Analysis of bayesian and frequentist strategies for sequential resource allocation

Emilie Kaufmann

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

TELECOM ParisTech

2014-ENST-0056

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal et Images »

présentée et soutenue publiquement par

Emilie KAUFMANN

le 1er octobre 2014

Analyse de stratégies bayésiennes et fréquentistes pour l'allocation séquentielle de ressources

Directeur de thèse : **Olivier CAPPE**
Co-encadrement de la thèse : **Aurélien GARIVIER** et **Rémi MUNOS**

Jury

M. Gérard BIAU, Professeur, LSTA, Université Pierre et Marie Curie	Examinateur
M. Thomas BONALD, Professeur, LTCI, Telecom ParisTech	Examinateur
M. Olivier CAPPE, Directeur de recherche, LTCI, CNRS & Telecom ParisTech	Co-directeur de thèse
M. Olivier CATONI , Directeur de recherche, DMA, CNRS & Ecole Normale Supérieure	Rapporteur
M. Nicolò CESA-BIANCHI, Professeur, Università degli Studi di Milano	Rapporteur
M. Aurélien GARIVIER, Professeur, IMT, Université Paul Sabatier	Co-directeur de thèse
M. Jean-Michel MARIN, Professeur, I3M, Université de Montpellier II	Examinateur
M. Rémi MUNOS, Directeur de recherche, INRIA, Equipe Sequel	Co-directeur de thèse

TELECOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

T H È S E

Remerciements

L'heure est venue de mettre la touche finale à ce manuscrit, en remerciant toutes les personnes sans qui ce dernier n'existerait pas, et qui ont fait de ces trois années de thèse une expérience inoubliable.

Tout d'abord un grand merci à mes directeurs de thèse. Vous avez su me proposer un sujet passionnant, et doser avec justesse encadrement et liberté de recherche tout au long de ces trois ans. Aurélien, merci d'avoir pu me consacrer des journées entières lors de nos séances de travail à Toulouse. Merci pour ton accueil chaleureux et pour tous tes conseils, qu'ils soient scientifiques ou non. Olivier, merci pour toutes nos discussions face à un tableau, un bon repas ou même une caméra, et ton aide au quotidien (même informatique) malgré ton emploi du temps chargé. Merci à vous deux pour vos encouragements, votre disponibilité et vos relectures attentives. Rémi, merci pour ton accueil fréquent à Lille lors de mes deux premières années de thèse, et pour notre travail sur le Thompson Sampling. Tu m'a aussi permis de collaborer avec Nathan, que je remercie dans ces lignes pour nos échanges nombreux et fructueux. I also want to thank Shivaram for pointing a new problem to me at ICML in 2012. It was great working with you, even from far, and I hope we will see each other again at some other conference!

Nicolò and Olivier, I am really honored that you accepted to review this manuscript and I warmly thank you for your careful reading. Merci également à Gérard, Thomas et Jean-Michel de s'être intéressés à mon travail, et d'avoir pris le temps d'assister à cette soutenance.

Le chemin de l'apprenti chercheur est fait de grandes joies (finir une preuve !) mais aussi de grandes frustrations (elle était fausse...). Pendant ces trois ans, cela aura été un plaisir de les partager (ou de les oublier) avec mes collègues doctorants et jeunes docteurs du 37, rue Dareau. Un merci tout particulier à mes co-bureau du DA320 pour leur bonne humeur. Certains sont maintenant docteurs (Joffrey, Alexandre, Sylvain, Yao), d'autres le seront bientôt (Adrien, Yasir et Andrés) et à Claire à qui je passe le flambeau, je souhaite trois belles années. Il me serait difficile de citer, sans oublier personne, tous ceux avec qui j'ai partagé pauses café, puis pauses thé, gâteaux ou autres bières post-séminaire, soirées gastronomiques ou encore un week-end incroyable en Roumanie. Ils se reconnaîtront, et sauront comme tout cela aura été précieux pour moi ! Une mention spéciale à Cristina, pour avoir été une colloc' géniale en plus d'une collègue formidable. A Amandine, qui m'a fait découvrir des thés délicieux (tout comme Ruocong), re-découvrir Hermann (et d'autres recettes), et aura toujours été d'une grande aide dans les formalités administratives. A Andrés pour une séance shopping mémorable (et d'autres à venir). A Emilie, pour ses astuces LaTeX et une séance de bricolage fort amusante rue Dareau. Au club des fans de Question Pour Un Champion (Sylvain^2 et Olivier). A Eric: non il n'y aura pas de concours de pot de thèse ! En plus de mes collègues, je souhaiterais remercier Janique et Laurence, interlocutrices toujours agréables et efficaces pour les ordres de mission, Fabrice et Sophie-Charlotte pour leur aide informatique (ainsi que Nicolas !), et Florence pour son aide dans tous les domaines ainsi que son soutien aux activités du Bureau des Doctorants. Un clin d'oeil à Georges, Simon, Dong-Bach et tous les autres pour cette belle aventure associative.

Ces trois années ont été enrichissantes en terme de rencontres et de voyages. D'abord au travers des séminaires et autres conférences (souvent) au soleil, où j'ai eu la chance de rencontrer des jeunes chercheurs fort sympatiques. Je pense en particulier aux doctorants et post-doc de l'équipe Sequel, aux randonneurs d'Aussois (et à ceux des îles Canaries), aux skieurs de Grenade, aux compagnons de tapas à Barcelone ou aux organisateurs d'YSP toujours motivés. J'ai aussi eu la chance de passer trois mois à l'Université de Princeton, ce qui m'a donné l'occasion de prendre la mesure de la recherche outre-Atlantique. Merci à Sébastien de m'avoir accueillie. J'ai beaucoup apprécié nos discussions et j'espère que nous aurons l'occasion de travailler ensemble dans le futur. Merci à tous les amis "américains" (qui ne le sont pas pour la plupart) qui ont rendu ce séjour agréable. Thanks to Rofoldo and Theo for making me feel at home during these three months. Thanks Aga for having been such a great office mate, but also Andrew, Flavia and Tana for our little Princeton gang. Thanks to the welcoming 'Frenchies' I met there and also to some graduate students at ORFE with whom we shared nice moments.

Mes remerciements vont bien sûr aussi à mes amis "parisiens" (qui ne le sont pas pour la plupart) et alsaciens (qui le sont pour la plupart) qui m'ont cotoyée et soutenue pendant cette aventure. Je pense notamment à Camille, Arthur et Sonia, Timon et Tiphaine (et Elise qui nous a rejoints en cours de route), Nina, Laure, Vincent, Elsa, Guillaume, Sarah (C.) et son Cricri, Pierre, Jean, et les autres amis rencontrés pendant les années Cachanaises, avec qui discuter de sa thèse est presque un passage obligé. Aux collocs, Evelyne et Yacine, pour les bons moments au 13, rue Bellier Dedouvre. A Sarah (L.), une autre alsacienne conquise par la capitale. A mes amis d'enfance (ou presque) pour qui la lecture d'une page au hasard de ce document est hilarante (et qui me croient désormais membre de la mafia avec tous ces bandits...), merci d'être toujours présents quand je rentre en Alsace ! Les "filles": Agnès, Marie-Madeleine, Martine, Marion, Marie, Céline, et tous ceux qui se sont rajoutés avec bonheur à notre petit groupe, c'est toujours un plaisir de vous retrouver !

Je ne serais pas arrivée jusqu'ici sans l'appui de ma famille. A mon papa, qui m'a familiarisée dès mon plus jeune âge avec les problèmes de remplissage de baignoires qui fuient ou autres croisements de trains, et à ma maman, qui a toujours eu peur que ses enfants détestent l'école, je veux dire un grand merci. Merci pour tout ce que vous m'avez transmis, pour votre soutien durant ces longues années d'études, et pour votre présence dans ce moment important. Merci à Stéphane pour cette complicité que nous partageons, et pour ta capacité à toujours croire en ta grande soeur. Tu as avant moi fait l'expérience de l'écriture d'un manuscrit qui intéressera à coup sûr de plus nombreux lecteurs que celui-ci ([Kaufmann, 2014]). Longue vie à l'imaginaire du roi Barry, et à ta muse Ombline. Plus largement, merci à tous les membres de ma famille pour tous ces bons moments partagés, autour d'une tarte flambée, d'une partie de carte ou autre, et pour tous ceux à venir. Un merci ému à Pépé et Mémé.

Romain, c'est à toi que s'adressent ces derniers mots. Quand nous nous sommes rencontrés tu achevais l'écriture de ta thèse, et une autre thèse plus tard –la mienne cette fois– ta présence à mes côtés est devenue une évidence. Merci pour tous ces beaux moments et pour ton soutien de chaque instant.

Contents

Introduction et présentation des résultats

Cette thèse s'est déroulée au sein du LTCI (Laboratoire Traitement et Communication de l'Information) à Telecom ParisTech, sous la co-direction d'Olivier Cappé et d'Aurélien Garivier. Elle a été agrémentée de quelques visites à l'Université Paul Sabatier à Toulouse, où Aurélien Garivier est désormais professeur. Cette thèse a également été co-encadrée par Rémi Munos, ce qui m'a amenée à travailler ponctuellement avec lui à l'INRIA Lille.

L'objectif de cette thèse est de proposer et d'analyser de nouvelles stratégies optimales pour des problèmes d'allocation séquentielle de ressources dans un environnement aléatoire. L'environnement est constitué de plusieurs options (certaines étant meilleures que d'autres) qui peuvent être testées, et produisent des résultats aléatoires. Nos ressources correspondent aux tests que nous pouvons effectuer, et le but est de déterminer des stratégies d'allocation de ce budget de test qui permettent de réaliser certains objectifs (par exemple identifier les meilleures options). Un modèle statistique naturel pour de telles situations est le *modèle de bandit stochastique à plusieurs bras*. L'objectif de ce chapitre est d'introduire les problèmes de bandits que nous avons considérés, et de présenter nos contributions. Celles-ci seront détaillées dans les chapitres suivants, où on trouvera les preuves des résultats énoncés, ainsi que des éléments bibliographiques plus précis.

La section 1 est consacrée à la présentation des *modèles de bandit* et des deux *problèmes de bandit* étudiés dans cette thèse : la maximisation des récompenses d'une part, et l'identification des meilleurs bras d'autre part. Dans chaque cas, nous nous attacherons en particulier à la définition d'un critère d'optimalité. La section 2 présente nos contributions relatives à la maximisation des récompenses. Dans des modèles de bandit paramétriques simples, ce problème est bien compris puisqu'il existe une borne inférieure asymptotique sur le *regret* d'un algorithme efficace, ainsi que des algorithmes atteignant cette borne. Bien que le regret soit une mesure de performance fréquentiste, nous montrons que deux algorithmes d'inspiration bayésienne, Bayes-UCB et Thompson Sampling, sont également (asymptotiquement) optimaux du point de vue du regret, sont plus simples d'implémentation que les algorithmes optimaux existants et se généralisent facilement à des modèles plus complexes, comme les modèles dits contextuels. La section 3 présente nos contributions à l'identification des meilleurs bras (ou exploration pure). Nous présentons et analysons deux algorithmes, KL-LUCB et KL-Racing, basés sur des intervalles de confiance construits à l'aide de la divergence de Kullback-Leibler, transposant au cadre de l'exploration pure des améliorations récentes obtenues pour la minimisation du regret. Nous proposons également une borne inférieure sur le nombre moyen d'échantillons des bras nécessaires pour identifier les m meilleurs bras, qui ne permet toutefois pas de prouver l'optimalité des algorithmes proposés. La complexité de l'identification des meilleurs bras est en effet moins bien comprise que celle de la minimisation du regret. Nous introduisons ici une notion de complexité, qui nous permettra, pour des exemples importants de modèles de bandit à deux bras, d'identifier des algorithmes optimaux.

Dans notre présentation, nous tâcherons de souligner les outils théoriques utilisés. Nous distinguons

principalement ceux liés à l'analyse d'algorithmes (inégalités de déviations, construction d'intervalles de confiance), et ceux liés à l'obtention de bornes inférieures sur la performances de ces algorithmes (liés aux changements de loi).

1 Présentation des problèmes de bandit étudiés

Un modèle de bandit stochastique à plusieurs bras (ou simplement modèle de bandit dans la suite) est une collection de K lois de probabilités $\nu = (\nu_1, \dots, \nu_K)$ supposées indépendantes, que l'on désigne par « bras » ou « options ». On note μ_a la moyenne du bras a, c'est-à-dire de la distribution associée ν_a, et on introduit $\mu^\star = \max_a \mu_a$ et $a^\star$ tel que $\mu^\star = \mu_{a^\star}$. Un agent, qui ne connaît pas ν, peut interagir avec ce modèle de bandit. Il choisit à chaque instant t un bras A_t et observe une réalisation de la loi sous-jacente, $X_t \sim \nu_{A_t}$. Le bras A_t est choisi en fonction des observations passées de l'agent, $A_1, X_1, \dots, A_{t-1}, X_{t-1}$

La suite de variables aléatoires $(A_t)_{t\in\mathbb{N}^*}$ est la *stratégie d'échantillonnage* des bras adoptée par l'agent, parfois appelée *politique* ou *algorithme de bandit*. Elle est notée $\mathcal{A}$ dans la suite. Bien évidemment, cette stratégie va fortement dépendre de l'objectif de l'agent, c'est-à-dire du *problème de bandit* considéré. Dans cette thèse il sera question de deux objectifs différents, qui visent globalement à identifier les meilleurs bras, mais sous des contraintes différentes.

Le terme « bandit » provient des bandits manchots, qui désignent les machines à sous. Le cadre probabiliste décrit ci-dessus peut en effet modéliser un casino, où l'on suppose que lorsqu'on tire son bras, chaque machine à sous délivre une récompense qui suit une certaine loi de probabilité (inconnue du joueur, bien évidemment). Un objectif naturel pour l'agent (le joueur) est de maximiser la somme des récompenses obtenues pendant son interaction avec le modèle de bandit (les machines à sous). Cette somme de récompenses étant aléatoire, un objectif raisonnable est de s'attacher à construire une stratégie maximisant pour un horizon T donné (le temps de jeu) l'espérance de la somme des récompenses obtenues. C'est cet objectif qui est considéré dans l'article de [Thompson, 1933] présentant le premier algorithme de bandit. Le cadre applicatif des essais cliniques qui y est décrit a véritablement motivé l'étude d'algorithmes de bandits (au contraire de l'exemple du casino, qui n'est qu'un prête-nom) et nous le présentons ici. Un médecin possède pour un symptôme donné K traitements possibles. Une probabilité de guérison inconnue, p_a, est associée au traitement a. Lorsque qu'il donne traitement A_t au t-ème patient, il observe la réponse du patient X_t, qui vaut 1 s'il guérit, 0 sinon (et qui constitue en quelque sorte la 'récompense' du médecin). On peut faire l'hypothèse que X_t suit une loi de Bernoulli de moyenne p_{A_t}, et que les réponses des différents patients sont indépendantes. L'objectif du médecin est de construire une stratégie d'allocation des traitements (A_t) qui maximise l'espérance du nombre de patients guéris, soit l'espérance de la somme de ses récompenses. La stratégie du médecin sera alors un compromis entre *exploration* (essai des traitements peu donnés pour estimer leur efficacité) et *exploitation* (tendance à privilégier le traitement qui a paru le plus efficace jusque-là).

Le fait d'interpréter les observations des bras comme des récompenses nous place naturellement dans le cadre plus général de l'apprentissage par renforcement, où l'interaction d'un agent avec son environnement est modélisée par un Processus Décisionnel de Markov (MDP pour *Markov Decision Process* en anglais). Un MDP est un quadruplet $(\mathcal{X}, \mathcal{A}, \mathcal{P}, \mathcal{R})$ où $\mathcal{X}$ désigne l'espace d'états, $\mathcal{A}$ l'espace d'actions, $\mathcal{P} : \mathcal{X}, \mathcal{A} \to \mathcal{M}_1(\mathcal{X})$ le noyau de transition et $\mathcal{R} : \mathcal{X}, \mathcal{A} \to \mathcal{M}_1(\mathcal{X})$ le noyau de récompenses. Lorsqu'il est dans l'état x, et qu'il choisit l'action a, l'agent reçoit une récompense $r \sim \mathcal{R}(\cdot|x, a)$ et effectue une transition vers un état $y \sim \mathcal{P}(\cdot|x, a)$. Le but de l'agent est de trouver une politique (une fonction indiquant quelle action choisir dans un état donné) qui maximise l'espérance de ses récompenses dans un MDP de paramètres inconnus. Le problème de bandit décrit ci-dessus peut donc être vu comme le problème

d'apprentissage par renforcement le plus simple : on a un seul état, x_0, et le le noyau de récompenses est donné par $\mathcal{R}(\cdot|x_0, a) = \nu_a$. Nous verrons au chapitre 3 que certains des algorithmes proposés dans cette thèse peuvent se généraliser au cadre de l'apprentissage par renforcement. Par ailleurs, la théorie des Processus Décisionnels de Markov nous sera utile au chapitre 1, où nous verrons que la formulation bayésienne d'un problème de bandit peut être modélisée par un MDP.

Un objectif différent peut être envisagé par l'agent : celui d'identifier le ou les meilleurs bras (c'est-à-dire ceux qui ont les moyennes les plus élevées), mais sans la contrainte de maximiser la somme des réalisations des bras X_t obtenues. Ces dernières ne sont alors plus perçues comme des récompenses. Pour bien comprendre la différence avec le problème précédent, considérons un autre exemple qui motive actuellement l'étude des modèles de bandits : celui de la publicité en ligne.

Un site Internet dispose d'un (ou plusieurs) emplacements publicitaires et son gestionnaire peut choisir les annonces qu'il veut y mettre parmi un panel de K publicités. Il est payé par les annonceurs en fonction du nombre de clics sur leur publicité. On peut modéliser de manière simple la réponse X_t du t-ème visiteur du site (clic ou non-clic sur la publicité qu'on lui présente) par une variable aléatoire de Bernoulli de moyenne p_a si on lui présente la publicité a. Le gestionnaire du site peut choisir de maximiser le nombre de clics -la somme des X_t- sans chercher à estimer précisément la probabilité de clic sur chacune des publicités, ce qui revient à maximiser ses récompenses dans le modèle de bandit associé. Le nombre de visiteurs du site étant grand, il peut aussi décider de procéder en deux phases : d'abord déterminer le (ou les) publicités ayant les probabilité de clic les plus élevées (en acceptant de présenter des 'mauvaises' publicités, et donc de perdre de l'argent pendant cette phase). Au terme de cette phase, il ne présentera plus que les meilleurs publicités sur son site. S'il adopte une telle stratégie, le gestionnaire du site dissocie la phase d'exploration de la phase d'exploitation.

Lorsque l'objectif de l'agent est d'identifier le(s) meilleur(s) bras, on parlera en effet d'*exploration pure*, par opposition au *compromis entre exploration et exploitation* auquel il faut parvenir lorsque l'objectif est la maximisation des récompenses. Si la maximisation des récompenses peut être vue comme un problème d'apprentissage par renforcement, l'identification des meilleur bras est plutôt un cas particulier de problème d'optimisation d'une fonction bruitée. Nous allons voir que les algorithmes pour ces deux objectifs, ainsi que leur complexité, sont de nature différente.

Quelques définitions et notations. Dans cette thèse nous allons considérer des classes de modèles de bandit à K bras, notées en général $\mathcal{M}$, pour lesquelles nous voudrons trouver des algorithmes efficaces pour l'ensemble des modèles de bandit de la classe $\mathcal{M}$. Par exemple on s'intéressera à la classe des *modèles de bandit binaires*, où le bras a est une distribution de Bernoulli de moyenne μ_a, $\mathcal{B}(\mu_a)$, qui permet de modéliser de nombreuses applications pratiques, comme on l'a vu plus haut.

Plus généralement, on considèrera les classes de *modèles de bandit paramétriques*, où la distribution du bras a dépend d'un paramètre θ_a : $\nu_a = \nu_{\theta_a}$, avec $\theta_a \in \Theta$. Un cas particulier important est celui des *modèles de bandit exponentiels*. Une classe $\mathcal{M}$ de modèles de bandit exponentiels est telle qu'il existe des fonctions A et b telle que pour tout $\nu \in \mathcal{M}$, la distribution ν_{θ_a} du bras a admet pour densité

$$f(x; \theta_a) = A(x) \exp(\theta_a x - b(\theta)). \tag{1}$$

En d'autres termes, les distributions des bras appartiennent à une famille exponentielle canonique à un paramètre. De telles distributions peuvent également être paramétrées par leur moyenne $\mu(\theta) = \dot{b}(\theta)$, ce qui permet d'introduire la fonction de divergence suivante, associée à une famille exponentielle donnée, qui correspond à la divergence de Kullback-Leibler entre deux distributions de cette famille, exprimée

en fonction de leurs moyennes :

$$d(\mu, \mu') = \mathrm{KL}(\nu_{\dot{b}^{-1}(\mu)}, \nu_{\dot{b}^{-1}(\mu')}), \tag{2}$$

où $\mathrm{KL}(p, q)$ désigne la divergence de Kullback-Leibler (ou KL-divergence) entre les distributions p et q, définie par

$$\mathrm{KL}(p, q) = \begin{cases} \int \log\left[\frac{dp}{dq}(x)\right] dp(x) \text{ si } q \ll p, \\ +\infty \text{ sinon.} \end{cases}$$

Les modèles de bandit binaires sont un cas particulier des modèles de bandit exponentiels, si la loi de Bernoulli de moyenne μ est paramétrée par son paramètre naturel $\theta = \log(\mu/(1-\mu))$. Pour les bandits binaires, on a

$$d(x, y) = \mathrm{KL}(\mathcal{B}(x), \mathcal{B}(y)) = x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y}.$$

1.1 Maximisation des récompenses : mesure de performance et objectifs

Soit $\nu = (\nu_1, \ldots, \nu_K)$ un modèle de bandit. On rappelle que $\mu^* = \max_a \mu_a$ désigne la moyenne du meilleur bras. Un algorithme de bandit $\mathcal{A} = (A_t)_{t\in\mathbb{N}}$ qui maximise les récompenses minimise de manière équivalente une quantité appelée *regret*, qui mesure l'écart entre la récompense moyenne obtenue si on n'avait tiré que le meilleur bras et la récompense moyenne effectivement obtenue par la stratégie. Le regret d'une stratégie $\mathcal{A}$ à l'horizon T est défini par

$$\mathrm{R}_\nu(T, \mathcal{A}) = \mathbb{E}_\nu\left[T\mu^* - \sum_{t=1}^{T} X_t\right].$$

Le regret peut aussi se réécrire de la manière suivante en introduisant $N_a(t)$, le nombre de tirages du bras a entre les instants 1 et t :

$$\mathrm{R}_\nu(T, \mathcal{A}) = \mathbb{E}_\nu\left[\sum_{t=1}^{T}(\mu^* - \mu_{A_t})\right] = \sum_{a=1}^{K}(\mu^* - \mu_a)\mathbb{E}_\nu\left[N_a(T)\right]. \tag{3}$$

La notion de regret a été introduite par [Lai and Robbins, 1985], qui donnent également une borne inférieure sur le regret d'une stratégie $\mathcal{A}$ vérifiant $\mathrm{R}_\nu(T, \mathcal{A}) = o(T^\alpha)$ pour tout $\alpha \in]0, 1[$ et tout modèle de bandit dans la classe $\mathcal{M}$ possédant un unique bras optimal. Un telle stratégie est dite *uniformément efficace*. Le résultat de Lai et Robbins est valable pour certaines classes de modèles de bandit paramétriques telles que les distributions des bras dépendent d'un paramètre réel. En particulier, il est vrai pour $\mathcal{M}$ une classe de modèles de bandit exponentiels, et nous l'énonçons dans ce cadre, en rappelant que la fonction $d(\mu, \mu')$ désigne la divergence de Kullback-Leibler (dans une famille exponentielle donnée) entre les distributions de moyennes μ et μ'.

