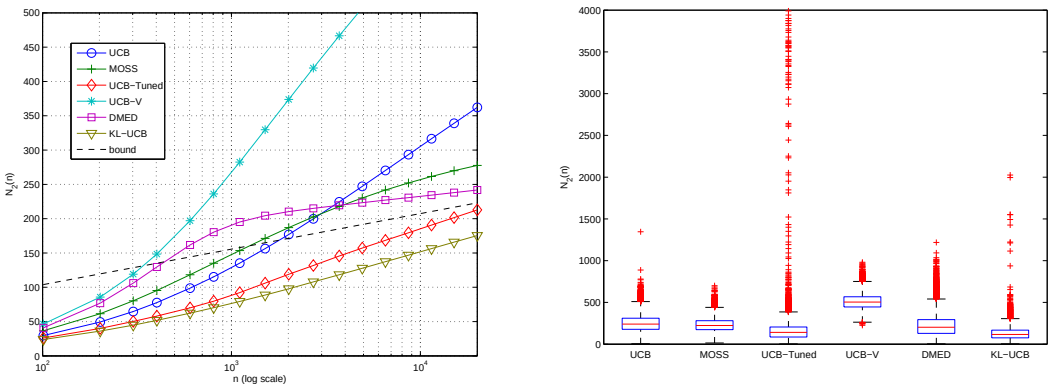


FIGURE 2.1 – Performance des différents algorithmes dans un scénario simple à deux bras. À gauche : nombre moyen de tirages du bras sous-optimal en fonction du temps ; à droite : distribution empirique du nombre de tirages sous-optimaux au temps $t = 5000$ pour $50,000$ expériences indépendantes.

En pratique, la comparaison avec ses principaux concurrents UCB, MOSS, UCB-V, DMED et même UCB-tuned tourne à l'avantage de KL-UCB. Nous avons mené des expériences dans différents scénarios de bandits bornés (voir un exemple figure 2.1), ainsi que pour des lois de récompenses exponentielles et poissoniennes. L'estimation du regret moyen nécessite des simulations numériques assez coûteuses en temps, puisqu'elles sont habituellement menées sur des horizons assez longs (plusieurs dizaines ou centaines de milliers d'étapes) et, par ailleurs, du fait de l'extrême dispersion du regret d'une simulation à l'autre : s'il est en effet le plus souvent d'ordre $\log(n)$, il arrive avec une probabilité faible mais non négligeable pour tous les algorithmes étudiés ici qu'il soit plutôt d'ordre n, ce qui peut peser sur l'espérance (voir par exemple [AMS09, SA11]). Concernant UCB-V, la raison de sa relativement faible efficacité est que le terme de second ordre

dans l'inégalité de Bernstein, qui est d'habitude négligeable, ne disparaît pas ici pour les bras sous-optimaux puisque ceux-ci sont tirés peu souvent, alors que le niveau de confiance demandé est de plus en plus grand. Quant à UCB-Tuned, il s'avère insuffisamment prudent.

2.2 Environnement non stationnaire

Bien que la formulation stationnaire du problème des bandits présente une jolie modélisation du dilemme exploration-exploitation, elle peut s'avérer insuffisante pour rendre compte des situations où les distributions de récompenses varient au cours du temps. Par exemple, dans le problème de la radio cognitive [LGJP11], un utilisateur opportuniste cherche à exploiter la disponibilité d'un canal dans un système multi-canaux ; la récompense est alors pour lui la disponibilité du canal, dont la loi de probabilité lui est inconnue. Une autre application est l'optimisation, en temps réel, de l'offre de contenu disponible sur un site web (une telle situation a fait l'objet d'un challenge organisé par le PASCAL en 2006, voir [HGB$^+$06], [KX08]). Ces exemples illustrent les limites du modèle des bandits stationnaires : la probabilité qu'un canal soit libre change au fil du temps, et les contenus intéressant un visiteur évoluent.

Pour modéliser de telles situations, des problèmes de bandits non stationnaires ont été considérés (voir [KS06, HGB$^+$06, SU08, YM09]), pour lesquels les distributions de récompenses peuvent changer. Motivés par les problèmes cités ci-dessus, et suivant en cela un paradigme largement répandu dans le domaine de la détection de rupture (voir [Fuh04, Mei06] et les références citées), Eric Moulines et moi avons étudié des environnements non stationnaires où les lois des récompenses connaissent des variations brusques. Comme on peut s'y attendre, nous montrons que les politiques habituelles se montrent incapables de suivre de tels changements, et qu'il faut donc des méthodes spécifiques.

Nous analysons deux algorithmes se rattachant au paradigme optimiste qui résolvent ce problème : D-UCB et SW-UCB. D-UCB (pour "Discounted-UCB") a été proposé dans l'article [KS06] avec des preuves d'efficacité empiriques, mais n'avait jusqu'alors jamais été analysé. SW-UCB (pour "Sliding-Windows-UCB") est une variante nouvelle qui se montre légèrement plus efficace dans le cadre considéré.

D-UCB s'appuie sur une borne supérieure de confiance $\bar{X}_t(\gamma, i) + c_t(\gamma, i)$ pour l'espérance de la récompense du bras i à l'instant t, où la moyenne empirique escomptée est

$$\bar{X}_t(\gamma, i) = \frac{1}{N_t(\gamma, i)} \sum_{s=1}^{t} \gamma^{t-s} X_s(i) \mathbb{1}\{I_s = i\} \ , \quad \text{avec } N_t(\gamma, i) = \sum_{s=1}^{t} \gamma^{t-s} \mathbb{1}\{I_s = i\},$$

et où le bonus d'exploration est $c_t(\gamma, i) = 2B\sqrt{\xi \log n_t(\gamma)/N_t(\gamma, i)}$, avec $n_t(\gamma) = \sum_{i=1}^{K} N_t(\gamma, i)$, pour un paramètre ξ approprié. Avec ces notations, D-UCB est décrit dans l'algorithme 2.

Algorithme 2 Discounted UCB

Pour tout $t \in \{1, \dots, K\}$ **Faire**
jouer le bras $I_t = t$;
Fin Pour
Pour tout t de $K+1$ à n **Faire**
jouer le bras

$$I_t = \arg\max_{1 \leqslant i \leqslant K} \bar{X}_t(\gamma, i) + c_t(\gamma, i).$$

Fin Pour

Pour estimer l'espérance de la récompense à un instant donné, l'algorithme D-UCB s'appuie sur une moyenne des récompenses passées calculée avec un facteur d'escompte qui donne plus de poids aux observations récentes. Nous proposons une version plus abrupte qui se contente de calculer une moyenne locale, en oubliant les observations trop anciennes. Plus précisément, cet algorithme s'appuie sur la borne de confiance supérieure $\bar{X}_t(\tau, i) + c_t(\tau, i)$. La moyenne empirique locale est

$$\bar{X}_t(\tau, i) = \frac{1}{N_t(\tau, i)} \sum_{s=t-\tau+1}^{t} X_s(i) \mathbb{1}\{I_s = i\} \ , \quad \text{avec } N_t(\tau, i) = \sum_{s=t-\tau+1}^{t} \mathbb{1}\{I_s = i\} \ ,$$

et le bonus d'exploration est $c_t(\tau, i) = B\sqrt{\xi \log(t \wedge \tau)/(N_t(\tau, i))}$, où $t \wedge \tau$ désigne le minimum entre t and τ, où ξ est un paramètre approprié. La politique est décrite dans l'Algorithme 3.

Algorithme 3 Sliding-Window UCB

Pour tout $t \in \{1, \dots, K\}$ **Faire**
jouer le bras $I_t = t$;
Fin Pour
Pour tout t de $K+1$ à n **Faire**
jouer le bras

$$I_t = \arg\max_{1 \leqslant i \leqslant K} \bar{X}_t(\tau, i) + c_t(\tau, i),$$

Fin Pour

Nous prouvons les bornes suivantes pour D-UCB et SW-UCB. Comme précédemment, il suffit d'étudier le nombre de tirages des bras sous-optimaux pour

en déduire immédiatement une borne de regret. Soit i_t^* l'indice du bras optimal à l'instant t (pour simplifier, on suppose qu'il est unique). On note $\tilde{N}_n(i) = \sum_{t=1}^n \mathbb{1}\{I_t = i \neq i_t^*\}$ le nombre de fois que le bras i a été tiré alors qu'il n'était pas optimal au cours des n premiers rounds, puis P_γ et $\mathbb{E}_\gamma$ la loi de probabilité et l'espérance sous la politique D-UCB utilisant le facteur d'escompte γ. Le théorème 1 affirme que, pour les paramètres $\xi \in (1/2, 1)$ et $\gamma \in (1/2, 1)$, pour tout $n \geqslant 1$ et pour tout bras $i \in \{1, \ldots, K\}$:

$$\mathbb{E}_\gamma \left[\tilde{N}_n(i)\right] \leqslant \mathrm{C}_1\, n(1-\gamma) \log \frac{1}{1-\gamma} + \mathrm{C}_2\, \frac{\Upsilon_n}{1-\gamma} \log \frac{1}{1-\gamma}\ ,$$

où

$$\mathrm{C}_1 = \frac{32\sqrt{2}B^2\xi}{\gamma^{1/(1-\gamma)}(\Delta\mu_n(i))^2} + \frac{4}{(1-\frac{1}{\mathrm{e}})\log\left(1+4\sqrt{1-1/2\xi}\right)}$$

et

$$\mathrm{C}_2 = \frac{\gamma-1}{\log(1-\gamma)\log\gamma} \times \log\left((1-\gamma)\xi \log n_K(\gamma)\right)\ .$$

Quand γ tend vers 1, $\mathrm{C}_2 \to 1$ et

$$\mathrm{C}_1 \to \frac{16\,\mathrm{e}\,B^2\xi}{(\Delta\mu_n(i))^2} + \frac{2}{(1-\mathrm{e}^{-1})\log\left(1+4\sqrt{1-1/2\xi}\right)}\ .$$

Suivant la façon dont Υ_n grandit avec n, γ doit être choisi différemment, et l'on obtient des bornes de regret à peu près conformes aux méthodes concurrentes.

L'algorithme SW-UCB montre un comportement similaire, mais l'absence de mémoire le rend un peu plus adapté aux ruptures brusques de l'environnement. En notant P_τ et $\mathbb{E}_\tau$ et la loi de probabilité et l'espérance sous la politique SW-UCB réglée avec la fenêtre de taille τ, la borne suivante est obtenue dans le théorème 2 : pour $\xi > 1/2$, $\tau \leqslant n$ et pour tout bras $i \in \{1, \ldots, K\}$,

$$\mathbb{E}_\tau \left[\tilde{N}_n(i)\right] \leqslant \mathrm{C}(\tau)\frac{n\log\tau}{\tau} + \tau\Upsilon_n + \log^2(\tau)\ ,$$

où

$$\begin{aligned}\mathrm{C}(\tau) &= \frac{4B^2\xi}{(\Delta\mu_n(i))^2}\frac{\lceil n/\tau\rceil}{n/\tau} + \quad \frac{2}{\log\tau}\left\lceil\frac{\log(\tau)}{\log(1+4\sqrt{1-(2\xi)^{-1}})}\right\rceil \\ &\to \frac{4B^2\xi}{(\Delta\mu_n(i))^2} + \frac{2}{\log(1+4\sqrt{1-(2\xi)^{-1}})}\end{aligned}$$

quand τ et n/τ tendent vers l'infini. En outre, le théorème 3 contient une borne inférieure pour la performance de n'importe quel algorithme dans un environnement pouvant changer brutalement un nombre fini de fois : celle-ci montre une la

quasi-optimalité de D-UCB et SW-UCB. Comme corollaire, il apparaît qu'un algorithme efficace dans le cas stationnaire ne peut avoir un regret d'ordre toujours plus faible que $n/\log(n)$ en présence de ruptures. Là encore, les preuves s'appuient sur des inégalités de déviations (quadratiques cette fois) auto-normalisées pour des moyennes escomptées (théorème 4). Ces bornes montrent en particulier qu'il n'y a pas de différence significative pour les regrets entre les cadres stochastiqes et adversariaux, contrairement au cas stationnaire. Dès lors, des comparaisons expérimentales avec les algorithmes soft-max adaptés (en l'occurrence, EXP3.S) sont pertinentes : elles montrent un net avantage pour les méthodes optimistes en terme de réactivité (voir figure 2.2), aux dépens bien sûr d'une moindre robustesse.

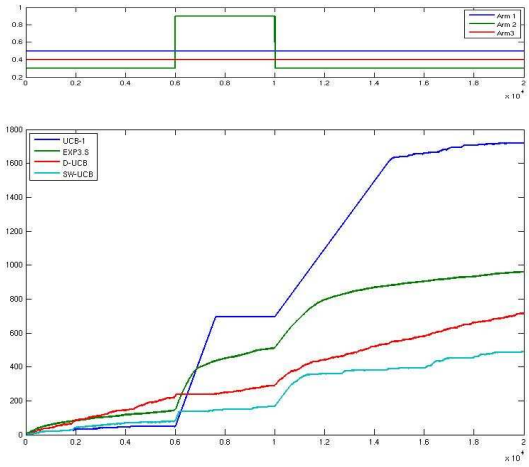


FIGURE 2.2 – Une expérience d'environnement non stationnaire, où le bras 1 est optimal sauf pendant une petite fenêtre de temps ; en bas est représentée l'évolution du regret en fonction du temps.

2.3 Bandits paramétriques : le cas exponentiel canonique

Récemment, des modèles de bandits plus structurés dans lesquels les bras ne sont pas supposés indépendants ont suscité un grand intérêt [DHK08, KSST08, LZ08a, PCA07, WKVP05]. Ces modèles sont motivés par de nombreuses applications, où 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. Les *bandits contextuels* permettent de considérer des modèles de bandits dans lesquels une information contextuelle est connue de l'agent. Cette

classe de problèmes de bandits recouvre deux types de modèles différents. Certains considèrent qu'une information est mise à la disposition de l'agent à chaque instant et que la récompense associée à chaque bras dépend de cette information [KSST08, LZ08a, WKVP05]. Dans ce cas, le bras optimal dépend de l'information contextuelle et est donc susceptible de changer à chaque instant. Nous nous intéressons à un autre type de bandits contextuel où une information, connue de l'agent au préalable, est associée à chaque bras. Le modèle linéaire, pour lequel les récompenses des bras ont une espérance qui dépend linéairement d'un vecteur de paramètres, a fait l'objet de plusieurs travaux récents [DHK08, RT08]. Comme dans le problème de bandits classique, les récompenses sont tirées de manières indépendantes selon une distribution inconnue ; néanmoins l'espérance de la récompense conditionnellement à l'action a est de la forme $m_a'\theta$ où m_a est l'information contextuelle représentée par un vecteur de $\mathbb{R}^d$ connu de l'agent et $\theta \in \mathbb{R}^d$ est un paramètre inconnu commun à toutes les actions.

Dans l'article [FCGS10], Sarah Filippi, Olivier Cappé, Csaba Szepesvari et moi étudions le cadre plus large des modèles linéaires généralisés dans lequel l'espérance de la récompense conditionnellement à l'action a est de la forme $\mu(m_a'\theta)$ où μ est une fonction croissante non linéaire. Ce type de modèles présente notamment l'avantage de conduire à des estimateurs $\hat{\theta}$ qui sont solutions de problèmes d'optimisation strictement convexes. Ces modèles permettent également d'aborder des cas intéressants où les récompenses sont des variables catégorielles ou binaires à travers, par exemple, la régression logistique. Une application très en vogue du modèle logistique est le problème d'optimisation de ressources publicitaires sur Internet. En effet, dans le modèle de facturation dit *pay per click*, le revenu est directement fonction du fait que l'utilisateur clique ou non sur l'annonce publicitaire qui lui est présentée [JS08], ce qui conduit à une récompense par utilisateur de type binaire. 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 problème de bandits contextuels, la sélection d'annonces jouant le rôle de bras et l'information contextuelle correspondant à des caractéristiques des annonces (par exemple une catégorisation sémantique : "sport", "cinéma", "loisirs", etc.). Dans ce type d'applications, la dimension d du vecteur de caractéristiques est typiquement petite par rapport au nombre K d'options de composition de pages disponibles (en particulier si chaque page comporte plusieurs annonces publicitaires choisies dans un pool d'annonces). Dans ce contexte, la modélisation de la loi des récompenses par une régression logistique semble assez naturellle, et en tout cas certainement plus naturellle que l'utilisation d'un simple modèle de régression linéaire qui ignore la nature binaire des récompenses. Nous présentons dans l'article des simulations montrant que les algorithmes basés sur les deux types de modèles sont effectivement susceptibles de conduire à des performances différentes.

Nous proposons un nouvel algorithme optimiste, GLM-UCB, qui étend à ce cadre l'approche d'UCB. L'idée est de construire directement une borne de confiance pour $\mu(m'_a\theta)$ sans passer par la construction d'une région de confiance pour θ. Nous présentons une analyse théorique des performances de cet algorithme en termes de regret à horizon fini, c'est-à-dire de comparaison entre l'espérance des récompenses reçues en suivant l'algorithme et l'espérance des récompenses reçues en jouant en permanence le bras optimal. La borne dépendant du problème est une fonction poly-logarithmique de l'horizon (théorème 1), tandis que la borne minimax est d'ordre $\sqrt{n}\log(n)^{3/2}$ (théorème 2). Encore une fois, ce sont des inégalités pour les déviations auto-normalisées qui font marcher la preuve (en l'occurrence, nous utilisons des résultats pour les martingales vectorielles de [DLPKL04]). En particulier, nous montrons que les performances dépendent de la dimension du vecteur de paramètres mais pas du nombre de bras. Pour les applications pratiques, nous proposons une façon de simple de régler les paramètres de l'algorithme basée sur des arguments asymptotiques (section 4.2), pour laquelle des garanties théoriques semblent hors d'atteinte, mais dont des simulations numériques montrent la pertinence en particulier dans le cas de la régression logistique.

2.4 Processus de décision markoviens

Les processus de décision markoviens (MDP) constituent un modèle stochastique plus riche, dans lequel l'agent se trouve à chaque instant dans un état évoluant de façon markovienne selon une loi de probabilité qui dépend de l'action choisie ; la récompense reçue est elle aussi une fonction aléatoire de l'état et de l'action. L'objectif est de trouver une politique optimale, c'est-à-dire une fonction (potentiellement randomisée) qui à chaque état associe l'action à choisir pour maximiser l'espérance des récompenses à venir, en moyenne ou avec un facteur d'escompte. La théorie de la programmation dynamique développée par Bellmann ([Bel56], voir [Ber95, BT96, Put94] pour des présentations complètes et modernes) caractérise les politiques optimales à travers les *équations de Bellman* et donne plusieurs façons, quand on connaît les paramètres du MDP, de résoudre celles-ci. Mais quand les paramètres (lois de transition, fonctions de récompense) sont inconnus, il n'existe pas de solution générale au problème qui, alors, généralise les modèles de bandits vus ci-dessus (ceux-ci pouvant être considérés comme des MDP avec un seul état). Avec Olivier Cappé et Sarah Filippi, nous avons exploré plusieurs pistes avant de nous intéresser spécialement aux algorithmes utilisant le paradigme optimiste (voir [AO07] et les références qu'il contient).

Considérons un MDP $\mathcal{M} = (\mathsf{S}, \mathsf{A}, P, r)$ à espace d'état S fini et à espace d'action A également fini. Notons $S_t \in \mathsf{S}$ et $A_t \in \mathsf{A}$ l'état dans lequel se trouve

l'agent et l'action choisie par lui au temps t. La probabilité qu'il saute de l'état S_t à l'état S_{t+1} est alors notée $P(S_{t+1}; S_t, A_t)$, et la récompense qu'il reçoit à l'instant t est la variable aléatoire $X_t \in [0,1]$ dont nous noterons l'espérance $r(S_t, A_t)$. L'objectif de l'agent est de choisir une séquence d'action qui maximise la somme des récompenses reçues ; ses choix sont résumés dans un politique stationnaire $\pi : \mathsf{S} \to \mathsf{A}$. Nous nous sommes essentiellement concentrés sur les MDP communicants, c'est-à-dire pour lesquels il existe une politique permettant de passer de n'importe quel état à n'importe quel autre en un temps d'espérance finie. Pour ces MDP, on sait (voir par exemple [Put94]) que la récompense moyenne [2] sous la politique stationnaire π est

$$\rho^\pi(\mathcal{M}) = \lim_{n\to\infty} \frac{1}{n} \mathbb{E}_{\mathcal{M},\pi} \left(\sum_{t=0}^{n} X_t \right) ,$$

et qu'elle ne dépend pas de l'état de départ. Notons $\pi^*(\mathcal{M}) : \mathsf{S} \to \mathsf{A}$ la polique optimale, et $\rho^*(\mathcal{M})$ la récompense moyenne associée, dont la dépendance dans le modèle $\mathcal{M}$ est soulignée par les notations : $\rho^*(\mathcal{M}) = \sup_\pi \rho^\pi(\mathcal{M}) = \rho^{\pi^*(\mathcal{M})}(\mathcal{M})$. Les équations d'optimalité de Bellman assurent que pour tout état $s \in \mathsf{S}$,

$$h^*(\mathcal{M}, s) + \rho^*(\mathcal{M}) = \max_{a\in\mathsf{A}} \left(r(s,a) + \sum_{s'\in\mathsf{S}} P(s'; s, a) h^*(\mathcal{M}, s') \right) ,$$

où le vecteur $h^*(\mathcal{M})$, de dimension $|\mathsf{S}|$, est appelé vecteur de biais, et n'est défini qu'à une constante près. Pour un MDP $\mathcal{M}$ fixé, la politique optimale $\pi^*(\mathcal{M})$ peut être trouvée en résolvant l'équation d'optimalité, et en définissant pour tout $s \in \mathsf{S}$,

$$\pi^*(\mathcal{M}, s) \in \arg\max_{a\in\mathsf{A}} \left(r(s,a) + \sum_{s'\in\mathsf{S}} P(s'; s, a) h^*(\mathcal{M}, s) \right) .$$

En pratique, la récompense moyenne optimale peut être calculée par exemple par l'algorithme d'itération sur les valeurs ([Put94]).

Nous nous sommes intéressés au cas où l'agent ne connaît pas les paramètres du MDP, c'est-à-dire les lois de transition et les lois de récompense, mais où il se sert d'estimations de celles-ci pour diriger ses choix : c'est ce que l'on appelle l'apprentissage basé sur le modèle, par opposition à des méthodes qui cherchent à évaluer directement l'effet à long terme du choix d'une action dans un état. Notons $\hat{P}_t(s'; s, a)$ l'estimateur du maximum de vraisemblance au temps t de la probabilité de transition de l'état s à l'état s' conditionnellement à l'action a, et notons $\hat{r}_t(s,a)$ la récompense moyenne observée s quand l'action a a été choisie :

$$\hat{P}_t(s'; s, a) = \frac{N_t(s,a,s')}{N_t(s,a)}, \quad \hat{r}_t(s,a) = \frac{\sum_{k=0}^{t-1} X_k \mathbb{1}\{S_k = s, A_k = a\}}{N_t(s,a)} ,$$

2. Nous ne parlerons ici que de récompense moyenne, et pas de récompense cumulée escomptée.

où $N_t(s,a,s') = \sum_{k=0}^{t-1} \mathbb{1}\{S_k = s, A_k = a, S_{k+1} = s'\}$ est le nombre de visites antérieures au temps t de l'état s qui ont été suivies par des visites à l'état s' alors que l'action a avait été choisie, et où $N_t(s,a) = \sum_{k=0}^{t-1} \mathbb{1}\{S_k = s, A_k = a\}$. Il est connu que la politique optimale dans le modèle estimé $\widehat{\mathcal{M}}_t = (\mathsf{S}, \mathsf{A}, \hat{P}_t, \hat{r}_t)$ n'est en général pas un bon choix, du fait de possibles erreurs d'estimation qui ne seront jamais compensées par de nouvelles observations si celles-ci ne sont pas provoquées : on sait que les politiques d'exploitation pure échouent avec une probabilité non nulle. Pour éviter cet écueil, les algorithmes *optimistes basés sur les modèles* considèrent l'ensemble $\mathcal{M}_t$ des MDP qui rendent les observations suffisamment vraisemblables (dont $\widehat{\mathcal{M}}_t$ fait notamment partie), et choisissent dans cet ensemble celui qui offre la plus grande récompense moyenne, et proposent de choisir son action conformément à une politique optimale de ce "MDP optimiste". L'ensemble $\mathcal{M}_t$ est choisi ainsi :

$$\mathcal{M}_t = \{\mathcal{M} = (\mathsf{S}, \mathsf{A}, P, r) : \forall s \in \mathsf{S}, \forall a \in \mathsf{A}, |\hat{r}_t(s,a) - r(s,a)| \leqslant \epsilon_R(s,a,t) \\ \text{et } d(\hat{P}_t(.;s,a), P(.;s,a)) \leqslant \epsilon_P(s,a,t)\} \ ,$$

où d désigne une pseudo-métrique sur l'espace des transitions de probabilités. Les *rayons de voisinage* $\epsilon_R(s,a,t)$ et $\epsilon_P(s,a,t)$ autour de la récompense estimée $\hat{r}_t(s,a)$ et des probabilités de transitions estimées $\hat{P}_t(.;s,a)$, décroissent avec $N_t(s,a)$. De telles méthodes ont été étudiées, notamment dans les articles [AO07, JOA10, BT09], où est choisie pour d une distance L^1 qui s'avère pratique à la fois dans le calcul du MDP optimiste et dans les preuves de regret. Dans notre article [FCG10], nous défendons l'idée, déjà présente dans l'article [BK97] (qui toutefois, contrairement à nous, supposait connue la structure du MDP), que la divergence de Kullback-Leibler fournit une bien meilleure pseudo-distance, tant en théorie qu'en pratique. Nous montrons en effet que ce choix conduit à un comportement significativement différent de l'algorithme, et à une performance améliorée, tandis que l'impact sur la complexité algorithmique est assez limité. La politique KL-UCRL, décrite dans l'algorithme 4, est donc une variante de ces méthodes. Elle s'appuie sur l'étape-clé qu'est la recherche du modèle optimiste (étape 8), détaillée ci-dessous dans l'algorithme 5.

KL-UCRL procède par phases : il est en effet nécessaire de suivre une politique pendant un certain temps avant de pouvoir en évaluer les effets, et il apparaît dans l'analyse que la présence de phases est difficilement évitable. En appelant t_j l'instant de début de la j-ième phase, la durée de celle-ci dépend du nombre de visites $N_{t_j}(s,a)$ dans chaque paire état-action (s,a) avant t_j, comparé au nombre $n_j(s,a)$ de visites à cette même paire pendant la j-ième phase. Plus précisément, un épisode s'achève dès que $n_j(s,a) \geqslant N_{t_j}(s,a)$ pour une quelconque paire état-action (s,a).

La politique π_j, suivie pendant cette phase, est la politique optimale du MDP

Algorithme 4 KL-UCRL

Initialisation : $j = 0$, $t_0 = 0$; $\forall a \in \mathsf{A}, \forall s \in \mathsf{S}, n_0(s,a) = 0$, $N_0(s,a) = 0$; politique initiale π_0.
Pour tout $t \geqslant 1$ **Faire**
  Observer S_t
  Si $n_j(S_t, \pi_j(S_t)) \geqslant \max(N_{t_j}(S_t, \pi_j(S_t)), 1)$ **Alors**
    Commencer une nouvelle phase : $j = j + 1$, $t_j = t$,
    Réinitialiser : $\forall a \in \mathsf{A}, \forall s \in \mathsf{S}$, $n_j(s,a) = 0$
    Calculer les estimateurs $\hat{P}_t$ et $\hat{r}_t$
    Trouver le modèle optimisite $\mathcal{M}_j \in \mathcal{M}_t$ et la politique correspondante π_j solution de l'équation (2.4) en utilisant l'algorithme 5
  Fin Si
  Choisir l'action $A_t = \pi_j(S_t)$
  Recevoir la récompense X_t
  Mettre à jour les compteurs de la phase courante :

$$n_j(S_t, A_t) = n_j(S_t, A_t) + 1$$

  Mettre à jour les compteurs globaux :

$$N_t(S_t, A_t) = N_{t-1}(S_t, A_t) + 1$$

Fin Pour

optimiste $\mathcal{M}_j = (\mathsf{S}, \mathsf{A}, P_j, r_j) \in \mathcal{M}_{t_j}$. Ce dernier est calculé en résolvant les *équations d'optimalité étendue* : pour tout $s \in \mathsf{S}$

$$h^*(s) + \rho^* = \max_{P,r} \max_{a \in \mathsf{A}} \left(r(s,a) + \sum_{s' \in \mathsf{S}} P(s';s,a) h^*(s') \right) \tag{2.4}$$

où le maximum est pris sur tous les paramètres P et r tels que

$$\forall s, \forall a, \quad KL(\hat{P}_{t_j}(.;s,a), P(.;s,a)) \leqslant \frac{C_P}{N_{t_j}(s,a)} ,$$
$$\forall s, \forall a, \quad |\hat{r}_{t_j}(s,a) - r(s,a)| \leqslant \frac{C_R}{\sqrt{N_{t_j}(s,a)}} ,$$