Théorème 1. *Soit $\mathcal{A}$ un algorithme uniformément efficace. Pour tout modèle de bandit $\nu \in \mathcal{M}$, et tout bras sous-optimal a,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\log(T)} \geq \frac{1}{d(\mu_a, \mu^*)}$$

En utilisant (3), on obtient la borne inférieure suivante sur le regret :

$$\liminf_{T\to\infty} \frac{R_\nu(T, \mathcal{A})}{\log(T)} \geq \sum_{a:\mu_a<\mu^*} \frac{(\mu^* - \mu_a)}{d(\mu_a, \mu^*)}. \tag{4}$$

Un algorithme dont le regret atteint la borne (4) est dit *asymptotiquement optimal*. Des raffinements successifs dans l'analyse des algorithmes de bandit ont conduit à l'introduction d'algorithmes asymptotiquement optimaux pour lesquels une *analyse à horizon fini* est proposée, c'est-à-dire une majoration non asymptotique de leur regret. Un exemple de tel algorithme est l'algorithme KL-UCB, introduit par [Cappé et al., 2013].

Cet algorithme s'inscrit dans la lignée des *politiques d'indices* dites de type UCB, c'est-à-dire utilisant un sommet d'intervalle de confiance (*Upper Confidence Bound* en anglais) (voir par exemple [Auer et al., 2002a, Audibert et al., 2009]). Une politique d'indices calcule pour chaque bras un indice ne dépendant que des observations passées de ce bras, et choisit le bras d'indice maximal. KL-UCB choisit à l'instant $t+1$ le bras $A_{t+1} = \text{argmax}_a u_a(t)$, où $u_a(t)$ est l'indice suivant, qui apparaît comme le sommet d'un intervalle de confiance basé sur la divergence de Kullback-Leibler :

$$u_a(t) = \sup\{q \geq \hat{\mu}_a(t) : N_a(t) d\left(\hat{\mu}_a(t), q\right) \leq \beta(t)\}, \quad \text{avec} \quad \beta(t) = \log t + 3\log\log t \tag{5}$$

où $\hat{\mu}_a(t)$ désigne la moyenne empirique des observations issues du bras a collectées entre les instants 1 et t, et d est la fonction de divergence associée à la famille exponentielle considérée (voir (2)).

Dans cette thèse, nous avons cherché à proposer de nouveaux algorithmes de bandits qui, tout en conservant la propriété d'optimalité asymptotique dans des modèles simples, ont de meilleures performances pratiques et un meilleur pouvoir de généralisation. Pour ce faire, nous avons adopté une approche bayésienne, discutée à la section 2.

1.2 Identification des meilleurs bras : mesure de performance et objectifs

Fixons $m \in \{1, \dots, K\}$ et supposons maintenant que l'agent cherche à identifier les m meilleurs bras. On note $(\mu_{[1]}, \dots, \mu_{[K]})$ le réarrangement décroissant des moyennes des bras, et on considère des classes $\mathcal{M}_m$ de modèles de bandit telles que pour tout $\nu \in \mathcal{M}_m$, $\mu_{[m]} > \mu_{[m+1]}$, de sorte que l'ensemble $\mathcal{S}_m^*$ des m bras ayant les plus grandes moyennes est défini sans ambiguïté.

Comme dans le cadre de la maximisation des récompenses, l'agent choisit séquentiellement les bras dont il veut obtenir des échantillons, selon une *règle d'échantillonnage* $(A_t)_{t\in\mathbb{N}}$, qui détermine quel bras tirer en fonction des observations passées. Mais il lui appartient aussi de décider quand arrêter son échantillonnage, selon une *règle d'arrêt* τ, qui est un temps d'arrêt par rapport à la filtration $\mathcal{F}_t = \sigma(A_1, X_1, \dots, A_t, X_t)$, et de choisir un ensemble de m bras $\hat{S}_m$ ($\mathcal{F}_\tau$-mesurable) selon une *règle de recommandation*. Le triplet $\mathcal{A} = ((A_t), \tau, \hat{S}_m)$ constitue sa stratégie.

La stratégie de l'agent peut être adaptée à deux contraintes différentes considérées dans la littérature, que l'on comprend bien à travers un exemple pratique. Imaginons qu'une entreprise cherche à déterminer les m meilleurs produits parmi K possibles afin de les lancer sur le marché, et puisse pendant une phase de test les proposer à des clients pour observer leur réaction (la réponse à chaque produit étant modélisée par des échantillons d'une distribution qui lui est associée). Pendant cette phase de test, l'entreprise perd de l'argent puisqu'elle accepte de présenter des mauvais produits à ses clients. Pour des raisons économiques, on peut donc imaginer qu'elle fixe le nombre de clients participants à l'étude de marché, et cherche alors à minimiser la probabilité de ne pas trouver les m meilleurs produits. Si ce « budget » est choisi trop petit, la probabilité d'erreur ne pourra peut-être pas être rendue très petite. Un autre type de contrainte possible est que l'entreprise fixe un seuil pour la probabilité d'erreur. Elle veut identifier les m meilleurs produits avec une probabilité supérieure à 0.95 par exemple : son but est alors d'atteindre ce seuil en minimisant le nombre de clients impliqués dans l'étude.

La formulation mathématique de ces deux contraintes est la suivante. Dans le cadre de l'identification des meilleurs bras *à budget fixé* (*fixed-budget setting* en anglais), le nombre de tirages des bras τ est fixé

à l'avance ($\tau = t$, où t est le budget) et on cherche à trouver une règle d'échantillonnage et une règle de recommandation qui minimisent la probabilité d'erreur, que l'on note $p_t(\nu) := \mathbb{P}_\nu(\hat{S}_m \neq \mathcal{S}_m^*)$. Dans ce cadre, une stratégie est dite *consistante* si pour tout problème de bandit $\nu \in \mathcal{M}_m$, $p_t(\nu)$ tend vers 0 lorsque t tend vers l'infini. Dans le cadre de l'identification des meilleurs bras *à niveau de confiance fixé* (*fixed-confidence setting* en anglais), on cherche à construire des stratégies dites δ-*PAC* (pour *Probably Approximately Correct* en anglais), dont la probabilité d'erreur est majorée par δ sur tous les modèles de bandit : $\forall \nu \in \mathcal{M}_m, \mathbb{P}_\nu(\hat{S}_m \neq \mathcal{S}_m^*) \leq \delta$. Dans ce cadre, l'objectif est de construire des stratégies δ-PAC qui minimisent le nombre moyen d'observations utilisées, $\mathbb{E}_\nu[\tau]$.

Par analogie avec le problème de minimisation du regret considéré plus haut, nous nous sommes posé la question suivante : comment définir des algorithmes optimaux pour l'identification des meilleurs bras avec un budget ou un niveau de confiance fixé ? La notion d'optimalité asymptotique via-à-vis du regret est en effet bien justifiée par la borne de Lai et Robbins, et on a montré dans la section précédente que, pour $\mathcal{M}$ une classe de modèles de bandit exponentiels, pour tout $\nu \in \mathcal{M}$,

$$\inf_{\text{stratégies } \mathcal{A}} \limsup_{T\to\infty} \frac{R_\nu(T,\mathcal{A})}{\log(T)} = \sum_{a:\mu_a<\mu^*} \frac{(\mu^* - \mu_a)}{d(\mu_a,\mu^*)}.$$

Le terme de droite peut s'interpréter comme un terme de complexité qui dépend du modèle de bandit ν et qui fait intervenir un quantité informationnelle (la divergence de Kullback-Leibler). Nous avons vu qu'il existe des algorithmes atteignant cette complexité (c'est-à-dire réalisant l'infimum ci-dessus), que l'on a qualifié d'asymptotiquement optimaux.

Pour l'identification des meilleurs bras, nous proposons les termes de complexité suivants pour un budget fixé ($\kappa_B(\nu)$) et un niveau de confiance fixé ($\kappa_C(\nu)$) :

$$\kappa_{\mathrm{B}}(\nu) = \inf_{\mathcal{A} \text{ consistant}} \left(\limsup_{t\to\infty} -\frac{1}{t}\log p_t(\nu) \right)^{-1} \quad \text{et} \quad \kappa_{\mathrm{C}}(\nu) = \inf_{\mathcal{A}\ \delta\text{–PAC}} \limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\log\frac{1}{\delta}}. \tag{6}$$

De manière heuristique, pour atteindre une probabilité d'erreur δ, un algorithme optimal (au sens des complexités ci-dessus) à budget fixé devrait utiliser un budget $t \simeq \kappa_B(\nu)\log(1/\delta)$ et un algorithme optimal à niveau de confiance fixé nécessite un nombre moyen d'échantillons $\mathbb{E}_\nu[\tau] \simeq \kappa_C(\nu)\log(1/\delta)$. On peut donc naturellement se poser la question de la comparaison entre $\kappa_B(\nu)$ et $\kappa_C(\nu)$.

La littérature abondante sur l'identification des meilleurs bras (nous renvoyons le lecteur au chapitre 5 pour une bibliographie détaillée) fournit des bornes supérieures et inférieures sur la probabilité d'erreur d'un algorithme consistant (pour un budget fixé) ou le nombre moyen d'échantillons utilisés par un algorithme δ-PAC (pour un niveau de confiance fixé), qui conduisent naturellement à des encadrement de $\kappa_B(\nu)$ et $\kappa_C(\nu)$ respectivement. Toutefois, un écart persiste entre les bornes supérieures et inférieures, ce qui ne permet pas d'identifier les complexités. Ces bornes font intervenir des constantes multiplicatives (non nécessairement explicites) et elles sont obtenues pour des modèles de bandit où les bras sont des distributions sous-gaussiennes [1], faisant intervenir la quantité

$$H(\nu) = \sum_{a=1}^{K} \frac{1}{\Delta_a^2} \quad \text{avec} \quad \Delta_a = \begin{cases} \mu_a - \mu_{[m+1]} & \text{pour } a \in \mathcal{S}_m^*, \\ \mu_{[m]} - \mu_a & \text{pour } a \in (\mathcal{S}_m^*)^c. \end{cases}$$

Pour des bandits binaires (qui forment un cas particulier de distribution 1/4 sous-gaussiennes), l'écart entre les moyennes de deux bras $(\mu_a - \mu_a')^2$ apparait d'après l'inégalité de Pinsker [2] comme une approximation de la divergence de Kullback-Leibler entre $\mathcal{B}(\mu_a)$ et $\mathcal{B}(\mu_a')$. Par analogie avec la minimisation

1. une distribution ν_a est dite σ^2 sous-gaussienne si $\forall \lambda \in \mathbb{R}$, $\mathbb{E}_{X\sim\nu_a}[e^{\lambda(X-\mathbb{E}[X])}] \leq \exp\left(\lambda^2\sigma^2/2\right)$
2. $\mathrm{KL}(\mathcal{B}(\mu_a),\mathcal{B}(\mu_a')) > 2(\mu_a - \mu_a')^2$

du regret, on pourrait donc s'attendre à des termes de complexité $\kappa_B(\nu)$ et $\kappa_C(\nu)$ qui dépendent de quantités informationnelles.

Nous présentons en section 3 nos contributions relatives à l'identification et la comparaison des complexités de l'identification des meilleurs bras pour un budget ou un niveau de confiance fixés. Nous chercherons en particulier à identifier les quantités informationnelles caractéristiques des problèmes d'identification des meilleurs bras.

2 Des algorithmes bayésiens pour la maximisation des récompenses

Cette section présente nos contributions relatives à l'analyse d'algorithmes de bandit d'inspiration bayésienne pour l'objectif de maximisation des récompenses. Celles-ci seront détaillées dans les chapitres 2 à 5 de ce document. La section 2.1 présente les modèles de bandit bayésiens, ainsi que le critère de performance qui leur est associé. Nous nous intéressons également à des approximations de la solution bayésienne du problème de bandit, comme une approximation basée sur les indices de Gittins à horizon fini. Notre objectif était de proposer de nouveaux algorithmes asymptotiquement optimaux du point de vue du regret. Bien que ce dernier soit une quantité fréquentiste, nous avons pu obtenir de telles garanties pour deux algorithmes bayésiens, Bayes-UCB et Thompson Sampling : nous présentons les analyses proposées en section 2.2. Ces deux algorithmes peuvent également être utilisés dans des modèles de bandit plus généraux, et nous discutons en particulier l'exemple des modèles contextuels linéaires à la section 2.3.

2.1 Deux approches probabilistes d'un problème de bandit

Nous considérons dans cette section des modèles de bandit paramétriques, de la forme $\nu = \nu_{\boldsymbol{\theta}} = (\nu_{\theta_1}, \dots, \nu_{\theta_K})$, où la distribution ν_a du bras a dépend d'un paramètre $\theta_a \in \Theta$. On note $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K) \in \Theta^K$ le paramètre global du modèle. Comme dans tout modèle paramétrique, deux approches sont possibles : l'approche fréquentiste où l'on considère que $\boldsymbol{\theta}$ est un paramètre inconnu, et l'approche bayésienne, où l'on considère que $\boldsymbol{\theta}$ est une variable aléatoire, qui suit une *loi a priori* Π_0.

Modèle de bandit fréquentiste	**Modèle de bandit bayésien**
- $\boldsymbol{\theta} \in \Theta^K$ est un paramètre inconnu	- $\boldsymbol{\theta}$ est tiré sous Π_0, une loi a priori sur Θ^K
- $\forall a$, $(X_{a,t})$ est i.i.d. de loi ν_{θ_a} et de moyenne μ_a	- $\forall a$, conditionnellement à θ_a, $(X_{a,t})$ est i.i.d. de loi ν_{θ_a} et de moyenne μ_a
- $(X_{a,t})_{a,t}$ est une famille indépendante	- conditionnellement à $\boldsymbol{\theta}$, $(X_{a,t})_{a,t}$ est une famille indépendante

Dans les deux modèles, l'agent choisit à l'instant t un bras A_t à tirer et observe la récompense $X_t = X_{A_t,t}$ issue du bras choisi. En introduisant la filtration

$$\mathcal{F}_t = \sigma(A_1, X_1, \dots, A_t, X_t),$$

si l'agent adopte une stratégie déterministe, la variable aléatoire A_t est $\mathcal{F}_{t-1}$-mesurable, alors que s'il adopte une stratégie randomisée, A_t est tirée selon une loi p_t sur $\{1, \dots, K\}$, et c'est le vecteur de probabilités p_t qui est $\mathcal{F}_{t-1}$-mesurable.

La notion de *regret* introduite en section 1.1 est une mesure de performance associée au modèle fréquentiste, puisqu'il dépend du modèle ν, ou de manière équivalente, du paramètre $\boldsymbol{\theta}$ dans notre cadre

paramétrique :

$$\mathrm{R}_{\boldsymbol{\theta}}(T, \mathcal{A}) = \mathbb{E}_{\boldsymbol{\theta}}\left[T\mu^\star - \sum_{t=1}^{T} X_t\right] = \sum_{a=1}^{K} (\mu^\star - \mu_a)\mathbb{E}_{\boldsymbol{\theta}}\left[N_a(T)\right].$$

$\mathbb{P}_{\boldsymbol{\theta}}$ et $\mathbb{E}_{\boldsymbol{\theta}}$ désignent la probabilité et l'espérance sous le modèle fréquentiste (qui dépend de $\boldsymbol{\theta}$). Notons $\mathbb{E}_{\Pi_0}$ et $\mathbb{P}_{\Pi_0}$ la probabilité et l'espérance sous le modèle bayésien. Une stratégie qui maximise l'espérance de la somme des récompenses sous le modèle bayésien minimise de manière équivalente le *risque bayésien* (ou regret bayésien), défini par

$$\mathrm{BR}_{\Pi_0}(T, \mathcal{A}) = \mathbb{E}_{\Pi_0}\left[T\mu^\star - \sum_{t=1}^{T} X_t\right] = \mathbb{E}_{\Pi_0}\left[\mathrm{R}_{\boldsymbol{\theta}}(T, \mathcal{A})\right].$$

La borne inférieure de [Lai and Robbins, 1985] nous a permis de définir des stratégies asymptotiquement optimales par rapport au regret, et nous en avons donné des exemple en section 1.1. Du point de vue du risque bayésien on peut aller plus loin et montrer qu'il existe une stratégie optimale, qui minimise le risque bayésien parmi toutes les stratégies possibles.

Ces deux formulations fréquentiste et bayésienne d'un même problème de bandit (celui de maximiser l'espérance de la somme des récompenses) peuvent être dissociées des outils qui leur sont associés. Dans la littérature liée à la minimisation du regret, les algorithmes proposés sont basés sur des estimateurs du maximum de vraisemblance des paramètres inconnus des bras, et sur des intervalles de confiances, que nous pouvons qualifier d'outils fréquentistes. A l'inverse, nous appellerons *algorithmes bayésiens* des algorithmes qui pour choisir le bras A_{t+1} se basent sur la loi a posteriori de $\boldsymbol{\theta}$,

$$\Pi^t(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}|A_1, X_1, \ldots, A_t, X_t),$$

qui est la loi conditionnelle de $\boldsymbol{\theta}$ sachant les observations obtenues jusqu'à l'instant t. Nous pouvons ainsi nous intéresser au risque bayésien d'un algorithme 'fréquentiste', et inversement, et c'est ce qui sera au cœur de cette thèse, à la performance d'algorithmes bayésiens évaluée en terme de regret.

Dans un premier temps, nous allons nous demander si la solution optimale du problème bayésien elle-même, ou certaines de ses approximations, peuvent fournir de nouveaux algorithmes asymptotiquement optimaux du point de vue du regret. Nous commençons par décrire ces premiers algorithmes bayésiens.

La solution bayésienne du problème de bandit. L'existence d'une solution à la minimisation du risque bayésien vient du fait qu'elle peut s'interpréter comme un problème de planification dans un Processus Décisionnel de Markov (MDP) associé. Pour simplifier la présentation, nous allons exhiber cette solution pour un modèle de bandit binaire, avec des loi a priori uniformes indépendantes sur chaque moyenne.

Soit $\nu = (\mathcal{B}(\mu_1), \ldots, \mathcal{B}(\mu_K)$ un modèle de bandit binaire. On fait l'hypothèse que pour tout a, $\theta_a \sim \mathcal{U}([0,1])$ et que les θ_a sont indépendantes. La loi a posteriori sur $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_K)$ à l'instant t prend la forme d'un produit de K marginales indépendantes $\Pi^t = (\pi_1^t, \ldots, \pi_K^t)$ et la loi a posteriori sur μ_a est donnée par

$$\pi_a^t = \mathrm{Beta}(S_a(t) + 1, N_a(t) - S_a(t) + 1),$$

où $S_a(t)$ est la somme des récompenses obtenues du bras a entre les instants 1 et t, et on le rappelle $N_a(t)$ est le nombre de tirages du bras a entre les instants 1 et t. La loi $\mathrm{Beta}(a, b)$ admet pour densité par rapport à la mesure de Lebesgue

$$f_{(a,b)}(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}\mathbb{1}_{[0,1]}(x).$$

L'histoire du jeu de bandit peut donc être résumée par une table $(A^a, B^a)_{a=1}^K \in (\mathbb{N}\times\mathbb{N})^K$ qui indique les paramètres des lois a posteriori Beta courantes associées à chaque bras. Cet état évolue dans le Processus Décisionnel de Markov dont l'espace d'état est $\mathcal{X} = (\mathbb{N}\times\mathbb{N})^K$, l'espace d'action est $\mathcal{A} = \{1,\dots,K\}$ et la transition est la suivante.

Supposons qu'à l'instant t on soit dans l'état $S_t = (A^a, B^a)_{a=1}^K$, ce qui signifie que la loi a posteriori π_a^{t-1} sur la moyenne du bras a est Beta(A^a, B^a). Si l'action $A_t = a$ est choisie, une récompense $X_t = X_{a,t}$ est tirée selon $\mathcal{B}(\mu_a)$ et l'état est mis à jour de la manière suivante :

$$\begin{aligned} A^a &\leftarrow A^a + X_t \\ B^a &\leftarrow B^a + (1 - X_t). \end{aligned}$$

et pour $i \neq a$, A^i et B^i ne changent pas. Les fonctions de transition et de récompense dans ce MDP sont connues puisqu'on a

$$\mathbb{P}_{\Pi_0}(X_t \mid S_t = (A^i, B^i)_{i=1}^K, A_t = a) = \frac{A^a}{A^a + B^a}$$

(qui est la moyenne d'une loi Beta de paramètre A^a et B^a). La théorie des MDP (voir par exemple [Sigaud and Buffet, 2008]) nous dit alors qu'il existe une politique $\phi^* : \mathcal{X}\times[0,T] \to \mathcal{A}$, indiquant quelle action choisir en fonction de l'état et de l'instant de jeu, qui maximise la somme des récompenses jusqu'à l'horizon T, $\mathbb{E}_{\Pi_0}^{\phi}\left[\sum_{t=1}^T X_t\right]$. C'est-à-dire que la stratégie d'échantillonnage $A_t = \phi^*(S_t, t)$ est solution du problème de bandit bayésien considéré. Cette politique ϕ^* est solution d'une équation de programmation dynamique, et peut dans le cas d'une horizon fini être calculée par récurrence.

Indices de Gittins et politique optimale. Pour des bandits binaires, le calcul de la politique optimale est théoriquement possible par récurrence, mais il est très couteux du fait de la taille de l'espace d'état et ne pourra être effectif que pour des petits horizons. Plus généralement, pour tout modèle de bandit paramétrique dans lequel les lois a priori sur les paramètres θ_a sont indépendantes, [Gittins, 1979] a montré que lorsqu'on cherche à maximiser la somme des récompenses actualisées, c'est-à-dire

$$\mathbb{E}_{\Pi_0}\left[\sum_{t=1}^{\infty} \alpha^{t-1} X_t\right],$$

pour un certain coefficient d'actualisation $\alpha \in]0,1[$, la solution bayésienne se réduit à une politique d'indices. Chacun de ces *indices de Gittins* $G_\alpha(\pi)$ (qui dépend d'une loi a posteriori π et du coefficient d'actualisation) peut être obtenu comme solution d'une équation de programmation dynamique dans un espace d'état réduit. La définition des indices de Gittins pour un critère actualisé peut naturellement être transposée à un critère à horizon fini, et nous introduisons les *indices de Gittins à horizon fini*, $G(\pi, n)$, qui dépendent de π, loi a posteriori sur un bras et du temps restant $n = T - t + 1$. Ces deux types d'indices peuvent être définis de la manière suivante :

$$G_\alpha(\pi) = \sup_{\tau>0} \frac{\mathbb{E}_{\theta\sim\pi}\left[\sum_{t=1}^{\tau} \alpha^{\tau-1} Y_t\right]}{\mathbb{E}_{\theta\sim\pi}\left[\sum_{t=1}^{\tau} \alpha^{\tau-1}\right]} \quad \text{et} \quad G(\pi, n) = \sup_{0<\tau\leq n} \frac{\mathbb{E}_{\theta\sim\pi}\left[\sum_{t=1}^{\tau} Y_t\right]}{\mathbb{E}_{\theta\sim\pi}\left[\tau\right]}.$$

où conditionnellement à θ, la suite (Y_t) est i.i.d. de loi ν_θ et où le supremum porte sur l'ensemble des temps d'arrêt τ par rapport aux (Y_t), bornés par n dans le second cas. $G(\pi, n)$ représente la récompense moyenne par unité de temps que l'on peut obtenir d'un bras dont la loi a posteriori courante est π, si on peut en collecter au plus n réalisations.