et où C_P et C_R désignent deux constantes qui contrôlent la taille des régions de confiance. La matrice de transition P_j et la récompense moyenne r_j du MDP optimiste $\mathcal{M}_j$ maximisent ces équations. Pour les calculer, la procédure d'*itération étendue sur les valeurs* est utilisée afin de résoudre l'équation de point fixe (2.4) (voir [Put94, AJO09]), comme décrit dans l'algorithme 5. A chaque étape de l'itération étendue sur les valeurs, le problème de maximisation (2.4) doit être résolu. Pour tout état s et pour toute action a, la maximisation de $r(s,a)$ sous la contrainte $|\hat{r}_{t_j}(s,a) - r(s,a)| \leqslant C_R/\sqrt{N_{t_j}(s,a)}$ est évidemment résolue en choisissant

$$r(s,a) = \hat{r}_{t_j}(s,a) + C_R/\sqrt{N_{t_j}(s,a)} ,$$

de sorte que la difficulté principale réside dans la maximisation du produit scalaire entre le vecteur de probabilité $q = P(.;s,a)$ et le vecteur de valeur $V = h^*$ sur la boule de Kullback-Leibler centrée autour du vecteur des probabilités de transition estimées $p = \hat{P}_{t_j}(.;s,a)$:

$$\max_{q \in \mathcal{S}} V'q \quad \text{s.t.} \quad KL(p,q) \leqslant \epsilon , \tag{2.5}$$

où V' désigne la transposée de V et où $\mathcal{S}$ est le simplexe de dimension $|S| - 1$. Le rayon $\epsilon = C_P/N_{t_j}(s,a)$ contrôle la taille de la région de confiance. Ce problème de maximisation convexe peut être résolu numériquement : on se voit ramené à trouver l'unique valeur ν pour laquelle $f(\nu) = \epsilon$, où f est la fonction convexe strictement décroissante définie de la façon suivante : pour tout $\nu \geqslant \max_{i \in \bar{Z}} V_i$, en notant $\bar{Z} = \{i : p_i > 0\}$,

$$f(\nu) = \sum_{i \in \bar{Z}} p_i \log(\nu - V_i) + \log \left(\sum_{i \in \bar{Z}} \frac{p_i}{\nu - V_i} \right) . \tag{2.6}$$

Comme on peut l'initialiser intelligemment, l'algorithme de Newton fournit efficacement la solution en quelques itérations.

Algorithme 5 MaxKL

Entrées: Vecteur valeur V, vecteur de probabilités p, réel $\epsilon > 0$
Sorties: Vecteur de probabilités q solution de (2.5)

Soit $Z = \{i : p_i = 0\}$, $\bar{Z} = \{i : p_i > 0\}$ et $I^* = Z \cap \arg\max_i V_i$
Si $I^* \neq \emptyset$ et s'il existe $i \in I^*$ tel que $f(V_i) < \epsilon$ **Alors**
    Soit $\nu = V_i$ et $r = 1 - \exp(f(\nu) - \epsilon)$.
    Choisir les q_i pour tout $i \in I^*$ tels que

$$\sum_{i \in I^*} q_i = r \ .$$

    Pour tout $i \in Z/I^*$, $q_i = 0$.
Sinon
    Pour tout $i \in Z$, $q_i = 0$, et $r = 0$.
    Trouver ν tel que $f(\nu) = \epsilon$.
Fin Si
Pour tout $i \in \bar{Z}$, prendre

$$q_i = \frac{(1-r)\tilde{q}_i}{\sum_{i \in \bar{Z}} \tilde{q}_i} \ ,$$

avec $\tilde{q}_i = \frac{p_i}{\nu - V_i}$.

Conformément à l'usage (voir [JOA10]), nous mesurons la performance de l'algorithme KL-UCRL par le regret R_n qui mesure, au temps n, la différence entre les récompenses cumulées et la meilleure politique en moyenne :

$$R_n = \sum_{t=1}^{n} (\rho^*(\mathcal{M}) - X_t) \ .$$

Nous avons adapté les preuves de regret connues pour l'algorithme UCRL2 au cas des voisinages de Kullback-Leibler, et nous obtenons des résultats similaires sans réussir à les améliorer sur ce plan. Soit

$$D(\mathcal{M}) = \max_{s,s'} \min_{\pi} \mathbb{E}_{\mathcal{M},\pi}(\tau(s,s')) \ ,$$

où $\tau(s,s')$ désigne le temps d'atteinte de l'état s', en partant de l'état s. La constante $D(\mathcal{M})$ (parfois appelée rayon de communication du MDP) apparaît dans les bornes de regret. Avec un bon réglage des constantes d'exploration C_P et C_R, le théorème 1 qu'avec probabilité $1-\delta$, pour tout $n > 5$, le regret de l'algorithme KL-UCRL satisfait

$$R_n \leqslant C D(\mathcal{M}) |\mathsf{S}| \sqrt{|\mathsf{A}| n \log(\log(n)/\delta)} \ ,$$

pour une constante $C \leqslant 24$ qui ne dépend pas du modèle.

Comme pour les problèmes de bandits, il est également possible de prouver une borne logarithmique pour le regret en espérance. Cette borne, décrite dans le Théorème 2, dépend du modèle à travers la constante $\Delta(\mathcal{M}) = \rho^*(\mathcal{M}) - \max_{\pi, \rho^\pi(\mathcal{M})<\rho^*(\mathcal{M})} \rho^\pi(\mathcal{M})$, qui mesure l'écart entre les politiques optimales et les politiques sous-optimales ; elle s'écrit :

$$\mathbb{E}(R_n) \leqslant CD^2(\mathcal{M}) \frac{|\mathsf{S}|^2|\mathsf{A}| \log(n)}{\Delta(\mathcal{M})} + C(\mathcal{M}) \ ,$$

où $C(\mathcal{M})$ dépend du modèle (voir [JOA10]).

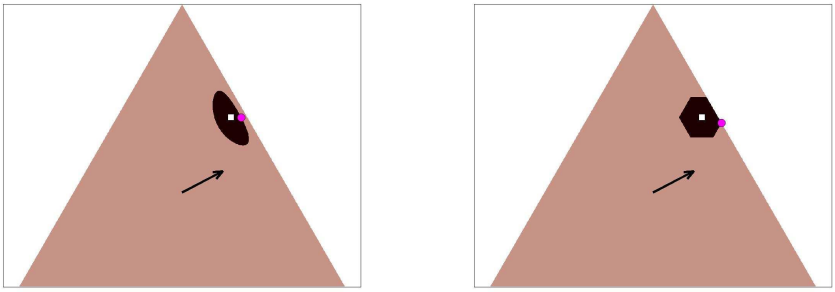


FIGURE 2.3 – Solution du problème de maximisation (2.5) pour des voisinages de Kullback-Leibler (gauche) et L^1 (droite) ; la flêche indique la direction de V, le point blanc représente p alors que le maximum de $V'q$ sur le voisinage (en noir) est atteint au point violet.

Contrairement à ce que l'on peut faire pour les problèmes de bandits les plus simples, toutes les bornes existantes pour ce type de problème s'avèrent en pratique très pessimistes, et ne reflètent la performance de KL-UCRL que qualitativement. C'est d'ailleurs la raison fondamentale pour laquelle l'intérêt de l'utilisation d'une pseudo-métrique de Kullback-Leibler ne peut être aussi clairement mise en évidence que, par exemple, dans cas des bandits bornés cité ci-dessus. En revanche, les expérimentations numériques sont très convaincantes. Nous avons fait différents essais sur des "benchmarks" bien connus dans la littérature, en particulier les environnements *RiverSwim* et *SixArms* proposés dans l'article [SL08] pour illustrer l'importance de l'exploration. L'utilisation de la divergence de Kullback-Leibler améliore très significativement à la fois le regret subi et aussi le temps pris par l'algorithme avant de se consacrer essentiellement à la stratégie optimale. Pour comprendre cette amélioration, nous avons plusieurs éléments : d'abord, la régularité des voisinages de Kullback-Leibler donne à la réponse de l'algorithme une continuité en fonction des observations qui fait défaut aux voisinages L^1.

Ensuite, il apparaît que l'optimisme d'UCLR peut l'amener à utiliser dans son MDP optimiste des probabilités nulles pour des transitions qui ont pourtant été observées (violant ainsi dans l'esprit le principe d'optimisme, voir figure 2.3). Enfin, l'optimisme d'UCRL l'enjoint à toujours augmenter la probabilité d'une transition de n'importe quel état vers celui qui est jugé le plus prometteur, alors que KL-UCB résout en pareil cas un arbitrage entre les preuves statistiques accumulées contre l'existence d'une telle transition et l'avantage qui pourrait en être tiré.

2.5 Observations partielles et "channel sensing"

Les processus de décision markoviens forment une classe suffisamment large pour modéliser un grand nombre de situations, de la gestion de stocks à la robotique en passant par la gestion des réseaux (voir par exemple [Ber95, BT96] et les références citées). Mais il existe aussi des situations où l'agent n'observe pas directement l'état dans lequel il se trouve, et doit se contenter d'observations indirectes. C'est notamment le cas dans le problème dit de l'accès spectral opportuniste pour la *radio cognitive* (voir [AWYVM08, Hay05, Mit00]), où l'on cherche à améliorer l'efficacité spectrale en faisant un meilleur usage des larges portions de bandes de fréquences qui restent inutilisées. Un des objectifs de la thèse de Sarah Filippi était d'élaborer des stratégies pour des utilisateurs secondaires de réseaux qui souhaitent utiliser des ressources sans avoir le droit de perturber les utilisateurs primaires. Ce problème a été modélisé de la façon suivante [LZ08b, ZTSC07] : à chaque instant t, un agent dispose de N canaux indépendants ; l'état $S_t(i)$ du canal numéro i est modélisé par une chaîne de Markov qui peut prendre les valeurs "libre" ou "occupée". L'agent souhaite utiliser ces canaux, mais ne peut transmettre un paquet que lorsque le canal qu'il choisi est libre ; à cause de limites matérielles et énergétiques, il ne peut pas tester l'état de tous les canaux à la fois [LEGJVP08, LZ08b, ZKL08] : il doit en choisir un à chaque instant, et n'observe donc pas totalement l'état du système. Dans ce modèle, l'accès au réseau peut être considéré comme un processus de décision markovien partiellement observé (POMDP), et dans ce cas particulier également comme un problème de bandits ininterrompus [LZ08b, ZTSC07].

Papadimitriou et Tsitsiklis [PT94] ont établi que la planification optimale dans ce problème est PSPACE-dure, et qu'il est donc sans espoir de trouver une politique optimale utilisable. Néanmoins, des publications récentes se sont focalisées sur des politiques d'indice [GM07, LNDF08, LZ08b] qui ont un coût algorithmique réduit tout en offrant des performances guère éloignées de l'optimum. Ces politiques d'indice sont inspirées des travaux de Whittle [Whi88], eux-mêmes inspirés pas les travaux de Gittins [Git79] évoqués plus haut au sujet des bandits

classiques : l'idée est toujours de calculer un indice par canal, et de choisir celui qui possède le plus grand indice. Pour calculer ces indices de Whittle, on doit résoudre un problème de planification dans un modèle plus simple contenant un seul canal, mais faisant également apparaître un coût d'utilisation [LNDF08, LZ08b] pour l'action d'observer le canal, qui est optionnelle.

Avec Sarah Filippi et Olivier Cappé, nous nous sommes donc intéressés à ce modèle à un canal avec observation coûteuse optionnelle. Il apparaît que les stratégies d'observation sont alors calculables, bien que non triviales. Mais les auteurs des travaux précédents faisaient le plus souvent l'hypothèse que l'on connaissait la loi d'évolution du canal par son utilisateur principal, c'est-à-dire les lois de transition entre les états des canaux [FCCM08, LZ08b, ZKL08, ZTSC07]. Cela est loin d'être toujours le cas en pratique : il faut que l'agent les apprenne au fur et à mesure, et qu'il puisse avoir une stratégie efficace même sans connaissance a priori des paramètres des canaux. Dans l'article [LGX+08] une règle heuristique basée sur des considérations asymptotiques d'estimateurs de ces paramètres avait été proposée. [LEGJVP08] avait également considéré ce problème, mais dans le cas plus simple (et peu réaliste) où les canaux sont sans mémoire.

On peut génériquement réduire les POMDP à des processus de décision markoviens, en choisissant pour espace d'état de ces derniers l'ensemble des lois de probabilités sur les états du POMDP : l'idée est de maintenir à jour un "état de croyance" comme pour le filtrage des chaînes de Markov cachées. Malheureusement, même pour le problème du canal que nous considérons ici, le problème d'apprentissage dans le MDP correspondant est trop complexe pour être résolu par les méthodes comme KL-UCRL vues ci-dessus. Une autre approche est donc nécessaire. Nous avons établi une politique procédant en deux temps : pendant une phase d'exploration, l'agent cherche à estimer suffisamment le canal pour déterminer sa politique, qu'il applique ensuite pendant la phase d'exploitation. Le problème principal est de trouver le bon équilibre entre exploration et exploitation, et donc d'achever la première phase dès que des éléments statistiques suffisants ont été accumulées. Cette méthode n'est pas spécifique au problème du canal : elle peut en fait s'appliquer à tout POMDP pour lequel

- l'état du canal est partiellement observable par une action d'observation,
- le POMDP est paramétré par un vecteur de petite dimension,
- et les transitions ne dépendent pas des actions choisies.

Les conditions précises (existence d'un estimateur $\sqrt{n}-$consistant, etc...) sont précisées dans notre article [FCG11]. L'*algorithme de pavage* que nous proposons a un principe assez simple : il consiste à diviser l'espace des paramètres en régions dans lesquelles on connaît la politique optimale. Pendant la phase d'exploration, on construit une région de confiance pour le paramètre qui se réduit au fur et à mesure : cette phase dure tant que la région de confiance n'est pas :

- soit entièrement incluse dans une de ces régions,

– soit entièrement incluse dans une des zones-tampons qui entourent les frontières entre ces régions (voir figure 2.4).

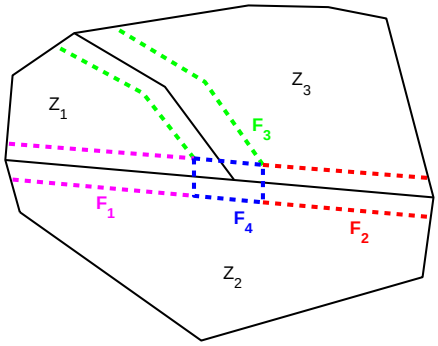


FIGURE 2.4 – Pavage de l'espace des paramètres pour un exemple présentant trois zones distinctes de politiques optimales, séparées par des zones-tampons.

Pour analyser l'algorithme, il faut donc contrôler le temps (aléatoire) que durera l'exploration, et conjointement régler l'épaisseur des zones-tampons. Pour en analyser la performance, nous considérons (comme d'habitude) le regret qui, au temps n, compare en moyenne les récompenses accumulées avec ce qu'un "oracle" connaissant la politique optimale à l'avance aurait obtenu. Nous montrons dans le théorème 1 des bornes de regret d'ordre $C(M)\log(n)$, où $C(M)$ dépend des paramètres du POMDP, ainsi que des bornes indépendantes des paramètres en $O\left(n^{2/3}(\log(n))^{1/3}\right)$. A notre connaissance, ce sont les seules disponibles pour ce problème. Des simulations numériques tendent à confirmer l'intérêt d'un tel algorithme comparativement aux autres algorithmes proposés pour le problème du canal.

2.6 Un algorithme optimiste pour la recherche de nouveauté

Ce chapitre s'achève par la description d'un problème qui ne relève pas réellement de l'apprentissage par renforcement, et qui ne présente qu'une certaine ressemblance apparente avec un problème de bandits, mais pour lequel Sébastien Bubeck (université de Princeton) et moi proposons une solution directement inspirée du paradigme optimiste dont nous montrons la surprenante efficacité. Ce problème nous a été posé lors des journées MAS de Bordeaux par Damien Ernst, chercheur en ingénierie électrique à l'Université de Liège, qui cherche à détecter des configurations problématiques dans des réseaux. Pour cela, il dispose de plusieurs simulateurs permettant d'échantilloner de façon aléatoire dans l'ensemble

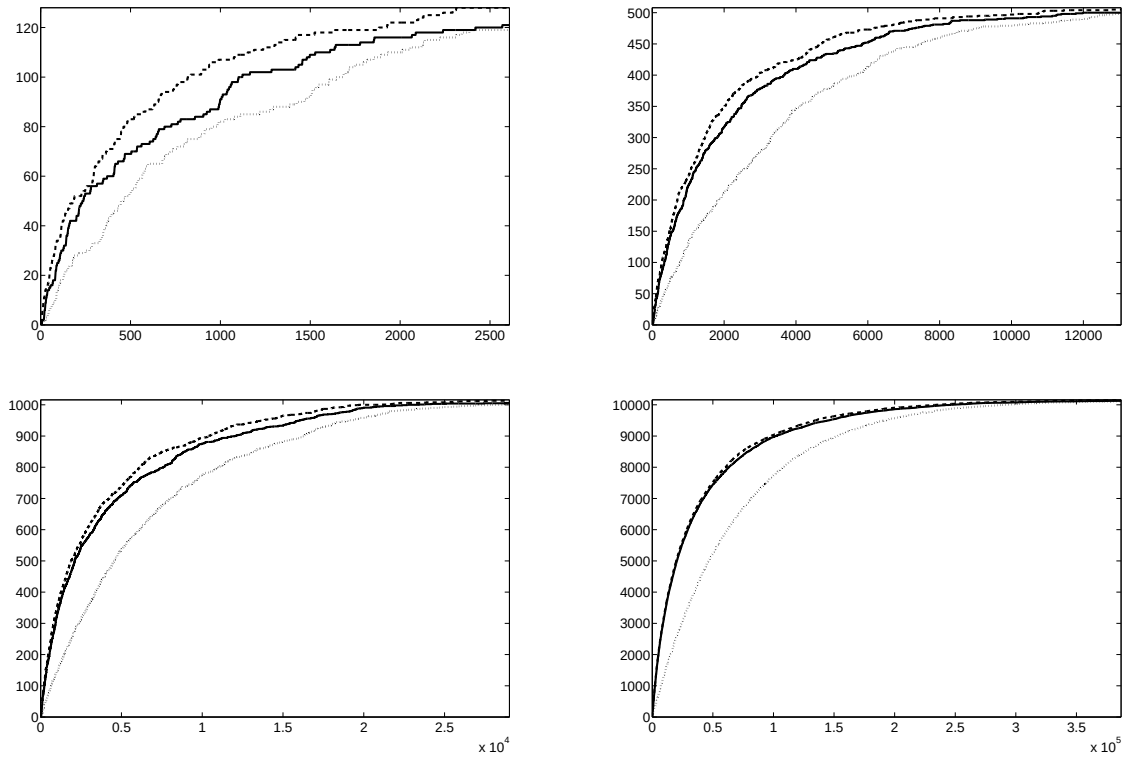
gigantesque des configurations possibles selon des lois plus ou moins concentrées sur les configurations recherchées. Une fois une configuration simulée, il peut tester rapidement si celle-ci est effectivement problématique ou pas. Son problème est de planifier ses appels aux différents simulateurs intelligemment, afin pour trouver un maximum de configurations problématiques différentes avec un minimum de simulations, alors qu'il ne connaît pas l'efficacité de chaque simulateur a priori. La principale différence avec un problème de bandits est qu'une fois une configuration trouvée il est parfaitement inutile de la retrouver : un simulateur qui fournirait constamment la même configuration (fût-elle recherchée) ne mériterait pas d'être utilisé plus d'une fois. Malgré cela, un algorithme d'inspiration optimiste s'avère ici remarquablement efficace.

Nous modélisons la situation de la façon suivante : supposons que $P^1, \dots, P^K$ sont des lois de probabilités de supports disjoints sur un ensemble $\mathcal{X}$ dénombrable dont un sous-ensemble B est constitué d'éléments remarquables. La probabilité de B sous chacune des lois P^i n'est pas connue : on sait seulement tester, pour chaque élément de x, son appartenance à B. On considère des variables $(X_t^i)_{1 \leqslant i \leqslant K, t \geqslant 1}$ indépendantes et telles que X_t^i suit la loi P_i. A chaque instant t, il faut choisir (en utilisant les observations passées) l'indice I_t d'une distribution afin de maximiser le nombre cumulé d'éléments remarquables trouvés parmi n tirages $S_n = \{X_1^{I_1}, \dots, X_n^{I_n}\}$.

L'idée de l'algorithme Good-UCB que nous proposons dans l'article [BEG11] est la suivante : grâce à l'estimateur de Good-Turing ([Goo53], voir aussi [MS00] pour une étude du risque ou [GS95] pour des applications en linguistique), on peut estimer pour chaque distribution $i \in \{1, \dots, K\}$ la "masse manquante" $P^i(B \setminus S_n)$ par un estimateur dont, grâce à l'inégalité de McDiarmid [McD89] ou pour plus de précision de Boucheron, Massart et Lugosi [BLM09], on peut contrôler le risque. Plutôt que d'utiliser directement cet estimateur, le principe d'optimisme enjoint de choisir à chaque instant la distribution pour laquelle une borne de confiance supérieure de la masse manquante est la plus grande.

Cet algorithme s'avère expérimentalement remarquablement efficace, faisant pour chaque valeur de n presque aussi bien qu'une stratégie "oracle" qui connaîtrait à chaque instant toutes les masses manquantes. Pour analyser ce phénomène, il nous a fallu nous détacher des schémas de preuve que l'on trouve dans l'analyse des problèmes de bandits, et même d'une analyse en termes de regret cumulé. En effet, on observe ici que l'écart avec la performance de l'oracle ne grandit pas avec n : il se stabilise très vite, et finit par s'estomper quand tous les éléments de B ont été trouvés. Cela s'explique notamment par le fait que les erreurs commises par un algorithme, qui à un instant donné ne choisirait pas la meilleure distribution de tirage, tendent non à se cumuler mais plutôt à se *compenser* automatiquement. Nous montrons, dans le cas particulier où les lois P^i sont uniformes sur leurs supports respectifs, la convergence uniforme presque-sûre de la proportion d'élé-

FIGURE 2.5 – Nombre de configurations recherchées trouvées, au cours du temps, par l'algorithme Good-UCB (trait plein), par la stratégie oracle (tirets), et par échantillonage équilibré (pointillés) des $K = 7$ distributions, pour des ensembles $\mathcal{X}$ de tailles $N = 128, N = 500, N = 1000$ et $N = 10000$.



ments remarquables trouvés par l'algorithme Good-UCB vers celle que trouve la stratégie oracle, dans une limite macroscopique illustrée à la figure 2.6 où la taille de l'ensemble $\mathcal{X}$ tend vers l'infini alors que la proportion d'éléments remarquables dans le support de chaque distribution se stabilise.

Chapitre 3

Filtrage particulaire et chaînes de Markov cachées

Les modèles de Markov cachés en général, et le filtrage particulaire en particulier, sont depuis plus de dix ans une spécialité du LTCI qui m'a accueilli après ma thèse. Grâce à l'enthousiasme d'Eric Moulines, Randal Douc et Steffen Barembruch (ce dernier effectuant sa thèse en collaboration avec la DGA), j'ai pu participer à la progression à la fois opérationnelle et théorique de leur compréhension.

3.1 Méthodes de Monte-Carlo séquentielles pour les chaînes de Markov cachées

L'inférence statistique dans les modèles de Markov cachés (HMM pour Hidden Markov Models, voir l'ouvrage de référence [CMR05]) repose sur la possibilité de calculer la loi a posteriori $\phi_{0:n}$ des états cachés $X_{0:n}$ conditionnellement aux observations $Y_{0:n}$, ou bien au moins certaines de ses marginales comme, pour $0 \leqslant t \leqslant n$, la loi $\phi_{t|n}$ de X_t conditionnellement à $Y_{0:n}$. Ces dernières se déduisent généralement des lois de filtrage ϕ_t de l'état caché X_t conditionnellement aux observations antérieures $Y_{0:t}$, par une opération de lissage. Ces calculs peuvent être effectués sans erreur principalement dans deux cas : dans le cas (étudié par Kalman) du modèle gaussien linéaire, et dans le cas où l'espace d'état $\mathcal{X}$ de la chaîne de Markov cachée est fini. Pour ce dernier cas, le célèbre algorithme de Baum-Welch peut être utilisé : c'est l'outil sur lequel Antoine Chambaz, Elisabeth Gassiat et moi nous étions appuyés pour le problème de l'estimation par pseudo-maximum de vraisemblance pénalisé [CGG09].

Mais en dehors de ces cas particulier, ou même, dans le cas fini, quand le nombre d'états possibles est très grand, des stratégies d'approximation doivent

être mises en place. Parmi elles, les *méthodes de Monte-Carlo séquentielles* jouent un rôle central (voir [CMR05] et les références citées). L'idée est de construire récursivement une approximation particulaire

$$\phi_t^N = \sum_{i=1}^{N} w_t^i \delta_{\xi_t^i}$$

de la loi de filtrage ϕ_t, où les $(w_t^i)_{1\leqslant i\leqslant N}$ désignent des poids, et où chaque particule $\xi_t^i \in \mathcal{X}$ est déduite de la génération antérieure $(\xi_{t-1}^j)_j$ par échantillonnage et, éventuellement, ré-échantillonnage.

Plaçons-nous pour simplifier dans un modèle complètement dominé, et notons $g_t(x)$ la vraisemblance de l'observation Y_t au point $x \in \mathcal{X}$, ainsi que $m(x,\cdot)$ la densité de la loi $M(x,\cdot)$, où M désigne le noyau de transition de la chaîne cachée $X_{0:n}$, et, conformément à l'usage, notons la densité des lois de filtrage/lissage de la même façon que ces lois elles-mêmes. Une première solution pour l'échantillonnage séquentiel consiste à donner à chaque particule ξ_{t-1}^i de la génération antérieure un descendant, tiré sous la densité de proposition $P_t(\xi_{t-1}^j,\cdot)$, et choisir pour la particule ξ_t^i ainsi créée un poids proportionnel à $w_{t-1}^i m(\xi_{t-1}^i,\xi_t^i) g_t(\xi_t^i)/P_t(\xi_{t-1}^i,\xi_t^i)$. Bien qu'il fonctionne correctement pour des séquences courtes, ce schéma ne s'avère toutefois pas assez stable pour traiter des données longues. En particulier, il apparaît un phénomène de *dégénérescence* qui fait que tout le poids de l'approximation particulaire se concentre au fur et à mesure sur un nombre de plus en plus restreint de particules (voir [CMR05], chapitre 7). Une solution, proposée par [GSS93], consiste alors à procéder régulièrement à une phase de *ré-échantillonnage* : après la phase d'échantillonage, on procède (au moins de temps en temps) à des tirages aléatoires parmi les particules qui visent à faire disparaître celles dont le poids est le plus faible, et à dupliquer les autres, afin d'égaliser les poids. Ce principe est utilisé pour le problème de l'idenfication aveugle des paramètres d'un canal à modulation linéaire présenté ci-dessous.

3.2 Forward Filtering, Backward Smoothing

Dans l'article [DGMO11], Randal Douc, Eric Moulines et moi avons considéré les schémas assez généraux dits de *filtre auxilaire* [PS99], où les particules pondérées (w_t^i,ξ_t^i) sont déduites de la génération précédente de la façon suivante : on commence par tirer, avec une probabilité $\vartheta_t(\xi_{t-1}^j)$, un ancêtre ξ_t^j pour notre nouvelle particule (la fonction ϑ_t est dite "poids multiplicateur d'ajustement"), puis l'on tire ξ_t^i selon la loi $P_t(\xi_{t-1}^j,\cdot)$, où P_t désigne le noyau de proposition. Le cas le plus simple est le *filtre auxiliaire bootstrap* : la fonction ϑ est prise constante, tandis que le noyau de proposition est constamment choisi égal à m, celui de la chaîne cachée. L'inconvénient majeur de ce choix est qu'il ne prend pas

en compte la nouvelle observation Y_t : une alternative consiste donc à choisir le filtre auxiliaire dit *complètement adapté*, pour lequel $\vartheta(x) = \int m(x, x')g_t(x')dx'$ et $P_t(x, x') = m(x, x')g_t(x')/\vartheta(x)$, qui a, lui, l'inconvénient d'être numériquement beaucoup plus coûteux.

A partir des approximations des lois de filtrage peuvent être déduites des approximations particulaires des lois de lissage, de plusieurs façons (voir [CMR05], chapitres 7-9). Deux algorithmes ont focalisé notre attention : ils reposent sur le constat que, conditionnellement aux observations, la suite $X_{0:n}$ constitue une chaîne de Markov inhomogène non seulement dans le sens du temps, mais également dans le sens inverse, avec un noyau de transition de densité b_{ϕ_t} où, pour toute densité de probabilité η le *noyau backward* b_η est défini par la relation :

$$b_\eta(x, x') = \frac{\eta(x')m(x', x)}{\int \eta(x')m(x', x)dx'} .$$

On en déduit des relations de récurrence permettant de calculer la loi de lissage jointe et ses marginales. Dès lors, il suffit d'utiliser l'approximation particulaire ϕ_t^N à la place du filtre ϕ_t pour obtenir de façon récursive une approximation de la loi de lissage : c'est l'idée qui dirige l'algorithme appelé *Forward Filtering Backward Smoothing* (FFBS).