Au chapitre 1, nous présenterons une définition équivalente des indices de Gittins, basée sur un problème de calibration associé à un bras. Cette interprétation permettra de fournir une méthode de calcul des indices de Gittins à horizon fini pour les bandits binaires, et de voir que contrairement au cadre actualisé, lorsqu'on fixe un horizon fini T, la politique d'indices associée aux indices de Gittins à horizon fini (appelée FH-Gittins pour *Finite-Horizon Gittins algorithm*), qui choisit à l'instant t

$$A_t = \text{argmax}_a G(\pi_a^{t-1}, T - t + 1),$$

ne coïncide pas en général avec la solution bayésienne présentée plus haut.

Toutefois, nous conjecturons que cette politique d'indices est une bonne approximation de la solution bayésienne. Cette conjecture est supportée par des expériences numériques montrant que pour des horizons courts où la solution optimale peut être calculée, le risque bayésien de l'algorithme FH-Gittins est très proche de celui de la stratégie optimale. Elle est aussi étayée par des approximations obtenues pour les indices de Gittins à horizon fini qui montrent que ceux-ci sont proches des indices utilisés par une variante de l'algorithme KL-UCB, que nous appelons KL-UCB-H$^+$. KL-UCB-H$^+$ est la politique d'indices associée à

$$u_a^{H,+}(t) = \sup\left\{q \geq \hat{\mu}_a(t) : N_a(t) d\left(\hat{\mu}_a(t), q\right) \leq \log\left(\frac{T}{N_a(t)}\right) + c \log\log\left(\frac{T}{N_a(t)}\right)\right\} \tag{7}$$

et [Lai, 1987] a prouvé, en fournissant une borne inférieure asymptotique sur le risque bayésien de tout algorithme, que cet algorithme constitue une bonne approximation de la solution bayésienne, pour des grandes valeurs de T.

Aux chapitres 2 et 3 nous proposerons également des expériences numériques montrant les bonnes performances de l'algorithme FH-Gittins en terme de regret, mais nous n'avons pas pu obtenir de garanties théoriques pour justifier ce constat empirique. De plus, cet algorithme reste difficile à implémenter, et nous n'avons pu le tester que pour des horizons $T \leq 1000$. Nous nous sommes donc focalisés dans la suite sur d'autres algorithmes bayésiens, plus facile d'implémentation, pour lesquels la propriété d'optimalité asymptotique vis-à-vis du regret sera établie.

2.2 Les algorithmes Bayes-UCB et Thompson Sampling

Nous introduisons dans cette section deux algorithmes bayésiens, pour lesquels nous proposons en particulier des analyses à temps fini montrant leur optimalité asymptotique dans des modèles de bandit binaires. Bayes-UCB, qui fait l'objet du chapitre 3, est une politique d'indices basée sur des quantiles bien choisi de la distribution a posteriori. L'échantillonnage de Thompson, ou Thompson Sampling, qui fait l'objet du chapitre 4, est un algorithme randomisé qui tire un bras selon sa probabilité a posteriori d'être optimal. Cette idée a été introduite par [Thompson, 1933] dans le premier article de bandit, mais les premiers résultats théoriques proposés pour cet algorithmes datent de la fin des années 2000, et la question de son optimalité asymptotique était encore ouverte.

BAYES-UCB

Bayes-UCB est une politique d'indices basée sur le principe d'optimisme qui a conduit à toute une famille de politiques d'indices fréquentistes. Ainsi l'algorithme UCB1 de [Auer et al., 2002a] (pour des bandits à support borné) choisit à l'instant $t+1$ le bras

$$A_{t+1} = \underset{a}{\text{argmax}}\ \hat{\mu}_a(t) + \sqrt{\frac{2\log(t)}{N_a(t)}},$$

et l'indice calculé peut être vu comme le sommet d'un intervalle de confiance obtenu avec l'inégalité de Hoeffding, alors que l'indice $u_a(t)$ utilisé par KL-UCB (5) est le sommet d'un intervalle de confiance construit avec l'inégalité de Chernoff (cf. chapitre 1). On parle de principe d'optimisme car pour chaque bras, un intervalle de confiance sur la moyenne inconnue est construit, et parmi tous les modèles statistiquement possibles, on agit optimalement de le meilleur des modèles possible (celui où les moyennes de tous les bras sont égales au sommet de leur intervalle de confiance). L'algorithme Bayes-UCB est basé sur ce même principe, mais les intervalles de confiance sont remplacés par des régions de confiance bayésiennes.

Soit $\nu = (\nu_{\theta_1}, \dots, \nu_{\theta_K})$ un modèle de bandit paramétrique. On suppose que les paramètres $(\theta_a)_{1\leq a‘K}$ sont tirés indépendamment selon des lois a priori $(\pi_a^0)_{1\leq a\leq K}$. Soit π_a^t la loi a posteriori du paramètre θ_a après t instants, et soit λ_a^t l'a posteriori sur la moyenne μ_a. Si a l'instant t le bras $A_t = a$ est choisi, les distributions a posteriori sont mises à jour de la manière suivante :

$$\pi_a^t(\theta) \propto \nu_\theta(X_t)\, \pi_a^{t-1}(\theta)\,, \quad \text{et pour tout } i \neq a,\ \pi_i^t = \pi_i^{t-1}. \tag{8}$$

L'algorithme Bayes-UCB dépend d'une famille de lois a priori $\Pi_0 = (\pi_a^0)_{1\leq a\leq K}$ et d'un paramètre réel c. Pour $t = 1, \dots, K$, l'algorithme tire les bras l'un après l'autre. Puis pour $t \geq K$, le bras choisi à l'instant $t+1$ est

$$A_{t+1} = \underset{a=1\dots K}{\operatorname{argmax}}\, Q\left(1 - \frac{1}{t(\log t)^c}; \lambda_a^t\right),$$

où $Q(\alpha, \pi)$ désigne le quantile d'ordre α de la distribution π, défini par $\mathbb{P}_{X\sim\pi}(X \leq Q(\alpha,\pi)) = \alpha$. Une illustration de l'algorithme est proposée sur la figure 1, pour des bandits binaires, où on le compare à KL-UCB. Dans les deux cas, on voit que le bras optimal est tiré la plupart de temps, conduisant à un intervalle de confiance resserré pour la moyenne de ce bras, ou à une loi a posteriori concentrée.

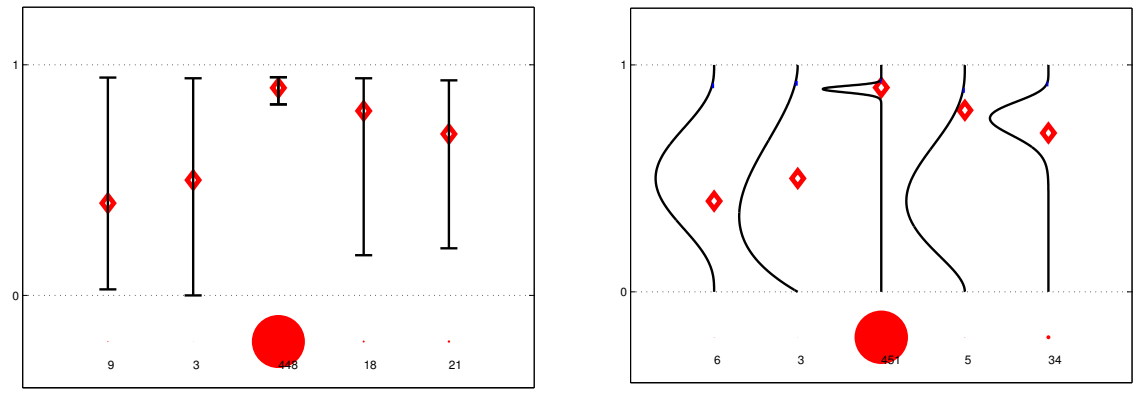

FIGURE 1 – Les intervalles de confiance utilisés par KL-LUCB (à gauche) et les lois a posteriori sur les moyennes utilisées par Bayes-UCB (à droite) après $T = 500$ instants, pour un modèle de bandit binaire à 5 bras (les losanges rouges représentent leurs moyennes).

Bayes-UCB peut être implémenté dans des classes de modèles de bandit exponentiels (définis par (1)) où les lois a posteriori sur les moyennes des bras ont des expressions explicites, sous réserve de choisir un a priori conjugué. Dans l'exemple des bandits binaires, paramétrés par leur moyenne, avec un a priori uniforme sur les moyenne, on a $\pi_a^t = \lambda_a^t = \text{Beta}(S_a(t)+1, N_a(t) - S_a(t) + 1)$ et l'algorithme

Bayes-UCB s'écrit

$$A_{t+1} = \underset{a=1\ldots K}{\operatorname{argmax}}\, Q\left(1 - \frac{1}{t(\log t)^c}; \text{Beta}(S_a(t)+1, N_a(t) - S_a(t)+1)\right).$$

Le résultat suivant, basé sur un encadrement précis de la queue d'une loi Beta, montre une connexion forte avec l'algorithme KL-UCB présenté en section 1.1. Bayes-UCB semble construire de manière automatique des intervalles de confiance basés sur la divergence de Kullback-Leibler.

Lemme 2. *Soit $d(x,y)$ la divergence de Kullback-Leibler entre deux lois de Bernoulli de paramètre x et y. La quantile d'a posteriori $q_a(t)$ utilisé par l'algorithme Bayes-UCB de paramètre c vérifie*

$$\tilde{u}_a(t) \leq q_a(t) \leq u_a(t)\ ,$$

avec

$$u_a(t) = \underset{x > \frac{S_a(t)}{N_a(t)}}{\operatorname{argmax}} \left\{ d\left(\frac{S_a(t)}{N_a(t)}, x\right) \leq \frac{\log(t) + c\log(\log(t))}{N_a(t)} \right\},$$

$$\tilde{u}_a(t) = \underset{x > \frac{S_a(t)}{N_a(t)+1}}{\operatorname{argmax}} \left\{ d\left(\frac{S_a(t)}{N_a(t)+1}, x\right) \leq \frac{\log\left(\frac{t}{N_a(t)+2}\right) + c\log(\log(t))}{(N_a(t)+1)} \right\}.$$

Le chapitre 2 montre d'autres situations où Bayes-UCB présente des similarités avec des algorithmes fréquentistes existants. Par exemple, lorsque les distributions des bras sont gaussiennes de moyenne et de variance inconnues, on retrouve un algorithme proche de UCB1-norm proposé par [Auer et al., 2002a], et plus efficace en pratique.

Dans le cadre des bandits binaires, nous avons pu montrer le résultat suivant. Le fait qu'il soit vrai pour tout $\epsilon > 0$ indique qu'on a bien

$$\limsup_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)]}{\log(T)} \leq \frac{1}{d(\mu_2,\mu_1)},$$

et donc que la borne inférieure de Lai et Robbins est atteinte.

Théorème 3. *L'algorithme Bayes-UCB avec un a priori uniforme sur les moyennes et un paramètre $c = 5$ vérifie, pour tout $\epsilon > 0$ et tout T tel que*

$$\log T + 5\log\log T \geq \frac{d(\mu_2,\mu_1)}{1+\epsilon} \exp\left(\frac{8}{(\mu_1(1-\mu_1))^2} \frac{(1+\epsilon)^2}{\epsilon^2 d(\mu_2,\mu_1)^2}\right),$$

$$\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)] \leq \frac{1+\epsilon}{d(\mu_a,\mu_1)}\log(T) + \sqrt{\log T + 5\log\log T}\sqrt{\frac{2\pi(1+\epsilon)^3 d'(\mu_a,\mu_1)^2}{d(\mu_a,\mu_1)^3}}$$
$$+ \left(\frac{1+\epsilon}{d(\mu_a,\mu_1)} + \frac{2e+3}{1-\mu_1}\right)\log\log T + 27 + 2(1+\epsilon)^2\left(\frac{d'(\mu_a,\mu_1)}{d(\mu_a,\mu_1)}\right)^2.$$

Ce théorème repose sur le lien avec des indices proches de ceux utilisés dans KL-UCB, ce qui permet d'adapter l'analyse en temps fini proposée par [Cappé et al., 2013]. Afin de comprendre quels outils mathématiques sont au cœur de la preuve, nous en présentons une esquisse.

Supposons sans perte de généralité que le bras 1 est un bras optimal, et soit a un bras sous-optimal. Du fait de la phase d'initialisation, nous avons

$$\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)] = 1 + \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{t=K}^{T-1} \mathbb{1}_{(A_{t+1}=a)}\right].$$

L'événement $(A_{t+1} = a)$ peut ensuite être décomposé de la manière suivante, en fonction de la position de la moyenne du bras optimal μ_1 par rapport à son indice $q_1(t)$ et en utilisant le fait que si a est tiré à l'instant $t+1$, on a $q_a(t) > q_1(t)$:

$$\begin{aligned}(A_{t+1} = a) &\subseteq (\mu_1 \geq q_1(t), A_{t+1}) \bigcup (\mu_1 \leq q_1(t), A_{t+1} = a) \\ &\subseteq (\mu_1 \geq q_1(t)) \bigcup (\mu_1 \leq q_a(t), A_{t+1} = a)\,.\end{aligned}$$

Cette décomposition est celle utilisée par [Cappé et al., 2013], mais pour l'analyse de Bayes-UCB nous remplacerons μ_1 par $\mu_1 - g_t$ où $g_t = \sqrt{2/\log t}$. Nous obtenons

$$\begin{aligned}(A_{t+1} = a) &\subseteq (\mu_1 - g_t \geq q_1(t)) \bigcup (\mu_1 - g_t \leq q_a(t), A_{t+1} = a) \\ &\subseteq (\mu_1 - g_t \geq \tilde{u}_1(t)) \bigcup (\mu_1 - g_t \leq u_a(t), A_{t+1} = a)\,,\end{aligned}$$

en utilisant les notations et résultats du Lemme 2. Finalement, on obtient

$$\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)] \leq 1 + \underbrace{\sum_{t=K}^{T-1} \mathbb{P}_{\boldsymbol{\theta}}\left(\mu_1 - g_t \geq \tilde{u}_1(t)\right)}_{A} + \underbrace{\sum_{t=K}^{T-1} \mathbb{P}_{\boldsymbol{\theta}}\left(\mu_1 - g_t \leq u_a(t), A_{t+1} = a\right)}_{B}.$$

Le terme A est négligeable devant $\log(T)$, car il est peu probable que l'indice $\tilde{u}_1(t)$, qui est une borne supérieure sur μ_1 soit plus petit que $\mu_1 - g_t$. Pour montrer ceci, nous devons majorer la probabilité

$$\mathbb{P}_{\boldsymbol{\theta}}\left(\mu_1 - g_t \geq \tilde{u}_1(t)\right) = \mathbb{P}_{\boldsymbol{\theta}}\left((N_1(t)+1)d^+\left(\frac{S_1(t)}{N_1(t)+1}, \mu_1 - g_t\right) \geq \log\left(\frac{t}{N_1(t)+2}\right) + 5\log\log t\right),$$

où $d^+(x,y) = d(x,y)\mathbb{1}_{(x<y)}$. Pour cela nous devons établir une *inégalité de déviation* dit *auto-normalisée* car le nombre d'observations $N_1(t)$ est lui même une variable aléatoire. De plus les déviations sont mesurées à l'aide de la divergence de Kullback-Leibler. [Garivier and Cappé, 2011] proposent le premier résultat de ce type (pour l'analyse de KL-UCB) qui ne résulte pas d'une borne de l'union, et nous avons pu l'adapter (grâce au terme g_t) à la présence d'un biais et au taux d'exploration $\log(t/(N_1(t)+2))$ au lieu de $\log t$.

Une réécriture classique (un peu astucieuse) du terme B permet de le débarrasser des quantités autonormalisées et en introduisant $\hat{\mu}_{a,s}$ la moyenne empirique des s premières observations du bras a, nous avons

$$(B) \leq \sum_{s=1}^{T} \mathbb{P}_{\boldsymbol{\theta}}\left(s d^+\left(\hat{\mu}_{a,s}, \mu_1 - g_s\right) \leq \log T + 5\log\log T\right).$$

Le nombre d'échantillons de a nécessaires pour que la moyenne $\mu_1 > \mu_a$ ne soit plus dans un intervalle autour de $\hat{\mu}_{a,s}$ construit à l'aide de la divergence de KL est environ $\log(T)/d(\mu_a, \mu_1)$ et on pourra montrer que $(B) = (1+\epsilon)\log(T)/d(\mu_a,\mu_1) + o(\log(T))$, ce qui justifie l'ordre de grandeur de la borne obtenue dans le théorème 3, dans laquelle nous tâchons d'expliciter les termes de second ordre.

THOMPSON SAMPLING

Soit toujours $\nu = (\nu_{\theta_1}, \dots, \nu_{\theta_K})$ un modèle de bandit paramétrique tel que les lois a priori sur chacun des paramètres, $(\pi_a^0)_{1\le a\le K}$, sont indépendantes. On note $\mu(\theta)$ la moyenne d'un bras paramétré par θ. L'échantillonnage de Thompson (ou Thompson Sampling) consiste à tirer à chaque instant un échantillon des loi a posteriori courantes π_a^t, et à choisir le bras ayant conduit à l'échantillon correspondant à la moyenne la plus élevée. Plus précisément, à l'instant $t+1$,

$$\begin{aligned} \forall a = 1 \dots K, \qquad & \theta_a(t) \sim \pi_a^t \\ A_{t+1} \quad = \quad & \operatorname*{argmax}_a \mu(\theta_a(t)) \end{aligned}$$

Cet algorithme peut toujours être interprété comme une politique d'indice, mais l'indice calculé pour chaque bras dépend des observations passées de ce bras *et* d'une randomisation externe. En particulier, l'indice obtenu n'est pas 'optimiste' au sens précédent, car ce n'est plus une borne de confiance supérieure pour la moyenne μ_a : avec probabilité de l'ordre de un demi, il est même plus petit que la moyenne a posteriori. Thompson Sampling implémente un optimisme un peu différent, qui consiste a tirer un modèle selon l'a posteriori courant, et à agir de manière optimale dans ce modèle échantillonné, ce qui correspond à tirer les bras selon leur probabilité a posteriori d'être optimal. Ce principe simple peut aussi être implémenté dans des modèles de bandit plus complexes comme on le verra.

La première borne supérieure logarithmique sur le regret de cet algorithme est donnée par [Agrawal and Goyal, 2012] pour les bandits binaires avec une loi a priori uniforme π_U sur les moyennes. Mais leur résultat ne fait pas intervenir les divergences de Kullback-Leibler entre les bras $d(\mu_a, \mu_1)$ et ne permet pas de monter l'optimalité asymptotique de l'échantillonnage de Thompson. Dans le même contexte, le théorème suivant que nous prouvons au chapitre 3 permet de montrer que Thompson Sampling est asymptotiquement optimal au sens de la borne de Lai et Robbins.

Théorème 4. *Soit $\epsilon > 0$ et soient b et C_b les constantes définies dans la Proposition 5 ci-dessous. Pour tout bras sous-optimal a, il existe des constantes $N(b)$ et $N(\epsilon, \mu_1, \mu_a)$ telles que pour $T \ge N(\epsilon, \mu_1, \mu_a)$,*

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}[N_a(T)] \le (1+\epsilon)\frac{\log T}{d(\mu_a,\mu_1)} + \sqrt{\log(T)}\sqrt{\frac{2\pi(1+\epsilon)^3 d'(\mu_a,\mu_1)^2}{d(\mu_a,\mu_1)^3}} \\ + \; 2(1+\epsilon)^2\left(\frac{d'(\mu_a,\mu_1)}{d(\mu_a,\mu_1)}\right)^2 + 5 + 2C_b + N(b). \end{aligned}$$

L'analyse à temps fini que nous proposons est assez proche des analyses que l'on peut proposer pour des politiques d'indices optimistes. Afin de pouvoir adapter de telles analyses, nous avons besoin du résultat suivant, qui montre que le bras optimal est souvent tiré par l'algorithme Thompson Sampling.

Proposition 5. *Il existe des constantes $b = b(\mu_1, \mu_2) \in (0,1)$ et $C_b < \infty$ telles que*

$$\sum_{t=1}^{\infty} \mathbb{P}_{\boldsymbol{\theta}}\left(N_1(t) \le t^b\right) \le C_b.$$

Les éléments clés de notre analyse. Comme notre analyse de Bayes-UCB, celle proposée pour Thompson Sampling n'est valable que pour des bandit binaires puisqu'elle s'appuie sur une propriété spécifique

des lois Beta à coefficients entiers. En effet, si on note $F^{\text{Beta}}_{a,b}$ la fonction de répartition d'une loi Beta(a,b) et $F^{\text{Bin}}_{j,\mu}$ la fonction de répartition d'une loi binomiale de paramètres j et μ, on a

$$F^{\text{Beta}}_{a,b}(y) = 1 - F^{\text{Bin}}_{a+b-1,y}(a-1).$$

Ce résultat peut être établi en remarquant que la a-ème statistique d'ordre parmi $a+b-1$ variables aléatoires uniformes suit une loi Beta(a,b).

Cette propriété avait déjà été utilisée pour établir un lien entre les quantiles d'a posteriori et les sommets d'intervalles de confiance basés sur la divergence de Kullback-Leibler (Lemme 2). Ce lien sera à nouveau être utile dans notre analyse de Thompson Sampling, où nous introduisons des quantiles d'a posteriori bien choisi, et utilisons le fait qu'avec forte probabilité, les échantillons $\theta_a(t)$ utilisés par l'algorithme sont inférieurs à ces quantiles. En couplant cette remarque à une nouvelle décomposition de l'événement $(A_{t+1} = a)$, on pourra avec l'aide de la Proposition 5 proposer une analyse proche de celle de Bayes-UCB. La preuve de cette dernière est plus complexe et utilise pleinement la nature randomisée de l'échantillonnage de Thompson.

Une généralisation. Au chapitre 3, nous présentons également une preuve de l'optimalité asymptotique de l'algorithme de Thompson pour des modèles de bandit exponentiels avec un choix particulier de distribution a priori : l'a priori de Jeffreys.