Mais quand on cherche à approcher la distribution a posteriori $\phi_{s:t|n}$ d'une plage $(X_s, \ldots, X_t)$ un peu large des états cachés, la complexité algorithmique de l'algorithme FFBS devient trop grande. L'idée est alors d'utiliser une approximation supplémentaire : au lieu d'intégrer sous les (approximations des) noyaux backward, on peut tout simplement tirer des trajectoires utilisant ces noyaux. L'algorithme résultant, appelé *Forward Filtering Backward Simulation* (FFBSi), a donc une variance un peu plus grande que FFBS, qui en apparaît comme une version "Rao-Blackwellisée". En échange, sa complexité algorithmique est grandement réduite. Nous montrons en particulier qu'un schéma de tirage simultané intelligent des trajectoires permet de les réaliser en un temps d'espérance linéaire à la fois en le nombre de trajectoires souhaitées et en la longueur $t-s$ de la plage, sous des hypothèses de bornitude de la densité de transition (propositions 1 et 2).

La démonstration de garanties théoriques de la qualité d'estimation d'une loi a posteriori $\phi_{s:t|n}$ par son homologue particulaire $\phi_{s:t|n}^N$ est compliquée par le fait que les poids de lissage à chaque instant dépendent de *toutes* les particules pondérées avant et après cet instant. Le théorème 5 et le corollaire 6 contiennent toutefois des bornes de déviations fonctionnelles sous-gaussiennes : il est montré, sous de bonnes hypothèses, l'existence de deux constantes $B > 0$ et $C < \infty$ telles que pour toute fonction bornée f définie sur $\mathcal{X}^{n+1}$ on ait :

$$P\left[\left|\phi_{0:n|n}(h) - \phi_{0:n|n}^N(h)\right| \geqslant \epsilon\right] \leqslant B \exp\left(-\frac{CN\epsilon^2}{\mathrm{osc}^2(h)}\right) ,$$

où $\phi(h)$ désigne l'espérance de h sous la loi ϕ, et où $\mathrm{osc}(h)$ désigne l'oscillation de la fonction h. Ces résultats sont complétés dans le théorème 8 par un théorème de la limite centrale.

Le reste de l'article [DGMO11] est consacré à l'obtention de bornes uniformes en temps : alors que les bornes précédentes faisaient apparaître des termes de variance croissants avec la taille des observations, on montre que l'erreur d'approximation finit par se stabiliser quand la chaîne possède certaines propriétés d'oubli. La démonstration de ces résultats est rendue possible par des hypothèses de mélange fort du noyau M, qui entraînent des propriétés analogues pour les noyaux conditionnels aux observations à la fois dans le sens du temps et dans le sens inverse. Là aussi sont montrés une inégalité exponentielle (théorème 11) et un théorème de la limite centrale (théorème 12) avec une variance asymptotique qui reste bornée quand n tend vers l'infini.

3.3 Déconvolution aveugle et quasi-maximum de vraisemblance

Le cadre de l'estimation aveugle des paramètres d'un canal dans un schéma de modulation linéaire (voir [SJ00, TV05], et les références citées), qui était l'objet d'étude de la thèse de Steffen Barembruch, fournit un exemple de situation où les méthodes particulaires montrent tout leur intérêt alors même que l'espace d'états est fini. Le modèle est le suivant : une suite $(X_k)_k$ de symboles est tirée uniformément et sans mémoire sur un alphabet $\mathcal{X} \subset \mathbb{C}$ fini. Ce message est observé à travers un schéma de modulation linéaire avec un bruit additif : si l'entier L désigne l'ordre du canal, et si $h = (h_0, \ldots, h_{L-1}) \in \mathbb{C}^L$ désigne ses coefficients, la k-ième observation est :

$$Y_k = \sum_{l=0}^{L-1} A_{k-l} h_l + \epsilon_k \ ,$$

où les variables ϵ_k sont indépendantes et de loi $\mathcal{N}_{\mathbb{C}}(0, \sigma^2)$. Ce modèle s'écrit sous forme de chaîne de Markov cachée, en notant $X_k = (A_k, \ldots, A_{k-L+1})$ le k-ième état caché et $W_k = (A_k, 0, \ldots, 0)$. On a alors :

$$\begin{aligned} X_k &= QX_{k-1} + W_k \ \text{ et} \\ Y_k &= h' X_k + \epsilon_k \ , \end{aligned}$$

où Q désigne la matrice de décalage dont les coefficients non nuls sont sur la sous-diagonale. Les paramètres inconnus sont les coefficients h du canal (paramètre d'intérêt) et le niveau de bruit σ (paramètre de nuisance). On cherche à

les estimer en utilisant la méthode dite Expectation-Maximization (EM), mais l'espace d'état $\mathcal{X}^L$ est en pratique trop grand pour que l'on puisse utiliser l'algorithme de Baum-Welch : $|\mathcal{X}|$ est typiquement de l'ordre de quelques dizaines (ou centaines), alors que L peut valoir une dizaine. Par contre, la chaîne de Markov est ici tellement structurée que l'approche particulaire se prête à un traitement particulier très efficace pour les phases de filtrage et de lissage des itérations EM : c'est le résultat auquel Steffen Barembruch, Eric Moulines et moi avons contribué (voir [BGM09], précédé des articles de conférence [BGM08a, BGM08b]). En effet, une population de N particules pondérées (ξ_t^i, w_t^i) peut donner naissance au maximum à Nm nouvelles particules distinctes, ce qui fait que l'on peut toutes les envisager. Pour éviter une explosion combinatoire, il suffit, parmi les $M \leqslant Nm$ "enfants" possibles, d'en choisir un sous-ensemble de taille N sans trop dégrader la qualité de l'approximation. Une fois ce problème de sélection résolu, il apparaît qu'en s'y prenant attentivement, le terme dominant dans la complexité du filtrage/lissage particulaire est seulement d'ordre $O(mn)$. Nous avons envisagé plusieurs méthodes pour ce filtrage/lissage, en particulier "fixed-lag" (directement issue de la première méthode de ré-échantillonnage présentée ci-dessus) ainsi que "two-filter" ou plutôt sa variante "joined two-filter" plus spécifique au problème. L'article présente en détail les algorithmes originaux permettant d'atteindre cette faible complexité.

On se voit donc ramené à devoir construire un schéma de sélection, potentiellement randomisé, qui remplace une loi $\phi = \sum_{i=1}^M w_i \delta_{\xi_i}$ par une loi $\hat{\phi} = \sum_{i=1}^M W_i \delta_{\xi_i}$ la plus proche possible, avec la contrainte que le nombre de coefficients W_i non nuls soit égal (presque sûrement, ou au moins en espérance) à $N \leqslant M$. Le résultat dépend bien sûr de la façon dont on mesure la (pseudo-) distance de d entre ϕ et $\hat{\phi}$. Le théorème 1 de [BGM08a] explicite, en fonction d (qui doit satisfaire quelques hypothèses assez générales), la solution de ce problème. On peut distinguer trois cas : une métrique de type L^1 conduit à ne conserver de façon déterministe que les particules qui ont les poids les plus importants ; une métrique de type L^2 ou Kullback-Leibler conserve toutes les particules dont le poids dépasse un certain seuil, mais également quelques autres avec une probabilité proportionnelle à leur poids ; enfin, une pseudo-distance de type chi-deux conserve les particules les plus faibles avec une probabilité proportionnelle à la racine carrée de leurs poids, favorisant ainsi les petites en comparaison des autres métriques.

Nous avons testé les différents schémas de sélection, et les différentes méthodes de filtrage/lissage, sur des essais simulés faisant intervenir des valeurs de m et de L suffisamment petites pour qu'une comparaison avec la loi a posteriori exacte soit possible : il apparaît qu'aucune méthode n'est supérieure aux autres pour tous les niveaux de bruit à la fois (voir la figure 3.1), mais qu'en revanche les schémas de sélection randomisés améliorent clairement la convergence.

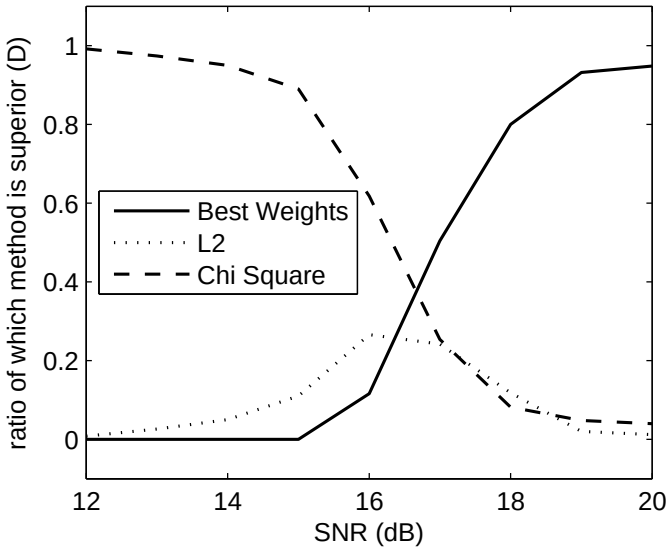


FIGURE 3.1 – Pour un canal $h = (0.63, 0.05, -0.3)$, comparaison de différents schémas de ré-échantillonnage pour l'approximation de la loi de lissage, en fonction du niveau de bruit.

Chapitre 4

Chaînes de Markov à mémoire variable

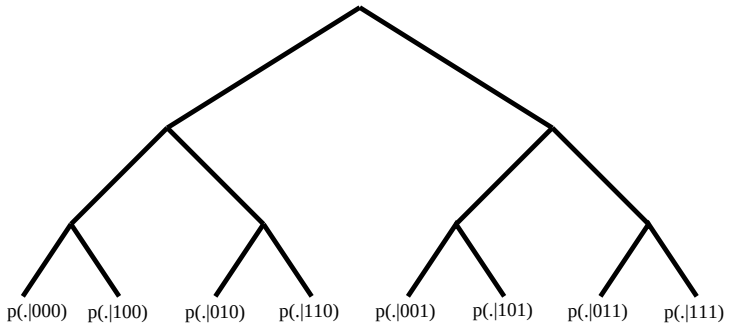
Les sources à arbres de contextes sont des modèles probabilistes qui ont été introduits par Jorma Rissanen pour généraliser les chaînes de Markov : elles ont depuis été beaucoup utilisées autant en probabilités appliquées qu'en statistiques, en théorie de l'information, en linguistique ou en bio-informatique. Ce succès s'explique par le compromis remarquable qu'elles réalisent entre puissance expressive et simplicité d'utilisation : guère plus difficiles à manipuler que les chaînes de Markov classiques, elles se révèlent infiniment plus flexibles et parcimonieuses en n'offrant de la mémoire que là ou cela est nécessaire, et en proposant un grand nombre de modèles par dimension.

Dans ces modèles, parfois appelés "chaînes de Markov d'ordre variable", la loi d'un symbole conditionnellement à son passé a la particularité de dépendre d'un certain nombre (fini ou infini) de ses prédécesseurs, ce nombre pouvant varier suivant la valeur prise par ce passé. J'avais commencé à travailler sur ces modèles pendant ma thèse : j'avais montré [Gar06c] qu'ils pouvaient être efficacement utilisés pour le codage des sources à longue mémoire par une méthode de double mélange appelée Context Tree Weighting [WST95], à condition de ne pas borner (comme il était alors d'usage) la profondeur des arbres utilisés, et j'avais étudié [Gar06a] la consistance d'un estimateur de type pseudo-maximum de vraisemblance pénalisé. J'ai continué à travailler sur ces modèles dans le cadre du projet de collaboration franco-brésilien COFECUB, avec des partenaires de l'université de Sao Paulo, sous un angle plus probabiliste et statisticien.

4.1 Simulation exacte

On appelle *noyau* sur un alphabet (considéré ici fini) A une application P définie sur l'ensemble $A^{-\mathbb{N}^*}$ des suites sur A indexées par des entiers négatifs

FIGURE 4.1 – Arbres de contexte correspondant à une chaîne de Markov d'ordre 3 sur l'alphabet binaire $A = \{0, 1\}$.



à valeur dans l'ensemble $M_1(A)$ des mesures de probabilités sur A, qu'il faut penser comme une famille de probabilités conditionnelles à un passé semi-infini. La terminologie d'*arbre* provient du fait que l'ensemble $A^{-\mathbb{N}^*}$ peut être représenté par un arbre complet infini dont les arêtes sont indexées par A, et dans lequel chaque suite $w \in A^{-\mathbb{N}^*}$ correspond à un chemin partant de la racine que l'on parcourt en suivant les arêtes indexées par $w_{-1}, w_{-2}, \ldots$, et auquel serait associée la loi de probabilité $P(w)$; la figure 4.1 montre un tel arbre sur l'alphabet binaire $A = \{0, 1\}$ tronqué à la profondeur 3.

D'un point de vue probabiliste, le premier problème concernant les sources à arbres de contextes est celui de leur existence : étant donné un noyau P, existe-t-il un processus stationnaire $(X_t)_{t\in\mathbb{Z}}$ dont P définisse des versions régulières des lois conditionnelles ? D'autre part, d'un point de vue pratique, la question qui se pose est de savoir comment simuler des trajectoires finies d'un tel processus. Ces deux questions sont en fait étroitement liées, et ont été considérées depuis lontemps ; les processus spécifiés par un noyau associant à toute suite infinie de symboles passés la loi de probabilité du symbole courant ont été appelées "chaînes à connections complètes" [OM35a, OM35b] ou "chaînes d'ordre infini"[Har55]. Dans un article plus récent [CFF02], Comets, Fernandez et Ferrari montrent, après [Lal86, Lal00, Ber87], comment exploiter la structure régénérative de tels processus. Sous de bonnes condition de non-nullité ainsi que sous des hypothèse de décroissance sommable de coefficients d'oubli, ils exhibent un algorithme de simulation parfaite par le passé qui converge en temps fini, réglant par là-même la question de l'existence (et, sous des conditions plus fortes, de l'unicité) du processus correspondant.