COMPARAISON NUMÉRIQUE

Nous proposons au chapitre 3 des expériences numériques comparant les performances de Bayes-UCB et de Thompson Sampling à celles d'autres algorithmes 'fréquentistes' de l'état de l'art, qui sont introduits avec plus de précisions au chapitre 1. La figure 2 présente une estimation de la distributions du regret cumulé des différents algorithmes obtenu à l'aide de $N = 50000$ répétitions d'un jeu de bandit jusqu'à un horizon $T = 20000$, pour une modèle de bandit binaire à 10 bras de moyennes $\mu = [0.1\ 0.05\ 0.05\ 0.05\ 0.02\ 0.02\ 0.02\ 0.01\ 0.01\ 0.01]$. Les quatre premiers algorithmes (dont le regret est représenté sur la même échelle) sont des variantes de l'algorithme UCB1 : UCB-Tuned ([Auer et al., 2002a]) et UCB-V ([Audibert et al., 2009]) utilisent des intervalles de confiances incorporant la variance empirique des distributions des bras, alors que MOSS ([Audibert and Bubeck, 2010]) remplace le $\log(t)$ dans l'indice UCB par $\log(t/(KN_a(t)))$. Ces quatre algorithmes ne sont pas asymptotiquement optimaux et on constate en effet un écart avec la borne inférieure de Lai et Robbins (en bleu). Les 6 algorithmes suivants (dont le regret est présenté sur la même échelle, différente de la précédente) possèdent au contraire la propriété d'optimalité asymptotique pour les bandits binaires. Des variantes de l'algorithme KL-UCB de [Cappé et al., 2013], KL-UCB-H$^+$ utilisant les indices (7), et KL-UCB+ où le $\log(T/N_a(t))$ dans (7) est remplacé par $\log(t/N_a(t))$, sont présentées, ainsi que l'algorithme DMED de [Honda and Takemura, 2010]. On peut constater que Thompson Sampling atteint les même performances que les meilleurs algorithmes fréquentistes, et que Bayes-UCB a des performances similaires à celles de KL-UCB. Par ailleurs, l'utilisation de ces deux algorithmes bayésiens présente également un avantage computationnel, car il est plus facile de calculer un quantile ou de produire un échantillon d'une loi Beta que de calculer l'indice qui intervient dans KL-UCB.

Nous proposons aussi au chapitre 3 des expériences pour une horizon plus court où l'on compare ces algorithmes asymptotiquement optimaux à l'algorithme FH-Gittins, montrant que les performances de ce dernier (quoique grossièrement comparables) sont plus variables en fonctions des modèles de bandit choisis. D'autres expériences où l'on estime le risque bayésien des algorithmes montrent que Thompson Sampling et Bayes-UCB semblent aussi asymptotiquement optimaux vis-à-vis de ce dernier.

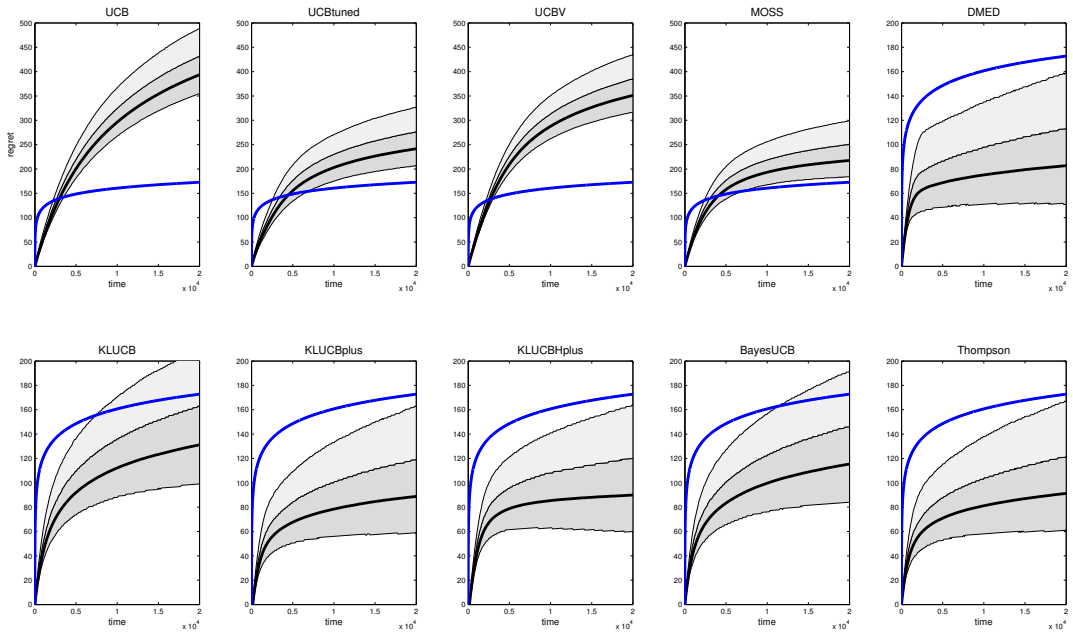

FIGURE 2 – Regret cumulé des divers algorithmes en fonction du temps. Sur chaque graphique, la courbe en bleu présente la borne inférieure, la courbe en gras présente le regret moyen, et les zones grisées claires et sombres correspondent respectivement aux 0.05% supérieurs et aux 99% centraux

2.3 Des algorithmes bayésiens pour des modèles plus généraux

Nous avons proposé des garanties théoriques pour Bayes-UCB et Thompson Sampling dans les modèles de bandit les plus simples, où les bras sont notamment indépendants. Si nous revenons à l'exemple de la recommandation de contenu (par exemple de publicité), de l'information contextuelle est disponible, et les réponses d'un utilisateur à deux publicités similaires seront sans doute très corrélées. Pour tenir compte de cette structure des actions, on peut considérer des *modèles de bandit contextuels*, qui sont présentés plus en détails au chapitre 4. Nous montrons ici que Bayes-UCB et Thompson Sampling peuvent être facilement appliqués dans ces modèles plus généraux, et nous donnons des bornes supérieures sur le regret (sous le modèle bayésien) nouvelles par rapport à la littérature, pour un modèle linéaire.

Nous nous intéressons au *modèle de bandit contextuel linéaire* suivant, où la notion d'action est remplacée par celle de contexte (ou d'action contexualisée). A chaque instant t, un ensemble $\mathcal{D}_t \subseteq \mathbb{R}^d$ de contextes est présenté à l'agent. Il doit choisir un contexte $x_t \in \mathcal{D}_t$ et reçoit la récompense

$$y_t = x_t^T \theta + \epsilon_t,$$

où ϵ_t est un bruit centré, $\theta \in \mathbb{R}^d$ est un paramètre de régression et x^T désigne la transposée du vecteur x. Revenant à l'exemple de la publicité en ligne, on peut imaginer que pour chaque utilisateur un vecteur de caractéristiques est disponible (provenant de son historique de navigation et des données disponibles sur lui). De même, on dispose de caractéristiques pour chaque publicité qu'on pourrait lui présenter. Pour chaque publicité, un vecteur de caractéristiques conjointes de la paire (utilisateur/publicité) peut être formé, ce qui donne l'ensemble de contextes disponibles. L'agent (le gestionnaire du site) choisit alors

la publicité à présenter et le modèle suppose une dépendance linéaire entre cette réponse et le contexte associé. L'hypothèse d'un ensemble $\mathcal{D}_t$ de contextes changeant à chaque instant correspond donc au fait qu'un utilisateur différent arrive à chaque fois, et aussi au fait que différentes publicités peuvent entrer ou sortir de la campagne. Une autre exploitation possible du modèle est la suivante : on suppose que les utilisateurs qui arrivent sont tous de même type (si une classification a été effectuée au préalable). Dans ce cas $\mathcal{D}_t$ est l'ensemble des vecteurs de caractéristiques des différentes publicités, et le vecteur θ indique les préférences de ce type d'utilisateur.

L'objectif de l'agent est de maximiser ses récompenses. De manière (quasiment) équivalente, il peut chercher à minimiser la quantité suivante appelée pseudo-regret (car contrairement au regret, qui est une espérance, cette quantité est aléatoire)

$$\mathcal{R}_\theta(T, \mathcal{A}) = \sum_{t=1}^{T} \left[(x_t^*)^T \theta - x_t^T \theta \right], \quad \text{où} \quad x_t^* = \underset{x \in \mathcal{D}_t}{\operatorname{argmax}}\, x^T \theta.$$

Comme précédemment, le paramètre de régression θ peut être vu comme un paramètre inconnu ou bien on peut faire l'hypothèse qu'il est tiré sous une loi a priori π_0 sur $\mathbb{R}^d$. Comme dans le cadre précédent, on peut définir des algorithmes utilisant des outils bayésiens et fréquentistes et s'intéresser aux performances de ces algorithmes sous le modèle fréquentiste (on notera $\mathbb{P}_\theta$ la probabilité associée) ou sous le modèle bayésien (on notera $\mathbb{P}$ la probabilité associée, qui dépend implicitement de l'a priori π_0). Le modèle bayésien que nous considérons ici fait l'hypothèse d'un a priori et d'un bruit gaussiens :

$$y_t = \theta^T x_t + \epsilon_t, \quad \text{avec} \quad \theta \sim \mathcal{N}\left(0, \kappa^2 I_d\right) \quad \text{et} \quad \epsilon_t \sim \mathcal{N}\left(0, \sigma^2\right). \tag{9}$$

Sous ces hypothèses, en introduisant

$$X_t = \begin{pmatrix} x_1^T \\ x_2^T \\ . \\ x_t^T \end{pmatrix} \in \mathcal{M}_{t,d}(\mathbb{R}), \quad Y_t = \begin{pmatrix} y_1 \\ y_2 \\ . \\ y_t \end{pmatrix} \in \mathbb{R}^t, \quad \text{et} \quad E_t = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ \epsilon_t \end{pmatrix} \in \mathbb{R}^t,$$

la loi a posteriori sur θ après t observations est gaussienne de moyenne $\hat{\theta}(t)$ et de covariance Σ_t, où

$$\begin{cases} \hat{\theta}(t) &= (B(t))^{-1} X_t^T Y_t \quad \text{avec} \quad B(t) = \frac{\sigma^2}{\kappa^2} I_d + X_t^T X_t \\ \Sigma_t &= \sigma^2 (B(t))^{-1}. \end{cases}$$

Les algorithmes. L'algorithme Bayes-UCB tel que nous l'avons présenté pour les bandits à bras indépendants peut être utilisé ici, en calculant pour chaque contexte $x \in \mathcal{D}_t$ un quantile de la loi a posteriori sur la moyenne associée, $x^T\theta$, qui a pour loi $\mathcal{N}\left(x^T\hat{\theta}(t), \|x\|_{\Sigma_t}\right)$, où $\|x\|_A = \sqrt{x^T A x}$. Si l'algorithme original utilise un quantile d'ordre $1 - 1/t$, nous le définissons ici en fonction d'un *taux d'exploration* $f(t, \delta)$. Bayes-UCB choisit à l'instant $t+1$ le contexte

$$x_{t+1} = \underset{x \in \mathcal{D}_{t+1}}{\operatorname{argmax}} \left[x^T \hat{\theta}(t) + \|x\|_{\Sigma_t} Q\left(1 - e^{-f(t+1,\delta)}; \mathcal{N}(0,1)\right) \right].$$

Les modèles de bandit linéaires tels que ceux que nous considérons ont été largement étudiés dans la littérature et le principe d'optimisme est appliqué de la manière suivante dans ces modèles. Etant donnée une région de confiance $C_t \subseteq \mathbb{R}^d$ pour le paramètre de régression, le contexte choisi est celui dont le produit scalaire avec un des paramètres de régression jugés possibles (dans C_t) est maximal :

$$x_{t+1} = \underset{x \in \mathcal{D}_{t+1}}{\operatorname{argmax}} \max_{\theta' \in C_t} x^T \theta'.$$

Dans les articles de [Auer, 2002, Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010, Chu et al., 2011, Abbasi-Yadkori et al., 2011], les régions de confiances qui sont construites sont de la forme $C_t = \left(\theta' \in \mathbb{R}^d : ||\hat{\theta}(t) - \theta'||_{\Sigma_t^{-1}} \leq \beta(t+1,\delta)\right)$. Ce sont des régions de confiance 'fréquentistes', c'est-à-dire vraies en forte probabilité sous $\mathbb{P}_\theta$, pour tout paramètre θ. $\hat{\theta}(t)$ ne s'interprète plus alors comme la moyenne de la loi loi a posteriori mais comme un estimateur des moindres carrés régularisé, où κ est un paramètre de régularisation. Les algorithmes associés se réécrivent

$$x_{t+1} = \underset{x \in \mathcal{D}_{t+1}}{\text{argmax}} \left[\hat{\theta}(t)^T x + ||x||_{\Sigma_t} \beta(t+1,\delta)\right]. \tag{10}$$

On peut noter que Bayes-UCB prend cette forme. Il est possible de définir une variante de cet algorithme, qui implémente le principe d'optimisme pour les bandits linéaire tel qu'introduit ci-dessus de manière bayésienne. En effet, une région de confiance pour θ sous le modèle bayésien (9) peut être construite de la manière suivante :

$$\mathbb{P}\left(||\theta - \hat{\theta}(t)||_{\Sigma_t^{-1}} \leq \sqrt{Q\left(1 - e^{-f(t+1,\delta)}; \chi_d^2\right)}\right) \geq 1 - e^{-f(t+1,\delta)}, \tag{11}$$

où un quantile d'une loi du chi-deux à d degrés de liberté apparait, car conditionnellement au t premières observations, $||\theta - \hat{\theta}(t)||^2_{\Sigma_t^{-1}} \sim \chi_d^2$. Nous définissons alors l'algorithme Bayes-LinUCB comme l'algorithme optimiste associé à cette région de confiance, qui choisit à l'instant $t+1$ le contexte

$$x_{t+1} = \underset{x \in \mathcal{D}_{t+1}}{\text{argmax}} \left[\hat{\theta}(t)^T x + ||x||_{\Sigma_t} \sqrt{Q\left(1 - e^{-f(t+1,\delta)}; \chi_d^2\right)}\right].$$

L'échantillonnage de Thompson peut également être facilement implémenté dans des modèles de bandit contextuels linéaires. Les bonnes performances de Thompson Sampling dans des modèles plus complexes, les modèles logistiques, avaient d'ailleurs été constatées en pratique avant que des garanties théoriques n'émergent pour les bandits binaires ([Scott, 2010, Chapelle and Li, 2011]). Dans notre modèle linéaire, à l'instant $t+1$, un échantillon $\tilde{\theta}(t)$ de la loi a posteriori (gaussienne) sur θ est tiré et le contexte choisi est

$$x_{t+1} = \text{argmax}_{x \in \mathcal{D}_{t+1}} \, x^T \tilde{\theta}(t).$$

Nos résultats. Nous montrons le résultat suivant pour les algorithmes Bayes-UCB et Bayes-LinUCB.

Théorème 6. *Avec le taux d'exploration $f(t,\delta) = \log \frac{K\pi^2 t^2}{3\delta}$, si pour tout t, $|\mathcal{D}_t| = K$ et si les contextes sont bornés par L, l'algorithme Bayes-UCB vérifie, sous le modèle (9),*

$$\mathbb{P}\left(\forall T \in \mathbb{N}, \mathcal{R}_\theta(T,\mathcal{A}) \leq \sqrt{Td}\sqrt{2C_1 \log\left(\frac{K\pi^2 T^2}{6\delta}\right)\log\left(1 + T\frac{L^2\sigma^2}{d\kappa^2}\right)}\right) \geq 1 - \delta.$$

Avec le taux d'exploration $f(t,\delta) = \log \frac{\pi^2 t^2}{6\delta}$, l'algorithme Bayes-LinUCB vérifie, sous le modèle (9) et avec des contextes bornés par L,

$$\mathbb{P}\left(\forall T \in \mathbb{N}, \mathcal{R}_\theta(T,\mathcal{A}) \leq d\sqrt{T}\sqrt{C_1 \log\left(1 + T\frac{L^2\sigma^2}{d\kappa^2}\right)\left(1 + \frac{2}{d}\log\frac{\pi^2 T^2}{6\delta} + 2\sqrt{\frac{1}{d}\log\frac{\pi^2 T^2}{6\delta}}\right)}\right) \geq 1 - \delta.$$

où $C_1 := \frac{4L^2\kappa^2}{\log(1+L^2\kappa^2\sigma^{-2})}$.

On a donc montré que le pseudo-regret de Bayes-UCB est en forte probabilité (sous le modèle bayésien) de l'ordre de $\tilde{O}(\sqrt{dT\log(K)})$ (où la notation $\tilde{O}$ ignore les facteurs logarithmiques en T) et celui de Bayes-LinUCB de l'ordre de $\tilde{O}(d\sqrt{T})$. Le pseudo-regret de l'algorithme optimiste qui utilise la région de confiance fréquentiste la plus fine, l'algorithme OFUL proposé par [Abbasi-Yadkori et al., 2011], est de l'ordre de $O(d\sqrt{T})$ en forte probabilité sous $\mathbb{P}_\theta$, pour tout θ. En revanche, on ne trouve pas dans l'état de l'art fréquentiste un résultat spécifique dans le cas où le nombre de contexte est fini et la dépendance en $\sqrt{dT\log(K)}$ obtenue pour Bayes-UCB dans ce cas est dans un certain sens optimale.

Pour Thompson Sampling, on peut déduire de l'analyse bayésienne de [Russo and Van Roy, 2014] que dans le cas où le nombre de contextes est fini, le risque bayésien, défini ici par $\mathrm{BR}_{\pi_0}(T,\mathcal{A}) = \mathbb{E}\left[\mathcal{R}_\theta(T,\mathcal{A})\right]$ de Thompson Sampling est de l'ordre de $\tilde{O}(\sqrt{dT\log(K)})$. Nous donnons également une borne générale (indépendante du nombre de contextes) qui montre que le risque bayésien de Thompson Sampling est de l'ordre de $\tilde{O}(d\sqrt{T})$.

Ainsi, en fonction de la relation entre la dimension d et le nombre d'actions K, une des deux variantes Bayes-UCB ou Bayes-Lin-UCB sera préférable ; celle qui conduit à la borne supérieure la plus petite sur le regret, ou aux meilleures performances pratiques. A l'inverse, l'échantillonnage de Thompson s'implémente d'une manière unique pour tous les modèles linéaires et son risque bayésien est majoré par la plus petite des deux bornes en $\tilde{O}(\sqrt{dT\log(K)})$ et $\tilde{O}(d\sqrt{T})$ obtenues.

3 Vers des algorithmes fréquentistes optimaux pour l'identification des meilleurs bras

On rappelle que contrairement à l'objectif de minimisation du regret, pour l'objectif d'identification des meilleurs bras, la notion d'algorithme (asymptotiquement) optimal n'existe pas dans la littérature. Nous avons proposé les deux notions de complexités suivantes pour un budget fixé ($\kappa_B(\nu)$) ou un niveau de confiance fixé ($\kappa_C(\nu)$) :

$$\kappa_{\mathrm{B}}(\nu) = \inf_{\mathcal{A}\ \text{consistant}} \left(\limsup_{t\to\infty} -\frac{1}{t}\log p_t(\nu) \right)^{-1}, \qquad \kappa_{\mathrm{C}}(\nu) = \inf_{\mathcal{A}\ \delta\text{–PAC}} \limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\log\frac{1}{\delta}}.$$

Afin d'évaluer ces deux complexités et de pouvoir les calculer, deux types de résultats sont nécessaires. Nous donnons des bornes inférieures sur le nombre moyen d'observations nécessaires à un algorithme pour identifier les m meilleurs bras avec une probabilité plus grande que $1-\delta$, ou sur la probabilité d'erreur d'un algorithme qui peut utiliser t observations des bras. Ensuite, nous présentons des algorithmes qui atteignent ces bornes inférieures. De tels algorithmes peuvent alors être qualifiés d'asymptotiquement optimaux.

La section 3.1 présente nos outils pour obtenir des bornes inférieures, ainsi qu'une première borne inférieure sur $\kappa_C(\nu)$, valable pour des distributions générales et pour toute valeurs de $m \geq 1$ (ce qui n'existait pas dans la littérature). Nous donnons ensuite en section 3.2 deux algorithmes pour l'identification des meilleurs bras à niveau de confiance fixé, basés sur des intervalles de confiances construits à l'aide de la divergence de Kullback-Leibler. La borne supérieure obtenue sur la moyenne du nombre d'échantillons utilisés par KL-UCB fait aussi intervenir des quantités informationnelle, mais n'atteint pas exactement la borne inférieure précédente. Dans le cas particulier de l'identification du meilleur bras parmi deux, nous proposons en section 3.3 de nouvelles bornes inférieures ainsi que des algorithmes asymptotiquement optimaux.

3.1 Une borne inférieure sur la complexité à niveau de confiance fixé

Toutes les bornes inférieure obtenues dans la littérature, que ce soit pour la minimisation du regret ([Lai and Robbins, 1985]) ou l'identification des meilleurs bras ([Mannor and Tsitsiklis, 2004, Audibert et al., 2010]) reposent sur des *changements de loi*. Un changement de loi relie la probabilité du même événement sous deux modèles de bandit différents, ν et ν'. Le lemme ci-dessous présente une nouvelle formulation synthétique pour un changement de loi, sous la forme d'une inégalité qui fait directement intervenir l'espérance du nombre de tirages de chaque bras.

Lemme 7. *Soient ν et ν' deux modèles de bandit tels que les distributions des tous les bras de ν et ν' soient absolument continues. Soit σ un temps d'arrêt par rapport à la filtration $(\mathcal{F}_t)$ et soit $A \in \mathcal{F}_\sigma$ tel que $0 < \mathbb{P}_\nu(A) < 1$. On a*

$$\sum_{a=1}^{K} \mathbb{E}_\nu[N_a(\sigma)]\mathrm{KL}(\nu_a, \nu'_a) \geq d(\mathbb{P}_\nu(A), \mathbb{P}_{\nu'}(A)),$$

où $d(x,y) := x\log(x/y) + (1-x)\log((1-x)/(1-y))$ désigne l'entropie relative binaire.

Soit $\mathcal{A}$ un algorithme δ-PAC. Pour minorer le nombre moyen d'observations $\mathbb{E}_\nu[\tau]$ pour un modèle de bandit ν fixé, on peut minorer $\mathbb{E}_\nu[N_a]$ pour chaque bras, où $N_a = N_a(\tau)$ désigne le nombre total d'observations du bras a. Une telle minoration peut être obtenue en appliquant le Lemme 7 avec le temps d'arrêt τ, en choisissant pour événement A l'événement d'erreur sous ν et pour ν' un modèle de bandit qui diffère de ν par le bras a uniquement, et qui a un ensemble de bras optimaux différent de celui de ν. Ceci permet de montrer, sous certaines hypothèses sur la classe $\mathcal{M}_m$, vérifiées par exemple pour une classe de modèles de bandit exponentiels, le résultat suivant.

Théorème 8. *Tout algorithme δ-PAC sur $\mathcal{M}_m$ vérifie, pour $\delta \leq 0.15$,*

$$\mathbb{E}_\nu[\tau] \geq \left[\sum_{a \in \mathcal{S}_m^\star} \frac{1}{\mathrm{KL}(\nu_a, \nu_{[m+1]})} + \sum_{a \notin \mathcal{S}_m^\star} \frac{1}{\mathrm{KL}(\nu_a, \nu_{[m]})} \right] \log\left(\frac{1}{2\delta}\right).$$

Notons que le Lemme 7 sera introduit dès le chapitre 1, car il permet également de donner une preuve simple de la borne inférieure de [Burnetas and Katehakis, 1996], qui généralise celle de Lai et Robbins.

3.2 Deux algorithmes : KL-LUCB et KL-Racing

Nous commençons par présenter deux algorithmes génériques pour l'identification des m meilleurs bras parmi K, pour tout $m \geq 1$, pour un niveau de confiance fixé. Ces algorithmes partagent l'utilisation d'intervalles de confiance, mais sont basés sur deux stratégies d'échantillonnage différentes. A chaque instant de jeu, indexé par t, ces algorithmes tirent entre 2 et K bras, et nous notons $\mathcal{I}_a(t) = [L_a(t), U_a(t)]$ un intervalle de confiance construit pour le bras a à l'instant t, ainsi que $N_a(t)$ (resp. $S_a(t)$) le nombre de tirages (resp. la somme des observations) du bras a entre les instants 1 et t, et $\hat{\mu}_a(t)$ la moyenne empirique.