Cet algorithme de simulation par le passé a depuis suscité beaucoup d'intérêt : plusieurs auteurs ont notamment essayé de généraliser ses résultats à des noyaux satisfaisant des conditions plus faibles (voir [FTC98, Gal10, DSP10], et les références citées). Gallo s'est attaché à montrer que les conditions de continuité de P (en tant qu'application de l'espace $A^{-\mathbb{N}^*}$ muni d'une topologie engendrée par

les cylindres vers l'espace $M_1(A)$ muni de la distance en variation totale) utilisées dans [CFF02] ne sont pas indispensables. D'autre part, le cas des chaînes de Markov d'ordre fini montre bien que la condition de non-nullité n'est pas nécessaire : Propp et Wilson, en particulier, ont proposé un algorithme de "couplage par le passé" [PW96] particulièrement élégant et rapidement célèbre, qui ne fait pas usage d'une telle propriété : c'est un temps de coalescence, étroitement lié au temps de mélange de la chaîne, qui contrôle la vitesse de convergence de l'algorithme.

Mon objectif a été de comprendre comment généraliser la méthode de Propp et Wilson aux chaînes d'ordre infini, en utilisant le moins possible de conditions sur le noyau, et tout en conservant un algorithme utilisable de complexité limitée. Il s'agissait de concevoir un algorithme qui, pour un noyau générique (infini), converge au moins aussi vite que celui de [CFF02], mais qui pour une chaîne de Markov (d'odre 1) se comporte exactement comme l'algorithme de Propp et Wilson. J'ai appris a posteriori que De Santis et Piccioni [DSP10] ont envisagé un objectif un peu semblable, en élaborant un algorithme "hybride" fonctionnant avec un régime de coalescence à court terme, et un régime de régénération à long terme. Dans l'article [Gar11], je propose plutôt de remplir le fossé entre le long et le court terme, en décrivant un algorithme un peu plus complexe mais proposant un traitement unifié de tous les régimes. Cet algorithme de couplage par le passé repose sur l'idée que la valeur X_t du processus au temps t peut être déterminée par la donnée d'une variable aléatoire U_t uniforme sur l'intervalle $[0, 1]$, ainsi que par la valeur d'un ensemble fini (dépendant de U_t) de valeurs du processus en des temps antérieurs, qui forme un *arbre* que l'on désigne sous le nom anglais de *trie* pour insister sur sa structure informatique, ou bien de *dictionnaire complet de suffixes* pour souligner sa fonction. L'algorithme proposé repose sur la simulation d'une *chaîne de Markov sur l'ensemble des tries* dont les feuilles sont étiquetées, qui s'achève quand un trie vide est atteint. Il serait trop long de décrire ici complètement les transitions de cette chaîne de Markov, dont le mécanisme de complétion/élagage est esquissé à la figure 4.2 pour éveiller la curiosité du lecteur. Précisons seulement que la loi de transition de cette chaîne de Markov est construite à partir d'un couplage original décrit dans la section 4 de l'article. L'algorithme lui-même est décrit dans la section 5.

Dans le cas des chaînes de Markov d'ordre 1, tous les tries qui apparaissent sont de profondeur 1, et seule importe donc l'étiquette des feuilles : on retombe exactement sur l'algorithme de Propp et Wilson. A l'autre extrême, l'algorithme de [CFF02] peut être interprété comme une procédure ne prenant en compte que la profondeur de pareils tries (et converge donc trajectoriellement moins vite, quand il converge : voir la section 6.1). D'un point de vue informatique, cet algorithme se montre intéressant même dans le cas de processus à mémoire finie d'odre d : il surpasse alors en termes de complexité de calcul la méthode de

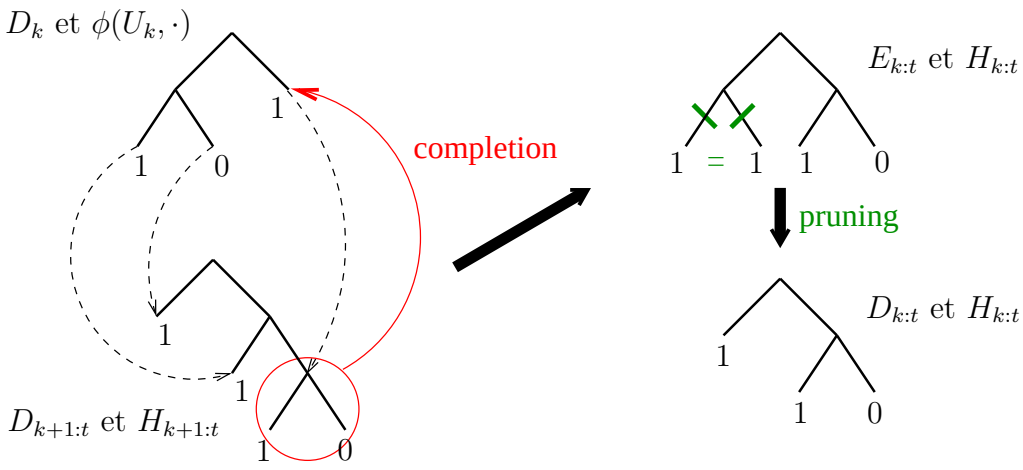


FIGURE 4.2 – Une itération de la chaîne de Markov sur les tries étiquetés utilisée par l'algorithme de simulation par le passé.

simulation de Propp et Wilson appliqué à la chaîne étendue $(X_{t-1}, \dots, X_{t-d})_t$ (voir la proposition 6). Sur un plan plus théorique, l'existence d'un tel algorithme illustre également le fait que les conditions de régularités et même de continuité ne sont pas nécessaires à l'existence d'un processus stationnaire compatible, ni même à la simulation parfaite : un exemple est fourni dans la section 7.

4.2 Estimation non asymptotique d'un modèle de mémoire

On dit qu'un noyau P sur un alphabet fini A admet un *contexte* $s \in A^k$ si

$$\forall (w, z) \in A^{-\mathbb{N}}, w_{-k:-1} = z_{-k:-1} = s \implies P(w) = P(z) \ ,$$

et si aucun suffixe de s n'a cette propriété. La connaissance d'un tel contexte est alors un élément crucial pour l'étude de P, puisqu'il détermine la valeur d'un passé fini au delà duquel il n'est pas nécessaire de revenir pour déterminer la loi d'un symbole. Dans l'arbre décrit plus haut, cela signifie que tous les chemins commençant par les symboles $s_{-1}, s_{-2}, \dots, s_{-k}$ peuvent être "fusionnés" en un seul, qui est alors étiqueté par la valeur commune prise sur eux par le noyau.

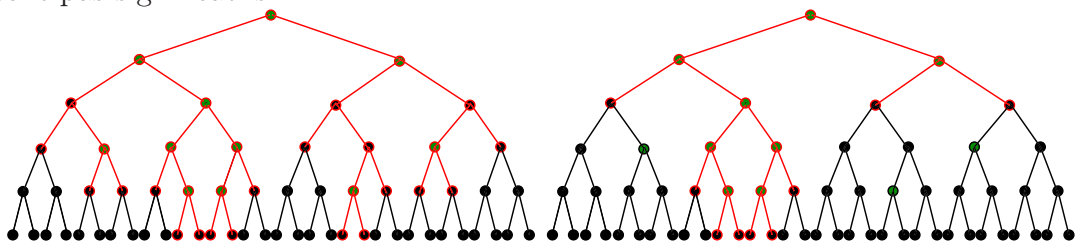
Pour l'utilisation pratique des sources à arbres de contexte, il est souvent nécessaire d'estimer les contextes de P à partir d'une séquence d'observation X_1^n distribuées suivant un processus $\mathbb{P}$ dont P définit une version des probabilités conditionnelles (par abus de langage, on dira aussi que s est un contexte de $\mathbb{P}$).

Un exemple, développé dans la section suivante, est donné par la linguistique, où (au delà de la valeur des transitions estimées) la structure même des arbres de mémoire apparaît comme un marqueur caractéristique d'une langue ou d'un dialecte (voir [GGG+10] et les références citées). Ce travail d'estimation a d'abord été entrepris dans le cas où le noyau P admet un dictionnaire complet de suffixes D *fini* : tout élément de $A^{-\mathbb{N}}$ possède alors un unique suffixe dans D. Dans ce cas, on peut fusionner les chemins partageant la même valeur, aboutissant ainsi à un arbre fini dont les feuilles correspondent exactement aux éléments D. Chaque arbre fini fournit un modèle, qui rassemble toutes les sources à arbres de contexte dont il forme l'ensemble des contextes. Dans le cas fini, la notion de consistance est claire : une suite d'estimateurs $\left(\hat{T}_n(X_{1:n})\right)_n$ est fortemement consistante si elle converge presque-sûrement vers le bon arbre quand n tend vers l'infini. Mais plus récemment, le cas de l'estimation avec des arbres infinis a également été considéré : l'objectif est alors d'estimer correctement les contextes finis de la source, et l'on a proposé d'appeler "fortement consistante" une suite d'estimateurs $\left(\hat{T}_n(X_{1:n})\right)_n$ telle que pour tout entier K, l'ensemble des éléments de $\hat{T}_n(X_{1:n})$ de taille inférieure à K converge presque-sûrement vers l'ensemble des contextes de la source de tailles inférieures à K.

On peut distinguer deux grandes familles d'estimateurs. La première, initiée par Rissanen [Ris83], se rattache à l'algorithme appelé *Context* qui est basé sur des règles d'élagage rappelant le principe de l'algorithme CART : en chaque noeud de l'arbre, on calcule une mesure de discrépance entre ses enfants qui déterminent si ceux-ci doivent ou non être conservés. Sous de bonnes hypothèses, la consistance de ces méthodes a été prouvée (voir par exemple [Ris83, BW99]). La seconde famille d'estimateurs est basée sur l'approche classique du maximum de vraisemblance pénalisé (PML) : pour chaque modèle (c'est-à-dire pour chaque dictionnaire complet de suffixes, ou encore pour chaque arbre fini), on calcule un (pseudo-)maximum de vraisemblance des observations qui mesure l'adéquation du modèle aux données, que l'on pénalise par une mesure de la complexité du modèle (typiquement proportionnelle au nombre de contextes qu'il contient). En théorie de l'information, de telles procédures s'interprètent comme des mises en oeuvre du principe de *Minimum Description Length* [BRY98]. Csiszar et Talata [CT06] ont montré que pour un choix pertinent de pénalités ces estimateurs sont fortement consistants si l'on se restreint aux modèles de petite taille ; dans le cas fini, j'ai levé cette restriction dans l'article [Gar06a].

Plus récemment est apparu le souci d'obtenir des bornes *non asymptotiques* pour la probabilité d'estimer correctement les contextes. Des résultats ont été obtenus pour l'algorithme Context [GMDS08, GL08] et pour les algorithmes PML [Leo10], faisant apparaître des bornes qui dépendent de propriétés inconnues du processus ; ces bornes utilisent des résultats récents de dépendance faible,

FIGURE 4.3 – Comparaison entre les estimateurs Context (à gauche) et PML (à droite) ; dans chaque algorithme, une mesure de discrépance dans chaque noeud est jugée significative (vert) ou non (noir). Les arbres estimés sont représentés en rouge : Context conserve tous les ancêtres des noeuds significatifs (ainsi que leurs enfants, et de manière à former un arbre complet), alors que PML ne conserve que les contextes reliés à la racines par des noeuds qui sont tous significatifs. La mesure de discrépance est locale pour Context, alors qu'elle est calculée de façon récursive pour PML, mais elles coïncident pour les noeuds dont les enfants ne sont pas significatifs.



et s'appuient fortement sur des hypothèses de mélange du processus.

Pour améliorer ces résultats, il convient en fait d'insister sur la distinction entre deux types possibles d'erreur d'estimation : la sous-estimation et la sur-estimation. L'arbre T_0 des contextes de la source est dit *sous-estimé* si l'un de ses éléments a un suffixe strict dans l'estimateur $\hat{T}_n$, et il est dit *sur-estimé*[1] quand un de ses contextes apparaît comme noeud interne de $\hat{T}_n$. Quand la taille de l'échantillon augmente, nos estimateurs évitent la sur-estimation et la sous-estimation pour des raisons différentes : tandis que la sous-estimation est évitée du fait de l'existence d'une distance strictement positive entre le vrai processus et tous ceux qui appartiennent à des modèles strictement plus petits, on échappe à la sur-estimation en contrôlant les fluctuations des lois de transitions empiriques.

Dans l'article [GL11], que j'ai écrit avec Florencia Leonardi, nous présentons une analyse unifiée des deux familles d'estimateurs, en montrant comment elles sont reliées (pour des paramètres comparables, l'estimateur PML est un sous-arbre de l'estimateur Context : voir la figure 4.2, ainsi que la proposition 2.13 de l'article). Sans hypothèse sur l'arbre (potentiellememnt infini) T_0, nous expliquons comment régler les deux procédures pour qu'elles fournissent avec grande probabilité des sous-arbres de T_0 : nous prouvons des bornes non asymptotiques plus précises que celles connues jusqu'alors pour la sur-estimation (théorème 2.14), qui de plus ne nécessitent pas les hypothèses faites dans [GMDS08, GL08, Leo10]. Ces résultats, une fois de plus et contrairement aux approches non-asymptotiques précédentes, s'appuient sur des inégalités de déviations auto-normalisées évoquées en introduction taillées sur mesure pour un schéma de preuve original reprenant

1. T_0 peut être à la fois sur-estimé et sous-estimé, pour des contextes différents.

l'idée de [Csi02] : il est préférable, pour estimer la probabilité conditionnelle $p(b|s)$ du symbole $b \in A$ conditionnellement à un passé $s \in \cup_k A^k$ suffisamment long par le rapport $N_n(sb)/N_n(s)$ du nombre d'occurrences de s suivies de b avec le nombre total d'occurrences de s dans l'échantillon, d'utiliser une approche martingale plutôt que d'étudier séparément la concentration de $N_n(sb)$ et celle de $N_n(s)$. D'autre part, nous exhibons des bornes pour la sous-estimation (théorème 2.18) sous des conditions classiques de mélange qui assurent qu'avec grande probabilité, tous les suffixes des éléments de T_0 de petite profondeur sont inclus dans $\hat{T}_n$. Ces résultats impliquent conjointement la consistance forte des algorithmes PML et Context (théorème 2.22) pour une plage de paramètres plus large que dans les articles précédents [Leo10, DGG06].

4.3 Estimation jointe de deux sources partiellements partagées

Ces procédures d'estimation de modèle de mémoire ont été utilisées notamment en linguistique, afin de caractériser certains motifs rythmiques des langues naturelles : voir par exemple [GGG+10] (et les références citées) pour une étude portant sur les variantes européenne et brésilienne du portugais. Dans cette application, toutefois, c'est plutôt pour la *comparaison* entre deux langues ou dialectes que les modèles à mémoire variable peuvent être pertinents : on cherche à mettre en évidence des caractéristiques partagées, et à faire ressortir des différences. L'approche naïve consistant à estimer séparément deux modèles, puis à comparer les estimateurs, peut toujours être utilisée ; mais quand Antonio Galves nous a présenté ce problème, il nous a semblé, avec Elisabeth Gassiat, qu'un estimateur construit précisément pour cet objectif comparatif pourrait s'avérer plus puissant.