L'algorithme Racing est basé sur un *échantillonnage uniforme couplé à des éliminations*. A chaque instant t, les bras sont partitionnés en trois ensemble : $\mathcal{R}_t$ est l'ensemble des bras restants en course, $\mathcal{S}_t$ l'ensemble des bras sélectionnés et $\mathcal{D}_t$ l'ensemble des bras éliminés. Chaque bras de $\mathcal{R}_t$ est tiré une fois, d'où le terme échantillonnage uniforme (on pose $\mathcal{R}_1 = \{1, \dots, K\}$ si bien que dans les premiers instants on tire tous les bras). Les bras de $\mathcal{R}_t$ sont ensuite triés par moyenne empirique décroissante, et si

le meilleur empirique $\hat{a}$ est tel que la borne inférieure $L_{\hat{a}}(t)$ de son intervalle de confiance est supérieure aux bornes supérieures $U_b(t)$ des $K - m - |\mathcal{D}_t|$ moins bons bras empiriques, on sélectionne $\hat{a}$ comme faisant partie des m meilleurs :

$$\mathcal{S}_{t+1} = \mathcal{S}_t \bigcup \{\hat{a}\}, \quad \mathcal{R}_{t+1} = \mathcal{R}_t \backslash \{\hat{a}\} \text{ et } \mathcal{D}_{t+1} = \mathcal{D}_t.$$

Si $\hat{a}$ n'a pas été sélectionné, on se donne aussi la possibilité d'éliminer le moins bon bras empirique $\hat{b}$ si la borne borne supérieure $U_{\hat{b}}(t)$ de son intervalle de confiance est inférieure aux bornes inférieures $L_a(t)$ des $m - |\mathcal{S}_t|$ meilleurs bras empiriques :

$$\mathcal{S}_{t+1} = \mathcal{S}_t, \quad \mathcal{R}_{t+1} = \mathcal{R}_t \backslash \{\hat{b}\} \text{ et } \mathcal{D}_{t+1} = \mathcal{D}_t \bigcup \{\hat{b}\}.$$

L'échantillonnage s'arrête au premier instant t tel que $|\mathcal{S}_t| = m$, et on recommande $\hat{S}_m = \mathcal{S}_t$. Cet algorithme est inspiré des algorithmes de la littérature basés sur des éliminations (e.g. [Jennison et al., 1982, Even-Dar et al., 2006, Heidrich-Meisner and Igel, 2009]), qui peuvent se réécrire comme l'algorithme Racing avec des intervalles de confiance particuliers.

L'algorithme LUCB est basé sur un *échantillonnage adaptatif*, et ne tire que deux bras bien choisis à chaque instant t. Il a été proposé par [Kalyanakrishnan et al., 2012] avec des intervalles de confiances basés sur l'inégalité de Hoeffding, et sa formulation générique est la suivante. A chaque instant t, l'ensemble $J(t)$ des m bras ayant les m plus grandes moyennes empiriques est formé. Les deux bras tirés sont celui parmi $J(t)$ dont l'intervalle de confiance a la plus petit borne inférieure, et celui parmi $J(t)^c$ dont l'intervalle de confiance a la plus grande borne supérieure. Ces deux bras l_t et u_t ont en effet de plus grandes chances d'être mal classifiés dans $J(t)$ et $J(t)^c$. L'échantillonnage s'arrête au premier instant t tel que $L_{l_t} > U_{u_t}$ (les intervalles de confiances des bras dans $J(t)$ et $J(t)^c$ sont séparés) et l'ensemble $J(t)$ est recommandé.

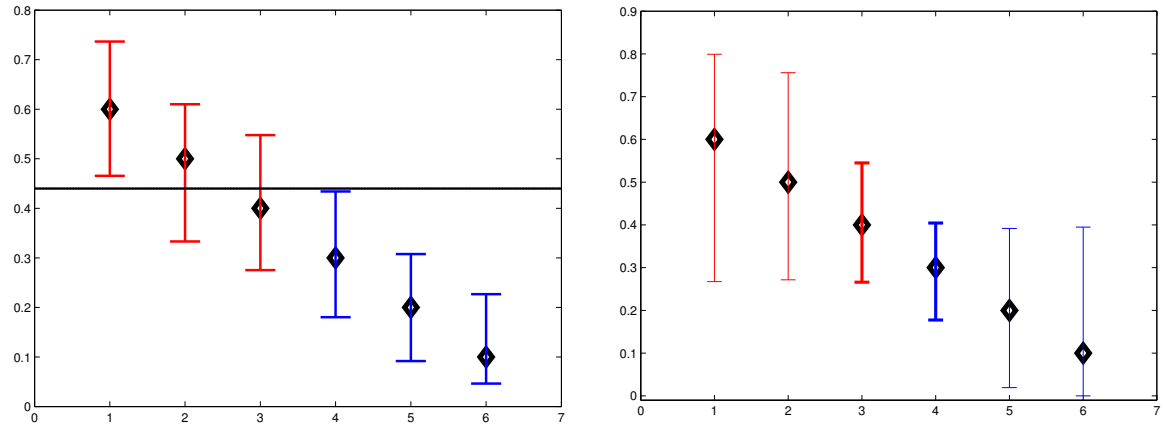

FIGURE 3 – Illustration des algorithmes pour $K = 6$, $m = 3$. Les losanges noirs représentent les moyennes inconnues. Les m meilleurs empiriques (resp. $K - m$ moins bons) sont en rouge (resp. en bleu). Les bras en gras sont ceux tirés à l'instant courant : à gauche KL-LUCB tire tous les bras, et le meilleur bras empirique va être éliminé. A droite KL-LUCB tire uniquement les bras l_t et u_t.

Les deux algorithmes décrits ci-dessus dépendent d'une famille d'intervalles de confiance. Nous analysons pour des modèles de bandit exponentiels les variantes appelées **KL-Racing** et **KL-LUCB** qui

utilisent les intervalles de confiance $\mathcal{I}_a(t) = [l_a(t)u_a(t)]$ avec

$$\begin{aligned} u_a(t) &:= \max\{q \in [\hat{\mu}_a(t), 1] : N_a(t)d(\hat{\mu}_a(t), q) \leq \beta(t,\delta)\}, \\ l_a(t) &:= \min\{q \in [0, \hat{\mu}_a(t)] : N_a(t)d(\hat{\mu}_a(t), q) \leq \beta(t,\delta)\}, \end{aligned}$$

où $d(x,y)$ est la fonction de divergence associée à la famille exponentielle (voir (2)) et $\beta(t,\delta)$ est au taux d'exploration. Une illustration de ces deux algorithmes est proposée sur la figure 3 pour des modèles de bandit binaires, où l'on a $d(x,y) = \mathrm{KL}(\mathcal{B}(x), \mathcal{B}(y))$. Une étude numérique sera présentée au chapitre 5 pour des bandit binaires, où l'on verra que comme dans le cadre de la minimisation du regret, l'utilisation de tels intervalles de confiance est préférable à celle d'intervalles proches de ceux utilisés par l'algorithme UCB1 et basés sur l'inégalité de Hoeffding. Par ailleurs, KL-LUCB semble nécessiter empiriquement moins d'observations que KL-Racing pour trouver les m meilleurs bras.

Analyse de KL-LUCB. Pour KL-LUCB nous avons obtenu une majoration de l'espérance du nombre de tirages de bras $\mathbb{E}_\nu[\tau]$ que nous présentons ici car elle conduit à une borne supérieure sur le terme de complexité $\kappa_C(\nu)$.

La présentation de notre résultat nécessite l'introduction d'une nouvelle quantité informationnelle, l'*information de Chernoff*. L'information de Chernoff entre les distributions de Bernoulli $\mathcal{B}(x)$ et $\mathcal{B}(y)$, notée $d^*(x,y)$, est définie par

$$d^*(x,y) = d(z^*, x) = d(z^*, y) \text{ où } z^* \text{ est l'unique } z \text{ tel que } d(z,x) = d(z,y). \tag{12}$$

Cette définition peut être généralisée aux modèles de bandit exponentiels.

Théorème 9. *L'algorithme KL-Racing avec* $\beta(t,\delta) = \log\left(\frac{k_1 K t^\alpha}{\delta}\right)$ *pour* $\alpha > 1$ *et* $k_1 > 1 + \frac{1}{\alpha - 1}$ *recommande les* m *meilleurs bras avec une probabilité plus grande que* $1 - \delta$.

Soit $c \in [\mu_{[m+1]}, \mu_{[m]}]$. *Si de plus on choisit* $\alpha > 2$ *dans l'expression ci-dessus du taux d'exploration, il existe une constante* C_α *telle que*

$$\mathbb{E}_\nu[\tau] \leq 4\alpha H_c^* \log\left(\frac{e k_1 K (H_c^*)^\alpha}{\delta} \log\left(\frac{k_1 K (H_c^*)^\alpha}{\delta}\right)\right) + C_\alpha,$$

avec

$$H_c^*(\nu) := \sum_{a \in \{1,\dots,K\}} \frac{1}{d^*(\mu_a, c)}.$$

En utilisant les résultats des théorèmes 8 et 9, on obtient l'encadrement suivant sur la complexité pour un niveau de confiance fixé :

$$\sum_{a \in \mathcal{S}_m^*} \frac{1}{d(\mu_a, \mu_{[m+1]})} + \sum_{a \notin \mathcal{S}_m^*} \frac{1}{d(\mu_a, \mu_{[m]})} \leq \kappa_C(\nu) \leq 8 \min_{c \in [\mu_{m+1}, \mu_m]} \sum_{a=1}^{K} \frac{1}{d^*(\mu_a, c)}.$$

Nous avons obtenu des bornes informationnelles sur le terme de complexité $\kappa_C(\nu)$, mais un écart demeure entre la borne inférieure et la borne supérieure obtenues. Le paramètre c qui apparaît dans la borne supérieure semble être un avatar de notre preuve, et on s'attendrait plutôt à une borne supérieure de la forme

$$\sum_{a \in \mathcal{S}_m^*} \frac{1}{d^*(\mu_a, \mu_{[m+1]})} + \sum_{a \notin \mathcal{S}_m^*} \frac{1}{d^*(\mu_a, \mu_{[m]})}$$

Quant à la présence de l'information de Chernoff, elle semble aussi provenir de raisons techniques (voir chapitre 5). Mais nous allons voir que cette quantité informationnelle sera en effet caractéristique de la complexité d'un problème d'identification du meilleur bras dans un modèle de bandit à deux bras.

3.3 Caractérisation de la complexité pour des modèles de bandit à deux bras

Une partie du chapitre 5 est dédiée au calcul des complexités dans des modèles de bandit à deux bras. On peut tout d'abord souligner l'intérêt pratique de tels modèles de bandit, qui fournissent un cadre théorique pour l'A/B Testing séquentiel. L'A/B Testing est une procédure utilisée par exemple pour l'optimisation des contenus web : deux versions d'une page web sont comparées en étant présentées à des utilisateurs. On présente à chaque utilisateur une seule des deux versions, $A_t \in \{1,2\}$, et l'utilisateur fournit une réponse X_t qui est un indice de la qualité de la page, modélisée comme une réalisation d'une loi de probabilité ν_1 ou ν_2. Un objectif commun est de déterminer quelle version a le plus grand taux de conversion (probabilité qu'un utilisateur devienne un consommateur) en collectant des réponses binaires des utilisateurs.

Dans des modèles de bandit à deux bras, les deux algorithmes que nous avons présentés, KL-Racing et KL-LUCB se réduisent au même algorithme, qui tire les deux bras de manière uniforme et utilisent un critère d'arrêt basé sur la séparation d'intervalles de confiances. On peut en particulier se demander si cet *échantillonnage uniforme* fait sens.

Des bornes inférieures plus fines. En utilisant le même outil qu'en section 3.1, le Lemme 7, mais on considérant des changements de lois différents (où en particulier les deux bras du modèle de bandit alternatif ν' sont modifiés par rapport au modèle ν), on peut montrer le résultat suivant. Ce résultat est présenté dans un cadre plus général au chapitre 5, mais nous l'énonçons ici pour des modèles de bandit où les bras sont paramétrés continument par leurs moyennes. Cela inclut le cas des modèles de bandit gaussiens de variances connues et le cas des modèles de bandit exponentiels.

Théorème 10. *Soit $\mathcal{M}_1$ une classe de modèles de bandit à deux bras paramétrés continument par leurs moyennes. Soit $\nu = (\nu_1, \nu_2) \in \mathcal{M}_1$. On a les résultats suivants*

Budget fixé	***Niveau de confiance fixé***
Tout algorithme consistant sur $\mathcal{M}_1$ vérifie	*Tout algorithme δ-PAC sur $\mathcal{M}_1$ vérifie, pour $\delta \le 0.15$,*
$\limsup_{t\to\infty} -\frac{1}{t}\log p_t(\nu) \le \mathrm{KL}^*(\nu_1,\nu_2)$	$\mathbb{E}_\nu[\tau] \ge \frac{1}{\mathrm{KL}_*(\nu_1,\nu_2)}\log\left(\frac{1}{2\delta}\right)$
avec $\mathrm{KL}^*(\nu_1,\nu_2) := \mathrm{KL}(\nu^*,\nu_1) = \mathrm{KL}(\nu^*,\nu_2)$.	*avec* $\mathrm{KL}_*(\nu_1,\nu_2) := \mathrm{KL}(\nu_1,\nu_*) = \mathrm{KL}(\nu_2,\nu_*)$.

On en déduit

$$\kappa_B(\nu) \ge \frac{1}{\mathrm{KL}^*(\nu_1,\nu_2)} \quad \text{et} \quad \kappa_C(\nu) \ge \frac{1}{\mathrm{KL}_*(\nu_1,\nu_2)}.$$

La quantité $\mathrm{KL}^*(\nu_1,\nu_2)$ dans la complexité à budget fixé peut être interprétée comme une information de Chernoff, alors que la quantité $\mathrm{KL}_*(\nu_1,\nu_2)$ apparait comme une quantité similaire, mais où les rôles des arguments sont inversés.

Algorithmes asymptotiquement optimaux. Dans la classe des bandits gaussiens, définie par

$$\mathcal{M}_1 = \{\nu = \left(\mathcal{N}\left(\mu_1,\sigma_1^2\right), \mathcal{N}\left(\mu_2,\sigma_2^2\right)\right) : (\mu_1,\mu_2) \in \mathbb{R}^2, \mu_1 \neq \mu_2\}$$

où les variances σ_1^2 et σ_2^2 sont connues (mais potentiellement différentes), la divergence de Kullback-Leibler est symétrique, et les deux bornes inférieures obtenues dans le théorème 10 sont égales. On

montre même que

$$\boxed{\kappa_B(\nu) = \kappa_C(\nu) = \frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}}$$

en proposant des algorithmes asymptotiquement optimaux. Pour un budget fixé, un algorithme qui échantillonne les bras proportionnellement à leurs écarts types (et non leurs variances) et recommande le meilleur empirique vérifie en effet

$$\liminf_{t\to\infty} -\frac{1}{t}\log p_t(\nu) \geq \frac{(\mu_1-\mu_2)^2}{2(\sigma_1+\sigma_2)^2},$$

et atteint donc la borne inférieure du théorème 10. Pour un niveau de confiance fixé, en couplant un échantillonnage séquentiel qui maintient la proportion de tirages du bras 1, $N_1(t)/t$, proche du ratio des écarts types $\alpha = \sigma_1/(\sigma_1+\sigma_2)$ ($A_t = 2$ ssi $\lceil \alpha t\rceil = \lceil \alpha(t-1)\rceil$) à une règle d'arrêt basée sur la différence des moyennes empiriques,

$$\tau = \inf\left\{t\in\mathbb{N} : |\hat{\mu}_1(t)-\hat{\mu}_2(t)| > \sqrt{2\sigma_t^2(\alpha)\log(t/\delta)}\right\} \quad \text{où} \quad \sigma_t^2(\alpha) = \frac{\sigma_1^2}{\lceil \alpha t\rceil} + \frac{\sigma_2^2}{(t-\lceil \alpha t\rceil)},$$

on obtient un algorithme δ-PAC qui vérifie

$$\mathbb{E}_\nu[\tau] \leq (1+\epsilon)\frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}\log\left(\frac{1}{\delta}\right) + \underset{\delta\to 0}{o_\epsilon}\left(\log\left(\frac{1}{\delta}\right)\right),$$

et atteint donc la borne du théorème 10.

Dans la classe des bandits binaires, définie par

$$\mathcal{M}_1 = \{\nu = (\mathcal{B}(\mu_1),\mathcal{B}(\mu_2)) : (\mu_1,\mu_2)\in]0;1[^2, \mu_1\neq\mu_2\},$$

les deux bornes inférieures du théorème 10 ne sont pas égales. En notant $d^*(\mu_1,\mu_2)$ l'information de Chernoff définie en (12) et $d_*(\mu_1,\mu_2)$ la quantité définie par $d_*(\mu_1,\mu_2) = d(\mu_1,\mu_*)$ où μ_* est l'unique élément vérifiant $d(\mu_1,\mu_*) = d(\mu_2,\mu_*)$, on a en effet

$$\kappa_B(\nu) \geq \frac{1}{d^*(\mu_1,\mu_2)} \quad \text{et} \quad \kappa_C(\nu) \geq \frac{1}{d_*(\mu_1,\mu_2)},$$

et nous avons $d_*(\mu_1,\mu_2) < d^*(\mu_1,\mu_2)$. Pour un budget fixé, pour chaque ν, on peut montrer qu'il existe $\alpha(\nu)$ tel qu'en allouant $\alpha(\nu)t$ échantillon au bras 1, on atteint la borne du théorème 10. Ceci permet de montrer que

$$\boxed{\kappa_B(\nu) = \frac{1}{d^*(\mu_1,\mu_2)} \quad \text{et} \quad \kappa_C(\nu) > \kappa_B(\nu).}$$

Par ailleurs, nous montrons que dans des modèles de bandit binaires, il y a peu à gagner à considérer des stratégies dont la règle d'échantillonnage n'est pas uniforme. En effet, nous avons pu obtenir des bornes inférieures sur le nombre d'échantillons (resp. la probabilité d'erreur) de stratégies échantillonnant les bras uniformément, qui sont très proches de celles du théorème 10. Pour un budget fixé, la stratégie qui tire les bras uniformément et recommande le meilleur empirique s'avère ainsi être une très bonne approximation de la règle optimale. Pour un niveau de confiance fixé, nous montrons qu'un échantillonnage uniforme couplé à une règle d'arrêt basée sur la différence des moyennes empiriques conduit à un algorithme sous-optimal, et nous proposons une règle d'arrêt plus sophistiquée.

4 Organisation du document

Dans cette thèse nous présentons des algorithmes pour deux problèmes d'allocation séquentielle de ressources : la maximisation des récompenses et l'identification des meilleurs bras dans des modèles de bandit. De bons algorithmes pour ces deux problèmes peuvent être obtenus en utilisant des intervalles de confiances basés sur la divergence de Kullback-Leibler. Nous avons transposé cette idée, proposée par [Cappé et al., 2013] pour l'objectif de maximisation des récompenses et donnant lieu à l'algorithme KL-UCB, à l'identification des meilleur bras, pour laquelle nous avons introduit l'algorithme KL-LUCB. Par ailleurs, pour l'objectif de maximisation des récompenses, nous avons proposé l'algorithme Bayes-UCB, basé sur une interprétation bayésienne du problème, et prouvé son optimalité asymptotique du point de vue du regret, une mesure de performance fréquentiste, dans des modèles de bandit binaires. Nous avons établi cette même propriété pour l'échantillonnage de Thompson, un autre algorithme bayésien, introduit en 1933 mais dont l'optimalité était encore une question ouverte.

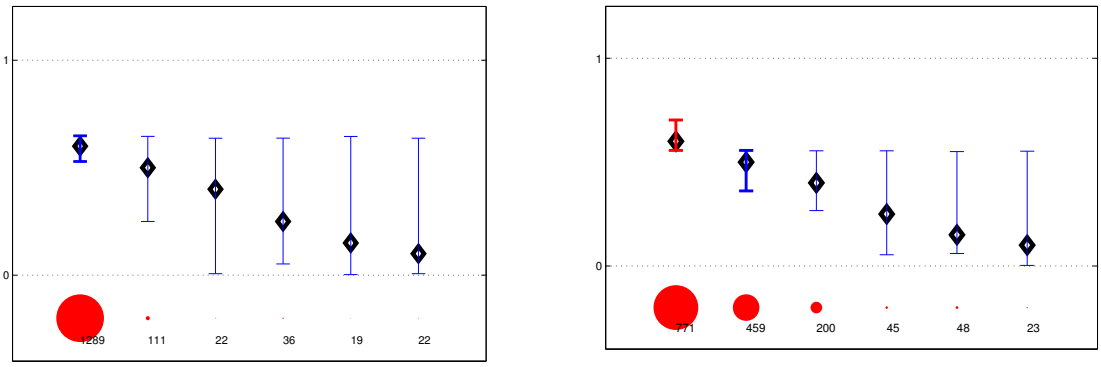

FIGURE 4 – KL-UCB (à gauche) et KL-LUCB pour l'identification du meilleur bras (à droite) après 1500 observations des bras environ.

Comme on peut le constater sur la figure 4, les algorithmes pour les deux problèmes considérés sont différents. KL-UCB tire massivement le bras optimal, et très peu les autres bras, ce qui conduit à une très bonne estimation de la moyenne du meilleur bras, mais une mauvaise estimation de celles des autres bras. A l'inverse pour identifier le meilleur bras (donc pour $m = 1$), KL-LUCB va tirer beaucoup plus les bras sous-optimaux et a donc une meilleure estimation de leur moyenne.

Si les algorithmes pour ces deux problèmes sont différents, leur complexité l'est également. Nous avons vu que la complexité du problème de minimisation du regret est bien connue puisque nous avons, dans le cadre des bandits binaires par exemple,

$$\inf_{\substack{\mathcal{A} \text{ uniformément} \\ \text{efficace}}} \limsup_{T\to\infty} \frac{\mathrm{R}_\nu(T,\mathcal{A})}{\log(T)} = \sum_{a:\mu_a<\mu^*} \frac{(\mu^*-\mu_a)}{d(\mu_a,\mu_{a^*})},$$

où des stratégies réalisant l'infimum sont par exemple KL-UCB, Bayes-UCB ou Thompson Sampling. Pour l'identification des meilleurs bras, avec un niveau de confiance $1-\delta$ fixé, nous avons montré que

$$\inf_{\mathcal{A}\ \delta\text{–PAC}} \limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)} \geq \sum_{a\in\mathcal{S}_m^*} \frac{1}{d(\mu_a,\mu_{[m+1]})} + \sum_{a\notin\mathcal{S}_m^*} \frac{1}{d(\mu_a,\mu_{[m]})}.$$

En revanche des résultats plus fins obtenus pour l'identification du meilleur parmi $K = 2$ bras montrent que cette borne inférieure ne peut pas être atteinte, et qu'en particulier des quantités informationnelles différentes, comme l'information de Chernoff, interviennent pour mesurer la complexité du problème.

CONTENU DE LA THÈSE

L'organisation du manuscrit est la suivante : les chapitre 1 à 4 présentent nos contributions relatives à la maximisation des récompenses, et le chapitre 5 est consacré à l'identification des meilleurs bras.

Plus précisément, le chapitre 1 introduit les deux approches bayésienne et fréquentiste pour la maximisation des récompenses. Nous y présentons les mesures de performance associées, ainsi qu'un état de l'art des algorithmes fréquentistes et bayésiens. Nous y étudions notamment un algorithme basé sur les indices de Gittins à horizon fini.

Le chapitre 2 introduit l'algorithme Bayes-UCB. Nous y donnons une analyse à temps fini de l'algorithme pour les bandits binaires, et présentons également des applications dans d'autres contextes, illustrées de premières simulations numériques.

Le chapitre 3 introduit le Thompson Sampling. Nous donnons une analyse à temps fini de l'algorithme pour les bandits binaires, esquissons une analyse pour des modèles de bandit exponentiels et présentons une étude numérique comparant Bayes-UCB et Thompson Sampling aux différents algorithmes bayésiens et fréquentistes discutés au chapitre 1.

Le chapitre 4 présente l'utilisation de Bayes-UCB et Thompson Sampling dans des modèles de bandit contextuels linéaires. Nous y donnons notamment de nouveaux éléments d'analyse bayésienne.

Le chapitre 5 présente le cadre de l'identification des meilleurs bras, propose une analyse des algorithmes KL-LUCB et KL-Racing ainsi qu'une illustration numérique de leur performance. Nous y prouvons également de nouvelles bornes inférieures sur les termes de complexités des différents problèmes et proposons des algorithmes asymptotiquement optimaux dans des modèles de bandit à deux bras.

PUBLICATIONS ASSOCIÉES

Les résultats présentés dans ce manuscrit sont le fruit de collaborations avec mes directeurs de thèse, Olivier Cappé, Aurélien Garivier et Rémi Munos, mais aussi avec Nathan Korda (ancien post-doctorant à l'INRIA de Lille avec Rémi Munos) ainsi que Shivaram Kalyanakrishnan (chercheur à *Yahoo ! Labs* à Bangalore). Ils ont fait l'objet des publications suivantes, dont certains des chapitres qui suivent (lorsque c'est indiqué) seront inspirés.

Articles de conférences avec actes :

- E. Kaufmann, O. Cappé et A. Garivier, *On Bayesian Upper Confidence Bounds for Bandit Problems*, AISTATS 2012
- E. Kaufmann, N. Korda et R. Munos, *Thompson Sampling : An Asymptotically Optimal Finite Time Analysis*, ALT 2012
- E. Kaufmann et S. Kalyanakrishnan, *Information Complexity in Bandit Subset Selection*, COLT 2013
- N. Korda, E. Kaufmann et R. Munos, *Thompson Sampling for one-dimensional Exponential Family Bandits*, NIPS 2013
- E. Kaufmann, O. Cappé et A. Garivier, *On the Complexity of A/B Testing*, COLT 2014

Article de journal soumis :

- E. Kaufmann, O. Cappé et A. Garivier, *On the Complexity of Best Arm Identification in Multi-Armed Bandit Models*. Soumis.

Chapter 1

Two probabilistic views on rewards maximization in bandit models

The first part of this thesis is dedicated to the design of strategies maximizing the sum of rewards in a bandit model. As discussed in the Introduction, this objective is motivated by many applications that range from clinical trials to the display of advertising. Different communities have worked on this bandit problem, and in the particular case of parametric bandit models, two different probabilistic points of view have been considered. This chapter aims at presenting these frequentist and Bayesian approaches. We introduce the two measures of performance associated to these two probabilistic modelings, namely the regret and the Bayes risk. We present state-of-the-art algorithms in each case. In particular, we discuss the use of a strategy based on Finite-Horizon Gittins indices.

Contents

1.1 Introduction

A stochastic multi-armed *bandit model*, denoted by $\nu = (\nu_1, \dots, \nu_K)$, is a collection of K arms, where each arm ν_a is a probability distribution with mean μ_a. An agent can interact with a bandit model by choosing at each round t an arm A_t to draw. This draw results in the observation of a realization X_t from the associated distribution ν_{A_t}. The *bandit problem* that we consider in the first four chapters of this thesis is the following. The samples (X_t) collected are perceived as rewards, and the agent has to choose the arms sequentially in order to maximize the sum of the rewards accumulated during his interaction with the bandit model. Of course, the bandit model ν is unknown to him, an so is the optimal arm a^*, such that $\mu_{a^*} = \max_a \mu_a$. The mean of the optimal is denoted by μ^*.

In this thesis, we mostly consider *parametric bandit models*, of the form $\nu = \nu_{\boldsymbol{\theta}} = (\nu_{\theta_1}, \dots, \nu_{\theta_K})$. The distribution ν_a of arm a depends on a parameter $\theta_a \in \Theta$, and we let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K) \in \Theta^K$ denote the global parameter of the model. As in every parametric model, two different points of view can be adopted: the frequentist point of view, in which $\boldsymbol{\theta}$ is seen as an unknown parameter, and the Bayesian point of view, in which $\boldsymbol{\theta}$ is a random variable, drawn from some prior distribution. This leads to two different probabilistic frameworks, described in Table 1.1.

Frequentist bandit model	**Bayesian bandit model**
- $\boldsymbol{\theta} \in \Theta^K$ is an unknown parameter	- $\boldsymbol{\theta}$ is drawn from Π_0, a prior distribution on Θ^K
- $\forall a$, $(X_{a,t})$ i.i.d. with distribution ν_{θ_a} and mean μ_a	- $\forall a$, conditionally to θ_a, $(X_{a,t})$ is i.i.d. with distribution ν_{θ_a} and mean μ_a
- $(X_{a,t})_{a,t}$ is an independent family	- conditionally to $\boldsymbol{\theta}$, $(X_{a,t})_{a,t}$ is an independent family

Table 1.1: Bayesian and frequentist bandit models

We denote by $\mathbb{P}_{\boldsymbol{\theta}}$ (resp. $\mathbb{E}_{\boldsymbol{\theta}}$) —or sometimes $\mathbb{P}_\nu$ (resp. $\mathbb{E}_\nu$) in a model that is not necessarily parametric— the probability (resp. expectation) under the frequentist modeling, and $\mathbb{P}_{\Pi_0}$ (resp.$\mathbb{E}_{\Pi_0}$) the probability (resp. expectation) under the Bayesian modeling. The subscripts might be omitted in some parts of this document when the associated probabilistic framework is clear.

In both settings, an agent interacts with the bandit model using a *sampling strategy*, sometimes called *policy* or *bandit algorithm*. This sampling strategy $\mathcal{A} = (A_t)_{t\in\mathbb{N}}$ is a sequence of choices of the arms based on previous outcomes. At time t, the agent draws the arm A_t and receives as a reward an observation from arm A_t, $X_t = X_{A_t,t} \sim \nu_{A_t}$. Introducing $\mathcal{F}_t$ the filtration defined by

$$\mathcal{F}_t = \sigma(A_1, X_1, \dots, A_t, X_t),$$

in a deterministic strategy A_t is assumed to be $\mathcal{F}_{t-1}$-measurable and in a randomized strategy A_t is drawn from a probability distribution p_t on $\{1, \dots, K\}$ such that the vector of probabilities p_t is $\mathcal{F}_{t-1}$-measurable. The bandit problem considered in the first part of this thesis is the following: the agent aims at building a strategy which minimizes the expected sum of rewards up to some horizon T.

In the frequentist view of the bandit problem, a strategy that maximizes the expected cumulated rewards equivalently minimizes the *regret*, defined for any strategy $\mathcal{A}$ and horizon T by

$$\mathrm{R}_{\boldsymbol{\theta}}(T, \mathcal{A}) = \mathbb{E}_{\boldsymbol{\theta}}\left[T\mu^* - \sum_{t=1}^{T} X_t\right] = T\mu^* - \mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{t=1}^{T} X_t\right].$$

This quantity depends on the parameter $\boldsymbol{\theta}$ (fixed and unknown). The notion of regret was introduced by [Lai and Robbins, 1985] in a frequentist, parametric setting. Regret represents the difference between the expected cumulative reward of the strategy drawing the (unknown) best arm at each round, and the expected cumulative reward obtained with the strategy $\mathcal{A}$. It can be rewritten as a function of the number of draws of each sub-optimal arm. Let $N_a(t)$ denote the number of draws of arm a between the instants 1 and t. Using that $\mathbb{E}_\nu[X_t|\mathcal{F}_{t-1}] = \mu_{A_t}$, one obtains the following useful expression of the regret, as a function of the expected number of draws of each arm (that is also defined for non-parametric bandit models, for which we use the subscript ν in place of $\boldsymbol{\theta}$):

$$\mathrm{R}_\nu(T,\mathcal{A}) = \mathbb{E}_\nu\left[\sum_{t=1}^{T}(\mu^* - \mu_{A_t})\right] = \sum_{a=1}^{K}(\mu^* - \mu_a)\mathbb{E}_\nu\left[N_a(T)\right]. \tag{1.1}$$

In the Bayesian view of the bandit problem there is an equivalent notion, called *Bayes risk* (a denomination introduced by [Lai, 1987]), or sometimes Bayesian regret. A strategy that maximizes the expected cumulative rewards equivalently minimizes the Bayes risk, defined for a strategy $\mathcal{A}$ and horizon T by

$$\mathrm{BR}_{\Pi_0}(T,\mathcal{A}) = \mathbb{E}_{\Pi_0}\left[T\mu^* - \sum_{t=1}^{T}X_t\right] = \mathbb{E}_{\Pi_0}\left[\mathbb{E}_{\Pi_0}\left[T\mu^* - \sum_{t=1}^{T}X_t\middle|\boldsymbol{\theta}\right]\right] = \mathbb{E}_{\Pi_0}\left[\mathrm{R}_{\boldsymbol{\theta}}(T,\mathcal{A})\right].$$

This quantity depends on the prior distribution Π_0. An algorithm minimizing the Bayes risk is good in average on all bandit models with parameters in Θ^K, whereas in the frequentist modeling we are looking for algorithms with small regret on every bandit model $\nu_{\boldsymbol{\theta}}$, for all $\boldsymbol{\theta} \in \Theta^K$.

The first bandit algorithm was introduced by [Thompson, 1933] in a Bayesian framework, and more generally the view adopted in the first works on bandit problems was mostly Bayesian. This could be explained by the fact that Bayes risk minimization has an exact solution (that can be obtained by dynamic programming, as explained in Section 1.3), whereas there exists no algorithm minimizing the regret on every bandit model. However the lower bound on the regret given by [Lai and Robbins, 1985] allows to define the notion of asymptotic optimality in the frequentist setting and paves the way to a more abundant frequentist literature at the end of the 1980's.

The Bayesian and frequentist modelings of the bandit problem (and the dedicated performance criteria) can be dissociated from the tools related to these two frameworks. For example, algorithms from the regret minimization literature rely on maximum likelihood estimates of the unknown parameters or confidence intervals (that we call frequentist tools), whereas algorithms from the Bayesian literature choose the next arm based on the current *posterior distribution*. At the end of round t, the posterior distribution of the parameter $\boldsymbol{\theta}$ is the conditional distribution of $\boldsymbol{\theta}$ given the observation, denoted by

$$\Pi_t(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}|A_1, X_1, \ldots, A_t, X_t).$$

One can evaluate the Bayes risk of an algorithm that uses frequentist tools or conversely focus on the regret of a Bayesian algorithm that uses a prior distribution and the associated posterior distributions in its routine. The latter objective is at the heart of this thesis, in which we show that two Bayesian algorithms, Bayes-UCB and Thompson Sampling, are asymptotically optimal with respect to the regret.

Before studying the regret of Bayesian algorithms in the next chapters, we present in this chapter state-of-the-art Bayesian and frequentist algorithms and discuss the link between regret and Bayes risk. Section 1.2 is dedicated to the frequentist framework. We start by presenting the lower bound on the regret of [Lai and Robbins, 1985], and the subsequent definition of asymptotically optimal algorithms. We propose a new short proof for this result. Then we present the recent improvements in the frequentist

literature that have led to the KL-UCB algorithm of [Cappé et al., 2013], that is asymptotically optimal with respect to Lai and Robbins' lower bound. We also present some tools for regret finite-time analysis that will be useful in the rest of this thesis.

In Section 1.3, we discuss the Bayesian optimal solution. It is often heard that Gittins ([Gittins, 1979]) solved the Bayesian bandit problem by exhibiting an optimal *index policy*. Index policies are bandit algorithms in which, at each round, one index for each arm is computed, based on the history of this arm only, and then the arm with highest index is chosen. We explain here that when the goal is to maximize the expected cumulative rewards up to some finite horizon —and not the discounted sum of rewards, as in Gittins' original paper— the corresponding index policy is not optimal. Nevertheless, we conjecture that this Finite-Horizon Gittins algorithm closely approximates the Bayesian optimal solution and we present some numerical experiments supporting this claim. Approximations of the Finite-Horizon Gittins indices also indicate similarities with indices used by a frequentist index policy, KL-UCB-H$^+$, that is proved to be asymptotically optimal in a Bayesian sense.

In a nutshell, this chapter introduces frequentist algorithms (asymptotically) optimal with respect to the regret, Bayesian algorithms optimal with respect to the Bayes risk, as well as a frequentist algorithm (asymptotically) optimal with respect to the Bayes risk. Table 1.2 summarizes the information of this chapter and can be completed with the contributions of this thesis relative to regret minimization (in bold), presented in Chapter 2 and Chapter 3.

	Regret	**Bayes risk**
Frequentist algorithm	KL-UCB	KL-UCB-H$^+$
Bayesian algorithm	**Bayes-UCB** **Thompson Sampling**	Dynamic programming FH-Gittins ?

Table 1.2: Optimal algorithms for each measure of performance

Notation. We introduce here useful notation that enables us to define and later analyse bandit algorithms. Quantities that are functions of t depend on the history of the game up to the end of round t. We introduce

- $N_a(t)$, the number of draws of arm a between instants 1 and t.
- $S_a(t) = \sum_{s=1}^{t} \mathbb{1}_{(A_s=a)} X_{a,s}$, the sum of rewards obtained from arm a between instants 1 and t.
- $\hat{\mu}_a(t) = \frac{S_a(t)}{N_a(t)}$, the empirical mean of the rewards obtained from arm a between instants 1 and t.

We also introduce

- $(Y_{a,k})_{k\in\mathbb{N}^*}$, the sequence of successive rewards obtained from arm a
- $\hat{\mu}_{a,s} = \frac{1}{s}\sum_{k=1}^{s} Y_{a,k}$ the empirical mean of the first s rewards obtained from arm a

If $(A_t = a)$, $Y_{a,N_a(t)} = X_{a,t}$, where $(X_{a,t})$ is the i.i.d. sequence associated to arm a introduced in Table 1.1). One has $S_a(t) = \sum_{s=1}^{N_a(t)} Y_{a,s}$ and $\hat{\mu}_a(t) = \hat{\mu}_{a,N_a(t)}$. The sequence $(Y_{a,k})_{k\in\mathbb{N}^*}$ is i.i.d. with distribution ν_a (conditionally to θ_a in the Bayesian modeling).

1.2 The frequentist approach

For the sake of comparison with Bayesian algorithms, we mostly focus our presentation on parametric bandit models, in particular on models in which the rewards belong to an exponential family.

However, we will see that algorithms developed for parametric bandits can sometimes be generalized to some non-parametric bandit models (e.g. bandits with bounded rewards, that are often considered in the frequentist literature).

1.2.1 Lower bounds on the regret

In the frequentist framework, the bandit problem was first considered by [Robbins, 1952], who introduces two strategies for two-armed bandit models. For Bernoulli bandit models, he proposes a strategy that changes the arm that is drawn if and only if a zero is observed. For more general two-armed bandit models, Robbins proposes a second strategy that relies on two disjoint increasing sequences of integers (a_n) and (b_n) fixed in advance, with $a_1 = 1$ and $b_1 = 2$. At time t, arm 1 is drawn if $t \in (a_n)_{n\in\mathbb{N}}$, arm 2 is drawn if $t \in (b_n)_{n\in\mathbb{N}}$. Otherwise, the arm with highest empirical mean of past rewards $\hat{\mu}_a(t-1)$ is chosen. This strategy modifies the 'greedy' strategy, that chooses $A_t = \text{argmax}_a \hat{\mu}_a(t-1)$, in a way that forces exploration. If the sequences (a_n) and (b_n) are chosen such that the proportion of integers in $\{1,\dots,t\}$ that belong to one of the two sequences goes to zero when t goes to infinity, [Robbins, 1952] shows that this strategy $\mathcal{A}_R$ is such that, for every two-armed bandit model ν,

$$\frac{\mathrm{R}_\nu(T,\mathcal{A}_R)}{T} \underset{T\to\infty}{\longrightarrow} 0.$$

The seminal paper of [Lai and Robbins, 1985] gives, in simple parametric cases, a lower bound on the regret of strategies having the following stronger consistency property:

$$\text{for all } \boldsymbol{\theta} \in \Theta^K \text{ such that in } \nu_{\boldsymbol{\theta}} \text{ there is a unique optimal arm, for all } \alpha \in]0,1],\ \mathrm{R}_{\boldsymbol{\theta}}(T,\mathcal{A}) = o(T^\alpha).$$

Algorithms satisfying this property are called *uniformly efficient*. For bandits whose arms are parameterized by a single parameter (i.e. $\Theta \subset \mathbb{R}$), under some conditions on Θ, Lai and Robbins show that a uniformly efficient algorithm has to draw each sub-optimal arm at least in a logarithmic fashion. More precisely,

$$\mu_a < \mu^* \ \Rightarrow\ \liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)]}{\log(T)} \geq \frac{1}{\mathrm{KL}(\nu_{\theta_a}, \nu_{\theta^*})},$$

where $\mathrm{KL}(p,q)$ is the Kullback-Leibler divergence between the distributions p and q, defined by

$$\mathrm{KL}(p,q) = \begin{cases} \int \log\left[\frac{dp}{dq}(x)\right] dp(x) \text{ if } q \ll p, \\ +\infty \text{ otherwise.} \end{cases}$$

From the regret decomposition (1.1), it follows that, for any uniformly efficient strategy $\mathcal{A}$,

$$\liminf_{T\to\infty} \frac{\mathrm{R}_{\boldsymbol{\theta}}(T,\mathcal{A})}{\log(T)} \geq \sum_{a:\mu_a<\mu^*} \frac{(\mu^*-\mu_a)}{KL(\nu_{\theta_a}, \nu_{\theta^*})}. \tag{1.2}$$

This lower bound leads to the definition of asymptotic optimality. An algorithm is termed *asymptotically optimal* if it satisfies, for every $\boldsymbol{\theta} \in \Theta$,

$$\sup_{T\to\infty} \frac{\mathrm{R}_{\boldsymbol{\theta}}(T,\mathcal{A})}{\log(T)} \leq \sum_{a:\mu_a<\mu^*} \frac{(\mu^*-\mu_a)}{\mathrm{KL}(\nu_{\theta_a}, \nu_{\theta^*})}.$$

Lai and Robbins' lower bound was later generalized by [Burnetas and Katehakis, 1996] to distributions that depend on multiple parameters. We give here a slightly more general result, that does not rely on parametric assumptions.

Definition 1.1. *A class $\mathcal{M}$ of bandit models is identifiable if it is of the form $\mathcal{M} = \mathcal{P}_1 \times \cdots \times \mathcal{P}_K$, where $\mathcal{P}_a$ is the set of possible distributions for arm a, and if for all a, $\mathcal{P}_a$ is such that*

$$\forall p, q \in \mathcal{P}_a, \quad p \neq q \;\Rightarrow\; 0 < \mathrm{KL}(p,q) < +\infty.$$

Theorem 1.2. *Let $\mathcal{M}$ be an identifiable class of bandit models. Let $\mathcal{A}$ be a bandit algorithm uniformly efficient on the class $\mathcal{M}$: for all $\nu \in \mathcal{M}$ with a unique optimal arm, for all $\alpha \in]0,1]$, $R_\nu(T,\mathcal{A}) = o(T^\alpha)$. Then, for all $\nu \in \mathcal{M}$,*

$$\mu_a < \mu^* \;\Rightarrow\; \liminf_{T\to\infty} \frac{\mathbb{E}_\nu[N_a(T)]}{\log(T)} \geq \frac{1}{\mathcal{K}^a_{inf}(\nu_a;\mu^*)}, \tag{1.3}$$

with

$$\mathcal{K}^a_{inf}(p;\mu) = \inf\left\{\mathrm{KL}(p,q) : q \in \mathcal{P}_a \text{ and } \mathbb{E}_{Y\sim q}[Y] > \mu\right\}.$$

All the distribution-dependent lower bounds derived in the bandit literature (e.g. [Lai and Robbins, 1985, Burnetas and Katehakis, 1996] but also [Mannor and Tsitsiklis, 2004, Audibert et al., 2010] in the literature relative to best arm identification, that will be presented in Chapter 5) rely on *changes of distribution*, and so does Theorem 1.2. A change of distribution relates the probabilities of the same event under two different bandit models ν and ν'. Lemma 1.3 below provides a new, synthetic, inequality from which all the lower bounds presented in this thesis will be directly derived. In other words, this result, whose proof is postponed to Section 1.5.1, encapsulates the technical aspects of the change of distribution. Lemma 1.3 could also be used to give simple proofs for the lower bounds of [Graves and Lai, 1997, Agrawal et al., 1989] in more general cases in which arms are not necessarily independent.

Lemma 1.3. *Let ν and ν' be two bandit models such that the distributions of all arms in ν and ν' are mutually absolutely continuous. Let σ be a stopping time with respect to $(\mathcal{F}_t)$ such that $(\sigma < +\infty)$ a.s. under both models. Let $A \in \mathcal{F}_\sigma$ be an event such that $0 < \mathbb{P}_\nu(A) < 1$. Then one has*

$$\sum_{a=1}^{K} \mathbb{E}_\nu[N_a(\sigma)]\mathrm{KL}(\nu_a,\nu'_a) \geq d(\mathbb{P}_\nu(A),\mathbb{P}_{\nu'}(A)),$$

where $d(x,y) := x\log(x/y) + (1-x)\log((1-x)/(1-y))$ is the binary relative entropy.