Dans l'article [GGG11], et dans le cadre du projet franco-brésilien COFECUB réunissant l'Université de Sao Paulo et Telecom Paristech, nous avons donc tous les trois considéré le modèle suivant : $(X_n)_n$ et $(Y_n)_n$ sont deux sources à arbres de contexte fini stationnaires, dont les noyaux P_X et P_Y sont partiellement identiques : il existe un ensemble σ_0 de contextes de P_X et de P_Y sur lesquels P_X et P_Y coïncident. On observe $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$. Le but est de construire un estimateur de σ_0, P_X et P_Y conjointement. Nous avons étudié un estimateur du maximum de vraisemblance pénalisé pour ce modèle. Sous des conditions usuelles de pénalisation, nous avons montré sa consistance forte (théorème 1). Il pourrait tout d'abord sembler qu'un tel estimateur soit surtout un objet théorique, compte tenu du nombre expontentiel de modèles en compétition, mais cela n'est pas le cas : il peut être calculé en un temps dépendant linéairement du nombre d'observations. Nous avons en effet montré dans la section 4 comment adapter le principe dit de "Context Tree Maximization" [WST95, CT06] pour obtenir un

algorithme qui, partant d'un arbre complet d'une certaine profondeur, l'élague progressivement en remontant des feuilles vers la racine. Conformément à nos attentes, nous avons montré sur des simulations numériques (présentées dans la section 5) combien cet algorithme peut être avantageux par rapport à une procédure d'estimation séparée quand des contextes sont effectivement partagés par les deux sources. Dans certains cas un peu déséquilibrés (où, par exemple, les deux échantillons ont des tailles très différentes), il apparaît un phénomène de compensation : l'estimation d'une des deux sources est légèrement dégradée au profit d'une amélioration de l'estimation de l'autre source.

Chapitre 5

Perspectives

Pour chacun des thèmes évoqués dans ce mémoire, les projets à accomplir ne manquent pas. La bonne compréhension que j'ai acquise des modèles élémentaires d'apprentissage par renforcement me permet d'envisager l'étude de problèmes plus ambitieux tout en conservant un haut niveau d'exigence quant à la précision des résultats obtenus (qui fait un peu défaut, pour l'instant, à l'état de l'art). Et l'élargissement de mes connaissances statistiques et probabilistes m'encourage à explorer de nouvelles pistes dans le champ des processus markoviens.

Arbres de Contextes Probabilisés

Une collaboration entamée avec Matthieu Lerasle a relancé un projet que j'ai depuis plusieurs années : importer en théorie de l'information l'approche de la sélection de modèles qu'ont popularisée Birgé et Massart [Mas03] ou Lepski [Lep91]. Ce projet est d'autant plus délicat que (contrairement au cadre classique de la sélection de modèles) les données sont ici très fortement dépendantes, et des difficultés techniques importantes en résultent ; d'autre part, des résultats classiques en théorie de l'information (et en particulier la possibilité d'agréger très efficacement les estimateurs des différents modèles) s'avèrent déjà très proches d'être optimaux (voir par exemple [WST95]), ce qui oblige une approche alternative à être particulièrement soignée pour justifier son intérêt. Nous plaçons pour cela beaucoup d'espoirs dans l'*heuristique de pente* [BM07], qui permet de choisir automatiquement la meilleure constante de pénalisation en fonction des données.

Par ailleurs, une piste différente et originale pour la sélection de modèles de mémoire est l'utilisation de pénalisations de type L^1. Si, du fait de la structure arborescente, une telle approche ne s'impose pas pour des raisons combinatoires (la pénalisation par la dimension est rendue possible par l'existence d'un algorithme glouton), on peut penser qu'elle présente des caractéristiques propres qui lui donnent un intérêt, au-delà de la simple application d'une idée très en vogue

depuis quelques années à un nouveau domaine. Je pense en effet qu'il est possible de faire porter la pénalisation sur l'opposé de constantes de couplage entre lois conditionnelles, ce qui aurait l'avantage de pouvoir se combiner avec les propriétés de continuité des processus correspondants.

Problèmes de bandits et MDP

Nous avons montré que des progrès très importants étaient possibles, même pour un problème aussi étudié que les bandits multibras ou encore pour l'estimation d'arbres de contextes, en utilisant la bonne notion de pseudo-distance : en l'occurrence, les inégalités de déviation auto-normalisées pour les divergences de Kullback-Leibler forment un outil dont toutes les possibilités n'ont pas encore été exploitées. En particulier, nous aimerions beaucoup montrer, au-delà de leur apport pratique qui a été mis en évidence par les simulations, que l'algorithme KL-UCRL permet d'obtenir des bornes de regret significativement meilleures que ses concurrents pour les processus de décision markoviens. En outre, pour les problèmes de bandits, nous souhaiterions étendre nos résultats à des classes de processus plus générales (que les familles exponentielles canoniques), potentiellement non paramétriques. Des travaux sur en ce sens sont en cours, en collaboration avec Gilles Stoltz (CNRS, ENS Ulm) et Odalric Ambrym-Maillard (INRIA Sequel, en contrat post-doctoral à l'université de Leoben).

Une piste, pour aller encore plus loin, consiste à utiliser un principe d'optimisme plus faible, non plus systématique mais bayésien. Pour les problèmes de bandits, cela revient à mettre un a priori sur toutes les lois de récompenses possibles pour les bras, puis à calculer séquentiellement pour chaque bras non plus une borne supérieure de confiance, mais plutôt un quantile bien choisi de la loi a posteriori. Des travaux en ce sens sont en cours avec Emilie Kaufmann, qui vient de débuter sa thèse à Telecom Paristech sous la co-direction d'Olivier Cappé, Rémi Munos et moi, et qui a achevé chez nous un stage de Master 2 très prometteur. Notre espoir, soutenu par de premières preuves, est que ces méthodes bayésiennes seront aussi puissantes, mais plus générales et surtout "automatiquement adaptées" pour résoudre les problèmes de bandits, tandis qu'elles seront strictement plus efficaces dans le cas plus riche des processus de décision markoviens. Une autre approche bayésienne, radicalement différente, consiste à considérer la solution exacte du problème de bandits bayésien dans un cadre paramétrique simple, où une loi a priori conjuguée est utilisée sur chacun des bras. Les travaux de Gittins [Git79] réduisent la tâche à un problème de planification dans un MDP à espace d'états infini, que l'on peut résoudre numériquement - il s'agit en fait de calculer l'enveloppe de Snell d'un certain processus. Notre objectif est d'obtenir des approximations de la solution qui la rende efficace en pratique, ainsi que des garanties théoriques fréquentistes pour le regret de cette méthode.

En collaboration avec Rémi Munos, nous espérons pouvoir obtenir ainsi des méthodes et des analyses qui se généralisent aisément à des problèmes de bandits plus complexes. Notre objectif est ensuite de considérer le cas où la loi des récompenses dépend d'un ensemble de covariables observées (problème des bandits contextuels), et tout particulièrement si seulement quelques-unes des covariables sont réellement utiles : se pose alors la question de savoir détecter lesquelles ; il faut élaborer une procédure de choix de modèle séquentielle, dans un cadre où les observations ne sont pas indépendantes ni identiquement distribuées mais déterminées, justement, par la procédure elle-même. Il s'agit dans un premier temps de voir dans quelle mesure les méthodes classiques (pénalisation, notamment LASSO, ou agrégation, approche bayésienne ou PAC-bayésienne) peuvent être adaptées ici, avant de peut-être devoir élaborer des solutions plus spécifiques.

Dérivations et extensions

Le problème des bandits multibras n'est pas seulement intéressant pour lui-même et pour les situations qu'il modélise : il fournit aussi une heuristique intéressante dans des situations apparemment différentes, mais auxquelles le paradigme optimiste peut être appliqué avec succès. En collaboration avec d'autres spécialistes de ces sujets, j'aimerais explorer en particulier quelques-unes de ces situations. En premier lieu, je travaille avec Gilles Stoltz, et avec Céline Lévy-Leduc dans le cadre de la thèse CIFRE financée par Orange de Marjorie Jala, portant sur l'étude de l'exposition au rayonnement électro-magnétique des téléphones mobiles, sur l'optimisation de fonctions bruitées en grande dimension. Le problème est le suivant : on cherche à trouver le maximum d'une fonction f, définie sur un sous-ensemble compact de R^p (avec p de l'ordre de quelques dizaines), en évaluant cette fonction en un nombre minimal de points (car chaque évaluation est numériquement extrêmement coûteuse) ; on peut choisir séquentiellement les points d'évaluation, en fonction des valeurs précédemment observées. L'idée est de reconnaître ici un problème de bandits, où les "bras" sont des points de $\mathbb{R}^p$ et les récompenses sont les valeurs observées de la fonction. Un algorithme de bandits efficace, en se concentrant rapidement sur le meilleur bras, identifiera correctement le maximum. Cette idée a été mise en oeuvre dans l'algorithme GP-UCB, présenté dans [SKSS10], de la façon suivante : en supposant que f est la réalisation d'un processus Gaussien, on peut calculer en chaque point $x \in \mathbb{R}^p$ la loi de $f(x)$ conditionnellement aux observations précédentes, ainsi que la variance conditionnelle ; muni de ces informations, on propose comme nouveau point celui qui maximise une borne supérieure de confiance. Des résultats de regret ont été obtenus, mais il paraissent largement perfectibles ; d'autre part, les hypothèses sur le noyaux gaussien qui sont faites dans l'article ne permettent de considérer que des fonctions très régulières. Enfin, nous souhaiterions étendre cette idée pour

rechercher non pas seulement le maximum de la fonction, mais plutôt la valeur des certains quantiles.

L'analyse du problème d'exploration pour lequel est décrit dans la section 2.6 l'algorithme Good-UCB peut être prolongée. Il conviendra d'abord de considérer des lois de probabilités non uniformes pour chaque distribution de tirage : la principale difficulté réside dans l'étude de la performance de la stratégie oracle, puisque l'estimateur de Good-Turing est totalement indifférent aux lois de tirages. Mais surtout, une autre asymptotique macroscopique est digne d'intérêt, où la taille de l'ensemble de tirages $\mathcal{X}$ tend vers l'infini alors que le nombre d'éléments remarquables présents dans le support de chaque distribution reste constant. Il est facile de voir que le régime limite sous la stratégie oracle est alors Poissonien, mais l'étude de la stratégie Good-UCB est nettement plus compliquée ; il semble notamment essentiel, dans ce cas, d'utiliser les inégalités de Boucheron, Massart et Lugosi [BLM09] pour faire apparaître le bon terme de variance.

Nouvelles approches

Un grand nombre de problèmes restent ouverts pour les trois grandes familles de problèmes (bandits, MDP, POMDP) présentées ci-dessus ; en particulier, l'écart reste abyssal entre les recherches méthodologiques s'intéressant à des problèmes concrets et les méthodes théoriques proposées et étudiées en détail, pour lesquelles sont souvent montrées des propriétés d'"optimalité" loin d'être évidentes en pratique (voir en particulier les bandits linéaires et non-linéaires) ; cela s'explique notamment par la grande difficulté qu'il y a à marier des inégalités de concentrations très fines et un peu inhabituelles avec des idées et des algorithmes d'exploration/exploitation de haut niveau. Au delà de la poursuite des réflexions présentées ci-dessus, il me semble intéressant de nous impliquer dans une approche qui consiste à réinterpréter le travail de planification comme un problème d'inférence dans un modèle de chaîne de Markov cachée. Si quelques travaux existent en ce sens (voir [TS06, HDdFJ07] et les références qu'ils contiennent), l'expertise de notre laboratoire en matière de filtrage particulaire laisse à penser que nous pourrons apporter à la fois de nouvelles méthodologies et des élements de preuve de leur consistance dans le cas des espaces d'états et d'actions continus, ou finis mais trop grands pour être traités par des méthodes de filtrage exactes.

Enfin, il reste un projet à long terme auquel mes travaux antérieurs me poussent naturellement : introduire et exploiter les processus à arbre de contexte en apprentissage par renforcement. Ce dernier fait un usage intensif des chaînes de Markov d'ordre un, ce qui s'explique par leur parenté avec les systèmes dynamiques. Toutefois, il existe des cas où introduire une mémoire plus grande, en particulier le long de certaines trajectoires, est susceptible d'améliorer beaucoup la modélisation ; c'est notamment le cas des processus de décision markoviens

continus lorsqu'ils sont traités par discrétisation. Il s'agirait donc de voir dans quelle mesure les techniques et algorithmes évoqués plus haut pour les arbres de contexte (simulation, estimation, etc...) peuvent être utilisés pour planifier ou apprendre dans de tels "processus de décision markoviens à mémoire variable".

Bibliographie

[ACBF02] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2) :235–256, 2002.

[Agr95] 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.

[AJO09] P. Auer, T. Jaksch, and R. Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in Neural Information Processing Systems*, 21, 2009.

[AMS09] J-Y. Audibert, R. Munos, and Cs. Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19), 2009.

[AO07] P. Auer and R. Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. *Advances in Neural Information Processing Systems : Proceedings of the 2006 Conference*, page 49, 2007.

[AWYVM08] 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.

[BEG11] Sébastien Bubeck, Damien Ernst, and Aurélien Garivier. Good-ucb : an optimistic algorithm for discovering unseen data, 2011.

[Bel56] R. Bellman. A problem in the sequential design of experiments. *Sankhyā : The Indian Journal of Statistics*, pages 221–229, 1956.

[Ber87] Henry Berbee. Chains with infinite connections : uniqueness and Markov representation. *Probab. Theory Related Fields*, 76(2) :243–253, 1987.

[Ber95] D.P. Bertsekas. *Dynamic Programming and Optimal Control, Two Volume Set*. Athena Scientific, 1995.

[BGG09] S. Boucheron, A. Garivier, and E. Gassiat. Coding on countably infinite alphabets. *IEEE Transactions on Information Theory*, 55(1) :358–373, 2009.

[BGM08a] S. Barembruch, A. Garivier, and E. Moulines. On approximate maximum likelihood methods for blind identification : How to cope with the curse of dimensionality. In *IEEE International Workshop on Signal Processing Advances for Wireless Communications*, 2008.

[BGM08b] S. Barembruch, A. Garivier, and E. Moulines. On optimal sampling for particle filtering in digital communication. In *IEEE International Workshop on Signal Processing Advances for Wireless Communications*, 2008.

[BGM09] S. Barembruch, A. Garivier, and E. Moulines. On approximate maximum likelihood methods for blind identification : How to cope with the curse of dimensionality. *IEEE Trans. on Signal Processing*, 57 :4247–4260, july 2009.