Proof of Theorem 1.2. Let $\nu = (\nu_1,\ldots,\nu_K)$ be a bandit model such that arm 1 is the unique optimal arm. Let $a \neq 1$ be a suboptimal arm. Consider the alternative bandit model ν' such that $\nu'_i = \nu_i$ for all $i \neq a$ and $\nu'_a \in \mathcal{P}_a$ is such that $\mathbb{E}_{Y\sim\nu'_a}[Y] > \mu_1$. Arm 1 is thus the unique optimal arm in the bandit model ν, whereas arm a is the unique optimal arm in the bandit model ν'. For every integer T, let A_T be the event defined by

$$A_T = \left(N_1(T) \leq T - \sqrt{T}\right).$$

Clearly, $A_T \in \mathcal{F}_T$. From Lemma 1.3, applied to the stopping time $\sigma = T$ a.s.,

$$\mathbb{E}_\nu[N_a(T)]\mathrm{KL}(\nu_a,\nu'_a) \geq d(\mathbb{P}_\nu(A_T),\mathbb{P}_{\nu'}(A_T)). \tag{1.4}$$

The event A_T is not very likely to hold under the model ν, in which the optimal arm should be drawn of order $T - O(\log(T))$ times, whereas it is very likely to happen under ν', in which arm 1 is sub-optimal and thus only drawn little. More precisely, Markov inequality yields

$$\begin{aligned}
\mathbb{P}_\nu(A_T) &= \mathbb{P}_\nu(T - N_1(T) \geq \sqrt{T}) \leq \frac{\sum_{i\neq 1}\mathbb{E}_\nu[N_i(T)]}{\sqrt{T}} \\
\mathbb{P}_{\nu'}(A^c_T) &= \mathbb{P}_{\nu'}(N_1(T) \geq T - \sqrt{T}) \leq \frac{\mathbb{E}_{\nu'}[N_1(T)]}{T-\sqrt{T}} \leq \frac{\sum_{i\neq a}\mathbb{E}_{\nu'}[N_i(T)]}{T-\sqrt{T}}
\end{aligned}$$

From the formulation (1.1), every uniformly efficient algorithm satisfies

$$\sum_{i\neq 1}\mathbb{E}_\nu[N_i(T)] = o(T^\alpha) \text{ and } \sum_{i\neq a}\mathbb{E}_{\nu'}[N_i(T)] = o(T^\alpha)$$

for all $\alpha \in]0,1]$. Hence $\mathbb{P}_\nu(A_T) \underset{T\to\infty}{\to} 0$ and $\mathbb{P}_{\nu'}(A_T) \underset{T\to\infty}{\to} 1$. Therefore, we get

$$\frac{d(\mathbb{P}_\nu(A_T),\mathbb{P}_{\nu'}(A_T))}{\log(T)} \underset{T\to\infty}{\sim} \frac{1}{\log(T)}\log\left(\frac{1}{\mathbb{P}_{\nu'}(A_T^c)}\right) \geq \frac{1}{\log(T)}\log\left(\frac{T-\sqrt{T}}{\sum_{i\neq a}\mathbb{E}_{\nu'}[N_i(T)]}\right).$$

The right hand side rewrites

$$1 + \frac{\log\left(1-\frac{1}{\sqrt{T}}\right)}{\log(T)} - \frac{\log\left(\sum_{i\neq a}\mathbb{E}_{\nu'}[N_i(T)]\right)}{\log(T)} \underset{T\to\infty}{\longrightarrow} 1$$

where we use the fact that $\sum_{i\neq a}\mathbb{E}_{\nu'}[N_i(T)] = o(T^\alpha)$ for all $\alpha \in]0,1]$. Finally, for every $\nu'_a \in \mathcal{P}_a$ such that $\mathbb{E}_{Y\sim\nu'_a}[Y] > \mu_1$ one obtains, using inequality (1.4)

$$\liminf_{T\to\infty}\frac{\mathbb{E}[N_a(T)]}{\log(T)} \geq \frac{1}{\mathrm{KL}(\nu_a,\nu'_a)}.$$

For all $\epsilon \in]0,1[$, ν'_a can then be chosen such that $\mathcal{K}^a_{\inf}(\nu_a,\mu_1) \leq \mathrm{KL}(\nu_a,\nu'_a) \leq \mathcal{K}^a_{\inf}(\nu_a,\mu_1)/(1-\epsilon)$, and the conclusion follows when ϵ goes to zero.

□

Distribution independent lower bounds on the regret. The lower bounds given in this section are *distribution-dependent*: for each bandit model ν and each algorithm $\mathcal{A}$, the regret $\mathrm{R}_\nu(T,\mathcal{A})$ is lower bounded by a quantity that depends on the arms distributions. More precisely, there exists a constant $C(\nu)$ –given in (1.2) for simple classes of parametric bandits– such that

$$\mathrm{R}_\nu(T,\mathcal{A}) \geq C(\nu)\log(T). \tag{1.5}$$

It is also possible to consider *minimax* performance bounds on the regret. Problem-independent upper and lower bounds on the regret have emerged in the literature on *adversarial bandits* (see [Cesa-Bianchi and Lugosi, 2006]), in which there is no stochastic assumptions on the arms. However, the distribution-independent lower bound first given by [Cesa-Bianchi and Lugosi, 2006] can also be formulated in our stochastic setting. Indeed, Theorem 3.5 of [Bubeck and Cesa-Bianchi, 2012] states that for every bandit algorithm $\mathcal{A}$, there exists a stochastic bandit model ν such that ν_a is supported in $[0,1]$ for all a and

$$\mathrm{R}_\nu(T,\mathcal{A}) \geq \frac{1}{20}\sqrt{KT}. \tag{1.6}$$

This *worst-case* result gives a regret lower bound for an algorithm belonging to $\mathrm{argmin}_{\mathcal{A}}\max_\nu \mathrm{R}_\nu(T,\mathcal{A})$. In this thesis, we rather aim at finding algorithms that are optimal for every bandit model ν, with respect to the distribution-dependent lower bound (1.5).

As already noted by [Bubeck and Liu, 2013], the proof of Theorem 3.5 of [Bubeck and Cesa-Bianchi, 2012] also yields a lower bound on the Bayes risk. Indeed, the authors lower bound the quantity

$$\frac{1}{K}\sum_{i=1}^K \mathbb{E}_{\nu_\epsilon^{(i)}}\left[\sum_{t=1}^T (X_{i,t} - X_{A_t,t})\right],$$

where $\nu_\epsilon^{(i)}$ is the bandit model in which all arms are Bernoulli distributions with mean $(1-\epsilon)/2$, expect arm i which is Bernoulli with mean $(1+\epsilon)/2$. Their result implies that for every bandit algorithm $\mathcal{A}$, there exists a prior distribution Π (uniform over a finite set of bandit models $\{\nu_\epsilon^{(1)}, \dots, \nu_\epsilon^{(K)}\}$) such that

$$\mathrm{BR}_\Pi(T, \mathcal{A}) \geq \frac{1}{20}\sqrt{KT}. \tag{1.7}$$

In Section 1.3 we will present the optimal strategy for any given prior distribution Π_0. For some specific prior distribution, we provide in Section 1.3.5 a lower bound on the Bayes risk $\mathrm{BR}_{\Pi_0}(T, \mathcal{A})$ that depends on the prior, that can be qualified as *prior-dependent*.

1.2.2 Examples of bandit models and associated tools to build bandit algorithms

Classes of *exponential bandit model*, such that the arms distributions all belong to the same *one-parameter canonical exponential family* form important examples in which the Lai and Robbins' lower bound holds. In such a class $\mathcal{M}$, there exists two functions A and b such that $\mathcal{M} = \{\nu = (\nu_{\theta_1}, \dots, \nu_{\theta_K}) : \forall a \in \{1, \dots, K\}, \theta_a \in \Theta\}$, with the distribution ν_θ having a density $f(\cdot;\theta)$ with respect to some reference measure λ given by

$$f(x;\theta) = A(x)\exp(x\theta - b(\theta)), \quad \theta \in \Theta \in \mathbb{R}. \tag{1.8}$$

The log-partition function $b : \Theta \to \mathbb{R}$ is supposed to be twice differentiable. Under this assumption, it can be shown that the mean of the distribution ν_θ is $\mu(\theta) = \dot{b}(\theta)$ and its variance is $\ddot{b}(\theta)$ (see e.g. [Cappé et al., 2013] for more details on exponential families). Thus, the mapping $\theta \mapsto \mu(\theta)$ is increasing and distributions belonging to a one-parameter canonical exponential family can be either parameterized by their *natural parameter* θ or by their mean μ. Hence, each exponential family induces a divergence on $(\dot{b}(\Theta))^2$, defined by

$$d(\mu, \mu') := \mathrm{KL}\left(\nu_{\dot{b}^{-1}(\mu)}, \nu_{\dot{b}^{-1}(\mu')}\right). \tag{1.9}$$

Besides, the Kullback-Leibler divergence between two distributions in an exponential family parameterized by the natural parameters θ and θ' respectively will be denoted by $\mathrm{K}(\theta, \theta') := \mathrm{KL}(\nu_\theta, \nu_{\theta'})$.

Many classical families of parametric distributions form a one-parameter canonical exponential family, like Gaussian distributions with known variance (*Gaussian bandit models*), Bernoulli distributions (*Bernoulli bandit models*), Poisson and exponential distributions (see Table 1.3). Compared to the definition (1.8), one can also consider (following the definition of [Bickel and Doksum, 2001] for example) slightly more general families, for which the density of ν_θ is given by $A(x)\exp(T(x)\theta - b(\theta))$, with $T(x) \neq x$. This generalization will be discussed in Chapter 3.

For a given class of exponential bandit models (for example Gaussian bandits or Bernoulli bandits), with $d(\mu, \mu')$ the associated divergence defined in (1.9), every uniformly efficient bandit algorithm $\mathcal{A}$ satisfies

$$\liminf_{T\to\infty} \frac{\mathrm{R}_{\boldsymbol{\theta}}(T, \mathcal{A})}{\log(T)} \geq \sum_{a:\mu_a<\mu^*} \frac{(\mu^* - \mu_a)}{d(\mu_a, \mu^*)}. \tag{1.10}$$

Deviation inequalities and confidence intervals for exponential families. In an exponential family, the relationship between the moment-generating function and Kullback-Leibler divergence given by Lemma 1.4 below is a crucial property that allows to build good confidence intervals. The logarithm of the moment-generating function of a random variable X and its Fenchel-Legendre transform are respectively defined by

$$\phi_X(\lambda) = \log \mathbb{E}[e^{\lambda X}] \quad \text{and} \quad \phi_X^*(x) = \sup_{\lambda\in\mathbb{R}} \{x\lambda - \phi_X(\lambda)\}.$$

	Density	θ	$b(\theta)$	$d(\mu,\mu')$
Gaussian distribution $\mathcal{N}(\mu,\sigma^2)$ (known variance)	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\frac{\mu}{\sigma^2}$	$\frac{\sigma^2\theta^2}{2}$	$\frac{(\mu-\mu')^2}{2\sigma^2}$
Bernoulli distribution $\mathcal{B}(\mu)$ (mean μ)	$\mu^x(1-\mu)^{1-x}\mathbb{1}_{\{0,1\}}(x)$	$\log\frac{\mu}{1-\mu}$	$\log(1+e^\theta)$	$\mu\log\frac{\mu}{\mu'}+(1-\mu)\log\frac{1-\mu}{1-\mu'}$
Poisson distribution $\mathcal{P}(\lambda)$ (mean λ)	$\frac{\lambda^x}{x!}e^{-\lambda}\mathbb{1}_{\mathbb{N}^*}(x)$	$\log(\lambda)$	e^θ	$\mu'-\mu+\mu\log\frac{\mu}{\mu'}$
Exponential distribution $\mathcal{E}(\lambda)$ (mean $1/\lambda$)	$\lambda e^{-\lambda x}\mathbb{1}_{\mathbb{R}^+}(x)$	$-\lambda$	$-\log(-\theta)$	$\frac{\mu}{\mu'}-1-\log\frac{\mu}{\mu'}$
Gamma distribution $\Gamma(k,\lambda)$ (mean k/λ)	$\frac{\lambda^k}{\Gamma(k)}x^{k-1}e^{-\lambda x}\mathbb{1}_{\mathbb{R}^+}(x)$	$-\lambda$	$-k\log(-\theta)$	$k\left(\frac{\mu}{\mu'}-1-\log\frac{\mu}{\mu'}\right)$

Table 1.3: Examples of exponential families and associated divergence

From Cramér's Theorem (Theorem 2.2.3 of [Dembo and Zeitouni, 2010]), this last quantity can be regarded as the optimal rate at which the empirical mean of i.i.d. samples concentrates around the true mean. More precisely, if (X_i) is an i.i.d. sequence with expectation μ, and if $\hat{\mu}_s = \frac{1}{s}\sum_{i=1}^s X_i$ denotes the empirical mean of the first s observations, for every $x > \mu$,

$$\mathbb{P}(\hat{\mu}_s \geq x) \leq e^{-s\phi^*_{X_1}(x)} \qquad \text{and} \qquad \lim_{s\to\infty} -\frac{1}{s}\log\mathbb{P}(\hat{\mu}_s \geq x) = \phi^*_{X_1}(x).$$

Because of this second statement, the deviation inequality stated first is optimal with respect to the large deviation principle. In an exponential family, Lemma 1.4 gives a close form for the rate function $\phi^*_{X_1}(x)$ that yields such an optimal deviation inequality for the empirical mean of i.i.d. samples.

Lemma 1.4. *If $X \sim \nu_\theta$, then $\phi^*_X(x) = d(x,\mu(\theta))$.*

Going a bit further, one can give a deviation inequality for the empirical mean of independent random variables 'dominated' by some distribution in an exponential family.

Lemma 1.5 (Chernoff inequality)**.** *Let (X_i) be a sequence of independent random variables such that*

$$\forall i \in \mathbb{N},\ \phi_{X_i}(\lambda) \leq \phi_Y(\lambda), \tag{1.11}$$

where $Y \sim \nu_\theta$ belong to an exponential family with associated divergence $d(\mu,\mu')$.

Let $\mu = \mathbb{E}[Y]$. Then if $\hat{\mu}_s = \frac{1}{s}\sum_{i=1}^s X_i$, one has

$$\begin{aligned} \text{for } x > \mu, \quad \mathbb{P}(\hat{\mu}_s \geq x) &\leq \exp(-sd(x,\mu)), \\ \text{for } x < \mu, \quad \mathbb{P}(\hat{\mu}_s \leq x) &\leq \exp(-sd(x,\mu)). \end{aligned}$$

Proof. The result follows from the Cramer-Chernoff method (see e.g. [Boucheron et al., 2013]). Using Markov inequality, the independence of the X_i and the upper bound on the $\phi_{X_i}(\lambda)$, one can write for any $\lambda > 0$,

$$\begin{aligned} \mathbb{P}\left(\frac{1}{s}\sum_{i=1}^s X_i \geq x\right) &= \mathbb{P}\left(e^{\lambda\sum_{i=1}^s X_i} \geq e^{\lambda sx}\right) \leq e^{-\lambda sx}\prod_{i=1}^s \mathbb{E}\left[e^{\lambda X_i}\right] = e^{-\lambda sx}\prod_{i=1}^s e^{\phi_{X_i}(\lambda)} \\ &\leq e^{-\lambda sx}\left(e^{\phi_Y(\lambda)}\right)^s = e^{-s(\lambda x-\phi_Y(\lambda))}. \end{aligned}$$

Optimizing in $\lambda \in \mathbb{R}^+$ to obtain the tightest possible inequality yields the result. Indeed, for $x > \mu$ it can be shown that

$$\sup_{\lambda \in \mathbb{R}^+} (\lambda x - \phi_Y(\lambda)) = \sup_{\lambda \in \mathbb{R}} (\lambda x - \phi_Y(\lambda)) = \phi_Y^*(x) = d(x, \mu),$$

using Lemma 1.4. The proof for $x < \mu$ follows the same lines.

□

Deviation inequalities for non-parametric distributions can be deduced from Lemma 1.5. The distribution of a random variable X is called σ^2-*subgaussian* if $\phi_X(\lambda) \leq \lambda^2\sigma^2/2$, which is the moment-generating function of the distribution $\mathcal{N}(0, \sigma^2)$. The associated divergence is $d(x, y) = (x - y)/(2\sigma^2)$ and one obtains from Lemma 1.5 that the empirical mean of i.i.d. samples of a σ^2 subgaussian distribution satisfies

$$\mathbb{P}(\hat{\mu}_s \geq x) = \mathbb{P}(\hat{\mu}_s \leq -x) \leq \exp\left(-s\frac{x^2}{2\sigma^2}\right).$$

From Hoeffding's lemma (see [Hoeffding, 1963]) every centered distribution with bounded support in $[a, b]$ is $\frac{(b-a)^2}{4}$-subgaussian. This remark yields Hoeffding's inequality.

Lemma 1.6 (Hoeffding's inequality)**.** *Let (X_i) be an i.i.d. sequence with mean μ that is supported in $[a, b]$. Then*

$$\mathbb{P}(\hat{\mu}_s \geq \mu + x) = \mathbb{P}(\hat{\mu}_s \leq \mu - x) \leq \exp\left(-s\frac{2x^2}{(b-a)^2}\right).$$

Moreover, when X_1 is supported in $[0, 1]$ and has mean μ, another result from [Hoeffding, 1963] shows that $\phi_{X_1}(x)$ is upper bounded by $\phi_{\mathcal{B}(\mu)}(x)$, the log moment-generating function of a Bernoulli distribution with same mean μ. Thus Lemma 1.5 also applies to bounded distributions with support in $[0, 1]$, taking $d(x, y) = \mathrm{KL}(\mathcal{B}(x), \mathcal{B}(y))$.

Refined confidence intervals for bounded and Bernoulli distributions. For Bernoulli distributions and for bounded distributions supported in $[0, 1]$, both Hoeffding and Chernoff inequalities can be used to build confidence intervals on the mean. Let $d(x, y)$ be the divergence associated to Bernoulli distributions. Let (X_i) be an i.i.d. sequence of Bernoulli (or bounded) random variables with mean μ and $\hat{\mu}_s$ be the empirical mean of the first s observations. Hoeffding's inequality (Lemma 1.6) yields the following confidence interval for the mean μ:

$$\mathbb{P}\left(\mu \in \left[\hat{\mu}_s - \sqrt{\frac{\log(1/\delta)}{2s}}; \hat{\mu}_s + \sqrt{\frac{\log(1/\delta)}{2s}}\right]\right) \geq 1 - 2\delta, \tag{1.12}$$

whereas using Chernoff inequality (Lemma 1.5) it can be shown that

$$\mathbb{P}\Big(sd(\hat{\mu}_s, \mu) \leq \log(1/\delta)\Big) \geq 1 - 2\delta. \tag{1.13}$$

Indeed, one has for example, letting x^* be defined by $x^* < \mu$ and $sd(x^*, \mu) = \log(1/\delta)$,

$$\mathbb{P}(\mu \geq \hat{\mu}_s, sd(\hat{\mu}_s, \mu) \geq \log(1/\delta)) = \mathbb{P}(\hat{\mu}_s \geq x^*) \leq e^{-sd(x^*, \mu)} = \delta.$$

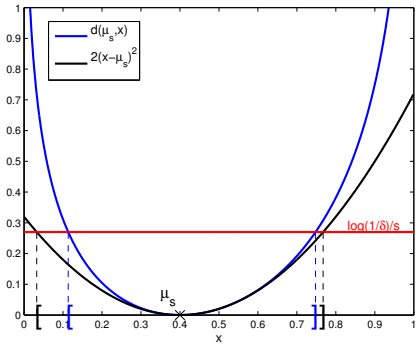

Figure 1.1: Two confidence intervals with the same coverage probability.

This second confidence region, (1.13), is implicitly defined through the function d. For example, the associated upper confidence bound can be written

$$u_s(\delta) = \max\{q > \hat{\mu}_s : sd(\hat{\mu}_s, q) \leq \log(1/\delta)\}.$$

The confidence region (1.12) also takes the form (1.13) if the function d is replaced by $\tilde{d}(x,y) = 2(x-y)^2$. From Pinsker's inequality, one has $d(x,y) > 2(x-y)^2$. While the two confidence intervals have the same coverage probability, Figure 1.1 illustrates that the confidence interval of (1.13), represented in blue, is contained in that of (1.12), represented in black. The use of these refined confidence intervals based on the divergence d will be crucial to build asymptotically optimal bandit algorithms.

1.2.3 Asymptotically optimal algorithms

In addition to their lower bound on the regret, [Lai and Robbins, 1985] provide the first asymptotically optimal bandit algorithms. These first optimal policies are index policies: at each round t, an index is computed for each arm (based on the past observations of this arm only), and the arm A_t chosen is the one with highest index. Index policies are reminiscent of the policy based on Gittins indices introduced earlier by [Gittins, 1979] in a Bayesian framework, that will be presented in Section 1.3. The form of the indices proposed by [Lai and Robbins, 1985] is however not very explicit, and the analysis proposed is asymptotic. [Agrawal, 1995] proposes simpler index policies and introduces the notion of *UCB-type algorithm* (for Upper Confidence Bound). Indeed, the indices used involve the empirical mean of past rewards and the number of draws of each arm and can be interpreted as upper confidence bound for the unknown mean of each arm. In the particular case of Gaussian bandits with known variance σ^2, [Katehakis and Robbins, 1995] propose the following fully explicit index policy and prove its asymptotic optimality. After an initialization phase in which each arm is drawn once, their policy chooses at time $t+1$

$$A_{t+1} = \operatorname*{argmax}_a U_a(t) \quad \text{with} \quad U_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{2\sigma^2 \log(t)}{N_a(t)}}.$$

UCB-type algorithms were popularized at the beginning of the years 2000, with the introduction of the UCB1 algorithm by [Auer et al., 2002a], for which the authors provide the first *finite-time analysis*. UCB1 is designed for the class of bandit models whose arms distributions have bounded support in

$[0,1]$, that we refer to as *bounded bandit models* in the sequel. UCB1 chooses at time $t+1$ the arm

$$A_{t+1} = \underset{a}{\operatorname{argmax}}\, U_a(t) \quad \text{with} \quad U_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{\beta(t)}{2N_a(t)}},$$

with an *exploration rate* $\beta(t) = 4\log(t)$. At round t, for arm a, the number of draws $N_a(t)$ being fixed, $U_a(t)$ can be interpreted as an Upper Confidence Bound obtained with Hoeffding inequality (see (1.12)) that holds with probability larger than $1 - 1/t^4$. UCB-type algorithms use the principle of *optimism in face of uncertainty*: among all the statistically plausible models, the arm chosen is the optimal arm in the best possible model (in which the mean of each arm is equal to its upper confidence bound). The theoretical guarantees for UCB1 are the following.

Theorem 1.7 ([Auer et al., 2002a],Theorem 1)**.** *Defining $\Delta_a := \mu^* - \mu_a$ as the squared gap between the optimal arm and arm a, if ν is a bounded bandit model,*

$$\mathrm{R}_\nu(T, \mathrm{UCB1}) \leq 8\left(\sum_{a:\mu_a<\mu^*} \frac{1}{\Delta_a}\right)\log(T) + \left(1 + \frac{\pi^2}{3}\right)\sum_{a=1}^{K} \Delta_a.$$

[Auer et al., 2002a] thus provide an efficient algorithm, together with a logarithmic upper bound on its regret, for the non-parametric class of bounded bandit models, which is interesting in itself. However, the class $\mathcal{M}$ of bounded bandit models is not identifiable (according to Definition 1.1), and Theorem 1.2 does not apply. There is no lower bound on the regret of algorithms that are uniformly efficient on this class $\mathcal{M}$ to which the result of Theorem 1.7 could be compared. But for Bernoulli bandit models, to which UCB1 can also be applied, this algorithm is sub-optimal with respect to the Lai and Robbins' lower bound. Indeed, Pinsker's inequality shows that $d(\mu_a, \mu^*) > 2\Delta_a^2$, hence the constant in front of $\log(T)$ is at least sixteen times bigger than the optimal constant prescribed by (1.10).