[BGR02] Bernard Bercu, Elisabeth Gassiat, and Emmanuel Rio. Concentration inequalities, large and moderate deviations for self-normalized empirical processes. *Ann. Probab.*, 30(4) :1576–1604, 2002.

[BK97] A.N. Burnetas and M.N. Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, pages 222–255, 1997.

[BLM09] Stéphane Boucheron, Gábor Lugosi, and Pacal Massart. On concentration of self-bounding functions. *Electron. J. Probab.*, 14 :no. 64, 1884–1899, 2009.

[BM07] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2) :33–73, 2007.

[BRY98] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44(6) :2743–2760, 1998. Information theory : 1948–1998.

[BT96] D. Bertsekas and J. Tsitsiklis. *Neuro-dynamic programming.* Athena Scientific, 1996.

[BT08] Bernard Bercu and Abderrahmen Touati. Exponential inequalities for self-normalized martingales with applications. *Ann. Appl. Probab.*, 18(5) :1848–1869, 2008.

[BT09] P.L. Bartlett and A. Tewari. REGAL : A Regularization based Algorithm for Reinforcement Learning in Weakly Communicating MDPs. *Annual Conference on Uncertainty in Artificial Intelligence*, 2009.

[BW99] P. Bühlmann and A. J. Wyner. Variable length Markov chains. *Ann. Statist.*, 27 :480–513, 1999.

[Cat10] O. Catoni. Challenging the empirical mean and empirical variance : a deviation study. arXiv :1009.2048, 2010.

[CBL06] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

[CFF02] F. Comets, R. Fernández, and P.A. Ferrari. Processes with long memory : regenerative construction and perfect simulation. *Ann. Appl. Probab.*, 12(3) :921–943, 2002.

[CGG09] A. Chambaz, A. Garivier, and E. Gassiat. A minimum description length approach to hidden Markov models with Poisson and Gaussian emissions. Application to order identification. *J. Statist. Plann. Inference*, 139(3) :962–977, 2009.

[CGT09] Stephan Clémençon, Aurélien Garivier, and Jessica Tressou. Pseudo regenerative block-bootstrap for Hidden Markov Models. In *IEEE Workshop on Statistical Signal Processing*, 2009.

[CMR05] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer-Verlag, 2005.

[Csi02] I. Csiszár. Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Trans. Inform. Theory*, 48 :1616–1628, 2002.

[CT06] I. Csiszár and Z. Talata. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inform. Theory*, 52(3) :1007–1016, 2006.

[DGG06] D. Duarte, A. Galves, and N.L. Garcia. Markov approximation and consistent estimation of unbounded probabilistic suffix trees. *Bull. Braz. Math. Soc.*, 37(4) :581–592, 2006.

[DGMO11] R. Douc, A. Garivier, E. Moulines, and J. Olsson. On the forward filtering backward smoothing particle approximations of the smoothing distribution in general state spaces models. *Annals of Applied Probability*, to appear 2011.

[DHK08] V. Dani, T.P. Hayes, and S.M. Kakade. Stochastic linear optimization under bandit feedback. *Conference on Learning Theory*, 2008.

[DLPKL04] V.H. De La Pena, M.J. Klass, and T.L. Lai. Self-normalized processes : exponential inequalities, moment bounds and iterated logarithm laws. *Annals of Probability*, 32(3) :1902–1933, 2004.

[DSP10] E. De Santis and M. Piccioni. A general framework for perfect simulation of long memory processes, April 2010.

[DZ10] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 2010. Corrected reprint of the second (1998) edition.

[EDMM02] E. Even-Dar, S. Mannor, and Y. Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *Conf. Comput. Learning Theory (Sydney, Australia, 2002)*, volume 2375 of *Lecture Notes in Comput. Sci.*, pages 255–270. Springer, Berlin, 2002.

[FCCM08] S. Filippi, O. Cappe, F. Clerot, and E. Moulines. A near optimal policy for channel allocation in cognitive radio. In *Lecture Notes in Computer Science*, Recent Advances in Reinforcement Learning, pages 69–81. Springer, 2008.

[FCG10] S. Filippi, O. Cappé, and A. Garivier. Optimism in reinforcement learning and Kullback-Leibler divergence. In *Allerton Conf. on Communication, Control, and Computing*, Monticello, US, 2010.

[FCG11] Sarah Filippi, Olivier Cappé, and Auélien Garivier. Optimally sensing a single channel without prior information : The tiling algorithm and regret bounds. *IEEE Journal of Selected Topics in Signal Processing*, 5(1) :68–76, Feb. 2011.

[FCGS10] S. Filippi, O. Cappé, A. Garivier, and C. Szepesvari. Parametric bandits : The generalized linear case. In *Neural Information Processing Systems (NIPS)*, 2010.

[FTC98] S. G. Foss, R. L. Tweedie, and J. N. Corcoran. Simulating the invariant measures of Markov chains using backward coupling at regeneration times. *Probab. Engrg. Inform. Sci.*, 12(3) :303–320, 1998.

[Fuh04] Cheng-Der Fuh. Asymptotic operating characteristics of an optimal change point detection in hidden Markov models. *Ann. Statist.*, 32(5) :2305–2339, 2004.

[Gal10] Sandro Gallo. Chains with unbounded variable length memory : perfect simulation and visible regeneration scheme, 2010.

[Gar06a] A. Garivier. Consistency of the unlimited BIC context tree estimator. *IEEE Trans. Inform. Theory*, 52(10) :4630–4635, 2006.

[Gar06b] A. Garivier. *Modèles contextuels et alphabets infinis en théorie de l'information*. PhD thesis, Université Paris Sud 11, Orsay, France, Nov. 2006.

[Gar06c] A. Garivier. Redundancy of the context-tree weighting method on renewal and Markov renewal processes. *IEEE Trans. Inform. Theory*, 52(12) :5579–5586, 2006.

[Gar09] A. Garivier. A lower-bound for the maximin redundancy in pattern coding. *Entropy*, 11(4) :634–642, 2009.

[Gar11] A. Garivier. A Propp-Wilson perfect simulation scheme for processes with long memory, 2011.

[GC11] A. Garivier and O. Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *23rd Conf. Learning Theory (COLT)*, Budapest, Hungary, 2011.

[GGG+10] A. Galves, C. Galves, J. Garcia, N.L. Garcia, and F. Leonardi. Context tree selection and linguistic rhythm retrieval from written texts. *ArXiv : 0902.3619*, pages 1–25, 2010.

[GGG11] A. Galves, A. Garivier, and E. Gassiat. Joint estimation of intersecting context tree models, 2011.

[Git79] J.C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, 41(2) :148–177, 1979.

[GL08] A. Galves and F. Leonardi. *Exponential inequalities for empirical unbounded context trees*, volume 60 of *Progress in Probability*, pages 257–270. Birkhauser, 2008.

[GL11] A. Garivier and F. Leonardi. Context tree selection : A unifying view. *Stochastic Processes and their Applications*, 121(11) :2488–2506, Nov. 2011.

[GM07] 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.

[GM11] A. Garivier and E. Moulines. On upper-confidence bound policies for non-stationary bandit problems. In *Algorithmic Learning Theory (ALT)*, volume 6925 of *Lecture Notes in Computer Science*, 2011.

[GMDS08] A. Galves, V. Maume-Deschamps, and B. Schmitt. Exponential inequalities for VLMC empirical trees. *ESAIM Probab. Stat*, 12 :43–45, 2008.

[Goo53] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40 :237–264, 1953.

[GS95] W. A. Gale and G. Sampson. Good-turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3) :217–237, 1995.

[GSS93] N. Gordon, D. Salmond, and A. F. Smith. Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F, Radar Signal Process.*, 140 :107–113, 1993.

[GWG02] M.R. Greenstreet, A. Winstanley, and A. Garivier. An event spacing experiment. In *ASYNC*, pages 42–51, 2002.

[Har55] Theodore E. Harris. On chains of infinite order. *Pacific J. Math.*, 5 :707–724, 1955.

[Hay05] S. Haykin. Cognitive radio : Brain-empowered wireless communications. *IEEE Journal of Selected Areas in Communications*, 23(2) :201–220, 2005.

[HDdFJ07] M. Hoffman, A. Doucet, N. de Freitas, and A. Jasra. Bayesian policy learning with trans-dimensional mcmc. In *NIPS*, 2007.

[HGB+06] C. Hartland, S. Gelly, N. Baskiotis, O. Teytaud, and M. Sebag. Multi-armed bandit, dynamic environments and meta-bandits, 2006. *NIPS-2006 workshop, Online trading between exploration and exploitation*, Whistler, Canada.

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301) :13–30, 1963.

[HT10] J. Honda and A. Takemura. An asymptotically optimal bandit algorithm for bounded support models. In T. Kalai and M. Mohri, editors, *Conf. Comput. Learning Theory*, Haifa, Israel, 2010.

[JOA10] T. Jaksch, R. Ortner, and P. Auer. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research*, 11 :1563–1600, 2010.

[JS08] W. Jank and G. Shmueli. *Statistical Methods in E-commerce Research*. Wiley-Interscience, 2008.

[KS06] L. Kocsis and C. Szepesvári. Discounted UCB. *2nd PASCAL Challenges Workshop*, Venice, Italy, April 2006.

[KSST08] S.M. 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.

[KX08] D. E. Koulouriotis and A. Xanthopoulos. Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems. *Applied Mathematics and Computation*, 196(2) :913–922, 2008.

[Lal86] S. P. Lalley. Regenerative representation for one-dimensional Gibbs states. *Ann. Probab.*, 14(4) :1262–1271, 1986.

[Lal00] S. P. Lalley. Regeneration in one-dimensional Gibbs states and chains with complete connections. *Resenhas*, 4(3) :249–281, 2000.

[LEGJVP08] L. Lai, H. El Gamal, H. Jiang, and H. Vincent 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.

[Leo10] F. Leonardi. Some upper bounds for the rate of convergence of penalized likelihood context tree estimators. *Brazilian Journal of Probability and Statistics*, 24(2) :321–336, 2010.

[Lep91] O.V. Lepskiĭ. Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatnost. i Primenen.*, 36(4) :645–659, 1991.

[LGJP11] L. Lai, H. El Gamal, H. Jiang, and H. Vincent Poor. Cognitive medium access : Exploration, exploitation, and competition, 2011.

[LGX+08] 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.

[LNDF08] J-L. 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.

[LR85] T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1) :4–22, 1985.

[LZ08a] 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.

[LZ08b] 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.

[Mas03] P. Massart. *Ecole d'Eté de Probabilité de Saint-Flour XXXIII*, chapter Concentration inequalities and model selection. LNM. Springer-Verlag, 2003.

[McD89] Colin McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Math. Soc. Lecture Note Ser.*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.

[Mei06] Yajun Mei. Sequential change-point detection when unknown parameters are present in the pre-change distribution. *Ann. Statist.*, 34(1) :92–122, 2006.

[Mit00] J. Mitola. *Cognitive Radio - An Integrated Agent Architecture for Software Defined Radio*. PhD thesis, Royal Institute of Technology, Kista, Sweden, May 8 2000.

[MS00] David A. McAllester and Robert E. Schapire. On the convergence rate of good-turing estimators. In *COLT*, pages 1–6, 2000.

[Nev72] J. Neveu. *Martingales à temps discret*. Masson, 1972.

[OM35a] O. Onicescu and G. Mihoc. Sur les chaînes de variables statistiques. *Bull. Sci. Math.*, 59 :174–192, 1935.

[OM35b] O. Onicescu and G. Mihoc. Sur les chaînes statistiques. *Comptes Rendus de l'Académie des Sciences de Paris*, 200 :511–512, 1935.

[PCA07] S. Pandey, D. Chakrabarti, and D. Agarwal. Multi-armed bandit problems with dependent arms. *International Conference on Machine learning*, pages 721–728, 2007.

[PS99] M. K. Pitt and N. Shephard. Filtering via simulation : Auxiliary particle filters. *J. American Statist. Assoc.*, 94(446) :590–599, 1999.

[PT94] CH Papadimitriou and JN Tsitsiklis. The complexity of optimal queueing network control. *Structure in Complexity Theory Conference*, pages 318–322, 1994.

[Put94] M.L. Puterman. *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc. New York, NY, USA, 1994.

[PW96] James Gary Propp and David Bruce Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. In *Proceedings of the Seventh International Conference on Random Structures and Algorithms (Atlanta, GA, 1995)*, volume 9, pages 223–252, 1996.

[Ris83] J. Rissanen. A universal data compression system. *IEEE Trans. Inform. Theory*, 29(5) :656–664, 1983.

[RT08] P. Rusmevichientong and J.N. Tsitsiklis. Linearly parameterized bandits. Arxiv preprint arXiv :0812.3465, 2008.

[SA11] Antoine Salomon and Jean-Yves Audibert. Deviations of stochastic bandit regret. In *Algorithmic Learning Theory (ALT)*, volume 6925 of *Lecture Notes in Computer Science*, 2011.

[SJ00] P. Schniter and C. R. Johnson. Bounds for the MSE performance of constant modulus estimators. *IEEE Transactions on Information Theory*, 46(7) :2544–2560, 2000.

[SKSS10] N. Srinivas, A. Krause, S.Kakade, and M. Seeger. Gaussian process optimization in the bandit setting : No regret and experimental design. In *ICML'10*, pages 1015–1022, 2010.

[SL08] A.L. Strehl and M.L. Littman. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8) :1309–1331, 2008.

[SU08] A. Slivkins and E. Upfal. Adapting to a changing environment : the brownian restless bandits. In *Proceedings of the Conference on 21st Conference on Learning Theory*, pages 343–354, 2008.

[TS06] M. Toussaint and A. Storkey. Probabilistic inference for solving discrete and continuous state Markov Decision Processes. In *ICML '06 : Proceedings of the 23rd international conference on Machine learning*, pages 945–952, New York, NY, USA, 2006. ACM.

[TV05] D. Tse and P. Viswanath. *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[Whi88] P. Whittle. Restless bandits : Activity allocation in a changing world. *Journal of Applied Probability*, 25 :287–298, 1988.

[WKVP05] C.C. Wang, S.R. Kulkarni, and H. Vincent Poor. Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3) :338–355, 2005.

[WST95] F.M.J. Willems, Y.M. Shtarkov, and T.J. Tjalkens. The context-tree weighting method : Basic properties. *IEEE Trans. Inf. Theory*, 41(3) :653–664, 1995.

[YM09] Jia Yuan Yu and Shie Mannor. Piecewise-stationary bandit problems with side observations. In *ICML '09 : Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1177–1184, New York, NY, USA, 2009. ACM.

[ZKL08] 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.

[ZTSC07] 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, 2007.