Successive refinements in the proof of Theorem 1.7, like the use of a smaller exploration rate of the form $\beta(t) = \alpha\log(t)$ with $\alpha > 1$ (see e.g. [Audibert et al., 2009, Bubeck, 2010]) lead to a smaller constant in front of the $\log(T)$, yet still expressed in terms of the squared gaps Δ_a. The UCB-V algorithm of [Audibert et al., 2009], still designed for bounded bandit models, uses for each arm a confidence interval built by taking into account the empirical variance of the arm (built using a empirical Bernstein bound). The regret bound obtained is no longer only a function of the means of the arms, but involves the quantities σ_a^2/Δ_a for each suboptimal arm, where σ_a^2 is the variance of arm a. The UCB-Tuned algorithm proposed by [Auer et al., 2002a] also uses variances estimates, but no theoretical guarantees for this algorithm are provided.

To obtain UCB-type algorithms matching Lai and Robbins' lower bound, one needs to use refined confidence intervals, based on Kullback-Leibler divergence. This idea, already suggested by [Lai and Robbins, 1985, Lai, 1987, Agrawal, 1995] reappeared with the DMED algorithm of [Honda and Takemura, 2010], designed for bounded bandit models. [Cappé et al., 2013] study a simpler index policy, that bears some similarities with DMED, called KL-UCB. While the general KL-UCB algorithm does not require any parametric assumption on the rewards distributions, the authors provide theoretical guarantees in two interesting parametric cases. They give a finite-time analysis for KL-UCB applied to exponential family bandit models (previously studied by [Garivier and Cappé, 2011]) and for bandit models whose arms distributions have a finite (known) support (previously studied by [Maillard et al., 2011]).

In an exponential bandit model, if $d(x,y) = \mathrm{KL}(\nu_{\dot{b}^{-1}(x)}, \nu_{\dot{b}^{-1}(x)})$ is the associated divergence, the KL-UCB algorithm is the index policy associated to

$$u_a(t) = \sup\{q \geq \hat{\mu}_a(t) : N_a(t)d\left(\hat{\mu}_a(t), q\right) \leq \beta(t)\},$$

with $\beta(t) = \log(t) + 3\log\log(t)$. This index appears as the upper confidence bound of a confidence interval built with Chernoff inequality (see (1.13)) that holds with probability $1 - 1/(t\log^3(t))$. This particular case of the algorithm is called kl-UCB by [Cappé et al., 2013]. In this document, the capital notation is kept to stand for the algorithm designed for rewards in an exponential family. One has the following result.

Theorem 1.8 ([Cappé et al., 2013], Theorem 1)**.** *In an exponential bandit model with associated divergence d, the KL-UCB algorithm satisfies, for any suboptimal arm a,*

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)] \quad \leq \quad & \frac{1}{d(\mu_a,\mu^*)}\log(T) + 2\sqrt{\frac{2\pi\sigma_{a,*}^2 d'(\mu_a,\mu^*)^2}{d(\mu_a,\mu^*)^3}}\sqrt{\log(T)+3\log\log(T)} \\ & + \left(4e + \frac{3}{d(\mu_a,\mu^*)}\right)\log\log(T) + 8\pi\sigma_{a,*}^2\left(\frac{d'(\mu_a,\mu^*)}{d(\mu_a,\mu^*)}\right)^2 + 6,\end{aligned}$$

where $\sigma_{a,*}^2 = \max\{Var[\nu_\theta] : \mu_a \leq E(\nu_\theta) \leq \mu^*\}$ and $d'(x,y) = \frac{\partial d(x,y)}{\partial x}$.

It follows from Theorem 1.8 and the regret decomposition (1.1) that KL-UCB is asymptotically optimal in every class of exponential bandit models. As explained by [Cappé et al., 2013], KL-UCB can also be used as it is for bounded bandit models (with support in $[0,1]$) with the divergence d associated to Bernoulli distributions, and the upper bound of Theorem 1.8 still holds true. This is (mostly) because the confidence region (1.13) also holds for this non-parametric class of bandit models. Similarly, it can be shown that KL-UCB with the divergence $d(x,y) = (x-y)^2/(2\sigma^2)$ associated to Gaussian distributions with variance σ^2 can also be applied to bandit models with σ^2-subgaussian rewards distributions, with the same theoretical guarantees. Thus, the analysis of [Cappé et al., 2013] also yields a new upper bound on the regret of UCB1 in bounded bandit models: this algorithm can indeed be seen as KL-UCB with the divergence $d(x,y) = 2(x-y)^2$ associated to Gaussian distributions with variance 1/4.

Sketch of a finite-time analysis. We do not provide the proof of Theorem 1.8, but we present the general structure of the finite-time analysis proposed by [Cappé et al., 2013], that can be applied to any optimistic index policy. We highlight the improvements proposed by the authors, some of which will be useful in the rest of this thesis.

Let $\nu = (\nu_1,\dots,\nu_K)$ be a stochastic bandit model. An optimistic index policy proceeds in the following way. After an initialization phase in which each arm is drawn once, the arm chosen at time $t+1$ maximizes some index $U_a(t) = U_{a,N_a(t),t}$, chosen such that for all $s \leq t$, $\mathbb{P}(U_{a,s,t} < \mu_a) \leq e^{-\beta(t)}$ for some exploration rate $\beta(t)$. To upper bound the regret of such a strategy, using the decomposition (1.1), it is sufficient to upper bound, for every suboptimal a, the quantity

$$\mathbb{E}_\nu[N_a(T)] = \mathbb{E}_\nu\left[\sum_{t=1}^{T}\mathbb{1}_{(A_t=a)}\right] = 1 + \mathbb{E}_\nu\left[\sum_{t=K}^{T-1}\mathbb{1}_{(A_{t+1}=a)}\right].$$

The event $(A_{t+1}=a)$ can then be decomposed in the following way, depending on whether the optimal arm μ_1 is or not under-estimated by its index $U_1(t)$. One also uses that if the suboptimal a is drawn at time $t+1$, one has in particular $U_a(t) > U_1(t)$. This leads to

$$\begin{aligned}(A_{t+1}=a) \quad & \subseteq \quad (U_1(t) < \mu_1) \cup (A_{t+1}=a, U_1(t) > \mu_1) \\ & \subseteq \quad (U_1(t) < \mu_1) \cup (A_{t+1}=a, U_a(t) > \mu_1).\end{aligned}$$

It follows that

$$\mathbb{E}_\nu[N_a(T)] \leq 1 + \underbrace{\sum_{t=K+1}^{T} \mathbb{P}_\nu\left(U_{1,N_1(t),t} \leq \mu_1\right)}_{\text{Term A: related to the possible under estimation of optimal arm}} + \underbrace{\sum_{t=K}^{T-1} \mathbb{P}_\nu\left(A_{t+1} = a, U_{a,N_a(t),t} > \mu_1\right)}_{\text{Term B: related to the possible over estimation of the suboptimal arm } a}$$

The 'under-estimation' term A is shown to be of order $o(\log(T))$. To upper bound this term, one needs to control the probability of the event $\left\{U_{1,N_1(t),t} \leq \mu_1\right\}$, that involves the *self-normalized* quantity $U_{1,N_1(t),t}$. This denomination comes from the fact that $U_{1,s,t}$ is already random, and the number of observations $N_1(t)$ is also random. A first idea to control this quantity is to use a union bound:

$$\mathbb{P}_\nu\left(U_{1,N_1(t),t} \leq \mu_1\right) \leq \mathbb{P}_\nu\left(\exists s \in \{1,\dots t\} : U_{1,s,t} \leq \mu_1\right) \leq \sum_{s=1}^{t} \mathbb{P}_\nu\left(U_{1,s,t} \leq \mu_1\right) \leq te^{-\beta(t)}.$$

[Cappé et al., 2013] advocate the use of more sophisticated *self-normalized deviation inequalities* to control this probability. To be able to analyse KL-UCB, the authors propose an *informational* self-normalized deviation inequality, that is expressed with the divergence function d, stated in Lemma 1.9. [Garivier, 2013] presents more general self-normalized inequalities, and discuss different applications.

Lemma 1.9. *Let (X_i) be a sequence of independent random variables such that $\phi_{X_i}(\lambda) \leq \phi_Y(\lambda)$ where $Y \sim \nu_\theta$ belongs to an exponential family with mean μ and associated divergence d. One has*

$$\mathbb{P}\left(\exists s \in \{1,\dots,t\} : s\, d^+\left(\frac{1}{s}\sum_{i=1}^{s} X_i, \mu\right) > \gamma\right) \leq e\lceil \gamma \log(t)\rceil \exp(-\gamma),$$

where $d^+(x,y) = d(x,y)\mathbb{1}_{(x<y)}$.

From Lemma 1.5, introducing also $d^-(x,y) = d(x,y)\mathbb{1}_{(x>y)}$, it easily follows that

$$\mathbb{P}(s\, d^+(\hat{\mu}_s, \mu) \geq \gamma) \leq e^{-\gamma} \quad \text{and} \quad \mathbb{P}(s\, d^-(\hat{\mu}_s, \mu) \geq \gamma) \leq e^{-\gamma} \tag{1.14}$$

Whereas a union bound, using (1.14) for all s, would upper bound the probability in Lemma 1.9 by $\boldsymbol{t}e^{\gamma}$, the bound in Lemma 1.9 is of order $\mathbf{log}\,(\boldsymbol{t})e^{-\gamma}$, and is thus significantly smaller. To obtain such a result, the union bound, summing over all the possible values of $N_1(t) \in \{1,\dots,t\}$ is replaced by a geometric 'peeling': one considers 'slices' on which $N_1(t) \in [(1+\eta)^{k-1}, (1+\eta)^k]$ for some parameter $\nu > 0$. On each slice, a maximal inequality for a well-chosen martingale is applied. This 'peeling-trick' will also be used in the analysis of Bayes-UCB and Thompson Sampling in the next two chapters. Lemma 1.9 will also be used directly several times in this document. We thus provide its proof in Section A.2 of Appendix A, in which we also present other self-normalized deviation inequalities used in this thesis.

The logarithmic factor in the regret comes from Term B. The upper bound proposed in Lemma 1.10 below allows to replace the self normalized quantity $U_{a,N_a(t),t}$ by the quantity $U_{a,s,T}$. This trick, introduced by [Garivier and Cappé, 2011], will be applied several times throughout this thesis, and leads to Lemma 1.10.

Lemma 1.10. *If the index $U_{a,s,t}$ is nondecreasing in t,*

$$\sum_{t=K}^{T} \mathbb{P}_\nu\left(A_t = a, U_{a,N_a(t),t} > \mu_1\right) \leq \sum_{s=1}^{T} \mathbb{P}_\nu\left(U_{a,s,T} > \mu_1\right);$$

Proof. The trick consists in upper bounding $U_{a,s,t}$ by $U_{a,s,T}$ and noting that only one term in the sum $\sum_{t=s}^{T} \mathbb{1}_{(A_t=a,N_a(t)=s)}$ can be different from zero. This sum can then be upper bounded by 1. More precisely, one writes

$$\sum_{t=K}^{T} \mathbb{P}_\nu \left(A_t = a, U_a(t) > \mu_1\right) = \mathbb{E}_\nu \left[\sum_{t=K}^{T} \sum_{s=1}^{t} \mathbb{1}_{(A_t=a,N_a(t)=s)} \mathbb{1}_{(U_{a,s,t}>\mu_1)} \right]$$
$$\leq \mathbb{E}_\nu \left[\sum_{s=1}^{T} \mathbb{1}_{(U_{a,s,T}>\mu_1)} \sum_{t=s}^{T} \mathbb{1}_{(A_t=a,N_a(t)=s)} \right] \leq \sum_{s=1}^{T} \mathbb{P}_\nu \left(U_{a,s,T} > \mu_1\right).$$

□

It remains to upper bound the right hand side of the inequality in Lemma 1.10. As $U_{a,s,T}$ is an upper bound on the mean $\mu_a < \mu_1$, for s larger enough, we will be able to show (with a Chernoff-type inequality for the UCB1 or KL-UCB algorithms) that the probability $\mathbb{P}\left(U_{a,s,T} > \mu_1\right)$ is very small. The critical value of s is of order $C_{1,a} \log(T)$, where $C_{1,a}$ is a constant that depends on the distribution of arms 1 and a and on the choice of upper bounds.

1.3 The Bayesian approach

Bandit models were introduced in a Bayesian framework to model medical trials with two possible treatments. As this particular case was considered in an overwhelming majority of the first bandit papers, we start by presenting Bayesian Bernoulli bandit models. Treatment 1 (arm 1) has a probability of success p, and treatment 2 (arm 2) has a probability of success q. One assume that (p, q) is drawn from a prior distribution with density $H(p, q)$ with respect to some reference measure. [Thompson, 1933] focuses on a particular case of product prior $H(p, q) = F(p)G(q)$, in which F and G are Beta distributions. He proposes a randomized approach, in which each arm is drawn according to its posterior probability of being optimal. In the two papers [Thompson, 1933, Thompson, 1935], the author focuses on the explicit computation of these posterior probabilities, but no analysis or performance study of the algorithm is proposed. More than 70 years will go by without advances in the study of this first bandit algorithm. Chapter 3 is focused on Thompson Sampling. We will provide therein bibliographic details from recent studies of this strategy and above all our own proof that this algorithm is asymptotically optimal in a frequentist sense.

Bandit problems were considered later in the work of [Robbins, 1952], who state the bandit problem in a frequentist setting. [Bradt et al., 1956] and [Bellman, 1956] study a particular Bayesian bandit model, in which the parameter of one of the arms is assumed to be known: $q = q_0$ is fixed, while p is drawn from a prior distribution $F(p)$. The joint prior distribution is thus of the form $H(p, q) = F(p)\delta_{q_0}(q)$. For a fixed horizon T, [Bradt et al., 1956] show that there exists a strategy minimizing the Bayes risk, that depends at time t on the remaining time to play $n = T - t$ and on the current posterior distribution F on p. The unknown arm is played if and only is some index $G(F, n)$ is larger than the mean q_0 of the known arm.

[Bellman, 1956] considers the same Bayesian bandit model in which one arm is known, but rather studies a *discounted bandit problem*. For a fixed $\alpha \in]0, 1[$, called the *discount factor*, the goal is to maximize the expectation of the sum of discounted rewards, defined by

$$\mathbb{E}_H \left[\sum_{t=1}^{\infty} \alpha^{t-1} X_t \right].$$

He shows that the optimal strategy for this problem is solution of some dynamic programming equations. The theory of dynamic programming was at the time still under development (see [Bellman, 1954]). Quite similarly to the undiscounted case, the optimal strategy plays the unknown arm if and only if some index is larger that q_0. In this case the index depends on the current posterior distribution F and on the discount factor α.

More general bandit problems were considered thereafter, with more than two arms, general rewards distributions, and even different objectives (e.g. maximizing the expected discounted rewards, where the sequence of discount is not necessary geometric). They are presented by [Berry and Fristedt, 1985], who give an overview of the bandit literature (mostly Bayesian) up to the 1980's. The two objectives considered above are notably presented: maximizing the expected sum of rewards up to some horizon T (i.e. minimizing the Bayes risk according to our definition) and maximizing the expected sum of discounted rewards (i.e. solving the discounted bandit problem). [Berry and Fristedt, 1985] explain that any such bandit problem appears as the solution of a dynamic programming equation. We present here this associated dynamic programming equation using elements from the theory of Markov Decision Processes (see e.g. [Puterman, 1994, Sigaud and Buffet, 2008]), developed more recently. A Markov Decision Process (MDP) is a 4-tuple $(\mathcal{X}, \mathcal{A}, \mathcal{P}, \mathcal{R})$ where $\mathcal{X}$ is a state space, $\mathcal{A}$ is an action space, $\mathcal{P} : \mathcal{X} \times \mathcal{A} \to \mathcal{M}_1(\mathcal{X})$ is a transition kernel and $\mathcal{R} : \mathcal{X} \times \mathcal{A} \to \mathcal{M}_1(\mathbb{R})$ is a reward kernel. When the agent is in state $x \in \mathcal{X}$ and chooses action $a \in \mathcal{A}$, a transition occurs: the agent receives a reward $r \sim \mathcal{R}(\cdot \mid x, a)$ and his new state is $y \sim \mathcal{P}(\cdot|x, a)$.

The interaction of an agent with a bandit model in a Bayesian framework can be modeled by the following MDP. Let $\boldsymbol{\theta}$ be drawn from a prior distribution Π_0.

- the current state is the current posterior distribution Π_t on $\boldsymbol{\theta}$: $\mathcal{X} \subset \mathcal{M}_1(\Theta^K)$
- there are K actions, corresponding to the draw of each arm: $\mathcal{A} = \{1, \ldots, K\}$
- when the agent chooses arm a in state Π, he observes a draw from arm a, $x \sim \nu_{\theta_a}$, receives the reward x, and computes the new posterior distribution Π' obtained by taking into account the new observation x. We let t_a^x be the operator such that $\Pi' = t_a^x(\Pi)$.

The agent's sampling strategy corresponds to a policy in this MDP. A deterministic policy g is a mapping that indicates which action is chosen in a state x_t and at time t. If (X_t^g) is a sequence of successive rewards obtained with g, one can consider the *value function* of this policy. Depending on the criterion considered (finite horizon or α-discounted rewards), value functions are defined by

$$V^g(\Pi, T) = \mathbb{E}_\Pi\left[\sum_{t=1}^{T} X_t^g\right] \quad \text{and} \quad V_\alpha^g(\Pi) = \mathbb{E}_\Pi\left[\sum_{t=1}^{\infty} \alpha^{t-1} X_t^g\right].$$

Bayes risk minimization is equivalent to solving the *planning problem* in the above MDP with a finite-horizon criterion, that is finding a policy maximizing $V^g(\Pi, T)$. Similarly, the discounted bandit problem corresponds to the planning problem in the same MDP but with a discounted criterion.

From the the theory of MDPs, there exists an optimal policy g^* such that its value function V^* satisfies a dynamic programming equation. For the finite horizon criterion, for all $k = 1, \ldots, T-1$, if Π^a denotes the a^{th} marginal distribution of Π, the optimal value function satisfies

$$\begin{aligned} V^*(\Pi, k) &= \max_{a=1\ldots K} \left(\mathbb{E}_{\theta_a \sim \Pi^a}\left[\mu(\theta_a)\right] + \mathbb{E}_{\substack{W \sim f(\cdot|\theta_a) \\ \theta_a \sim \Pi^a}}\left[V^*(t_a^W(\Pi_a), k-1)\right] \right) \\ V^*(\Pi, 0) &= 0 \end{aligned} \tag{1.15}$$

The optimal value function can thus be computed by induction and the optimal policy in state Π if the

remaining time to play is k chooses an action $g^*(\Pi, k)$ that realizes the maximum in (1.15). The optimal strategy for the Bayesian bandit problem with finite horizon is therefore $A_t = g^*(\Pi_{t-1}, T-t+1)$.

For the discounted criterion, the optimal value function is solution of the following equation:

$$V_\alpha^*(\Pi) = \max_{a=1\dots K} \left(\mathbb{E}_{\theta_a \sim \Pi^a}[\mu(\theta_a)] + \alpha \mathbb{E}_{\substack{W \sim f(\cdot|\theta_a) \\ \theta_a \sim \Pi^a}} \left[V_\alpha^*(t_a^W(\Pi)) \right] \right).$$

Due to the potentially very large state space, usual techniques to compute the optimal policy (value or policy iteration, induction when the horizon is finite) are often intractable. However, in some particular cases, a more practical description of the optimal policy can be obtained.

The solution of the Bayesian discounted or undiscounted bandit problem strongly depends on the prior distribution Π_0. [Feldman, 1962] manages to exhibit this solution, for a finite horizon T, in a very special case of two-armed binary bandit model. The prior distribution he considers captures the fact that the means of the arms $p_1 > p_2$ are known: $\mathbb{P}((p,q) = (p_1,p_2)) = \zeta$ and $\mathbb{P}((p,q) = (p_2,p_1)) = 1-\zeta$. If ζ_t denotes the posterior probability of the event $\{(p,q) = (p_1,p_2)\}$ at the end of round t, [Feldman, 1962] shows that the optimal policy chooses arm 1 at time t if and only if $\zeta_{t-1} > 1/2$. This *myopic* strategy chooses at each time the arm with highest posterior mean reward. It is not true that such a strategy is in general optimal.

In this particular example, the marginal distributions of each arm are very correlated. On the contrary, [Gittins, 1979] considers prior distributions with independent arms and shows that the solution of the discounted bandit problem reduces to an index policy. The computation of the Gittins index for each arm also resorts to dynamic programming but for a reduced state space. In Section 1.3.2, we define both the discounted and undiscounted Gittins indices (or Finite-Horizon Gittins indices) and we discuss the optimality of the associated index policies in Section 1.3.3. Unlike what happens in the discounted case, the index policy associated to the Finite-Horizon Gittins indices does not coincide with the Bayesian solution of the bandit problem with a finite horizon. However we will show that the FH-Gittins algorithm performs well in practice, and that approximations of the FH-Gittins indices presented in Section 1.3.4 show similarities with asymptotic approximations of the optimal strategy, discussed in Section 1.3.5.

Notation. Gittins indices are defined in Bayesian bandit models with independent arms, that we consider in the rest of this chapter. In this case, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ is drawn from a product prior $\Pi_0 = (\pi_1^0, \dots, \pi_K^0)$ such that π_a^0 is the prior distribution on θ_a, and the distributions $(\pi_a^0)_{a=1\dots K}$ are independent. The posterior distribution on $\boldsymbol{\theta}$ after t observations is thus a product prior,

$$\Pi_t = (\pi_1^t, \dots, \pi_K^t),$$

with the following notation:

- $\pi_{a,s}$ is the posterior distributions on θ_a after the first s observations of arm a:

$$\pi_{a,s} = \mathcal{L}(\theta_a | Y_{a,1}, \dots, Y_{a,s}).$$

- $\pi_a^t = \pi_{a,N_a(t)}$ is the posterior distribution on θ_a at the end of round t

1.3.1 Some examples of Bayesian bandit models.

In the particular case of exponential bandit models, introduced in Section 1.2.2, we now give examples of prior distributions that can be used, as well as a more explicit presentation of the associated

Distribution	Mean	Prior distribution on the mean	Posterior distribution on the mean after k observations whose sum is s
$\mathcal{B}(\mu)$	μ	Beta(a,b)	Beta$(a+s, b+k-s)$
$\mathcal{N}(\mu,\sigma^2)$	μ	$\mathcal{N}(\mu_0, m_0^{-1})$	$\mathcal{N}\left(\frac{m_0\mu_0+\sigma^{-2}s}{m_0+k\sigma^{-2}}, (m_0+k\sigma^{-2})^{-1}\right)$
$\mathcal{P}(\lambda)$	λ	$\Gamma(c,d)$	$\Gamma(c+s, d+n)$
$\mathcal{E}(\lambda)$	$1/\lambda$	Inv$\Gamma(c,d)$	Inv$\Gamma(c+n, d+s)$
$\Gamma(k,\lambda)$	k/λ	Inv$\Gamma(c,d)$	Inv$\Gamma(c+kn, d+ks)$

Table 1.4: Conjugate prior on the mean and associated posterior distributions.

Markov Decision Process in the Bernoulli case. Recall that the distribution of arm a conditionally to θ_a has a density of the form

$$f(x|\theta_a) = A(x)\exp(\theta_a x - b(\theta_a)).$$

A nice property of this family of distributions is that if the prior distribution π_a^0 has a density h_a^0, the density of the posterior distribution on θ_a, $\pi_{a,k}$ takes the following simple parametric form:

$$p(\theta_a|Y_{a,1},\ldots,Y_{a,k}) \propto \exp\left(\theta_a \sum_{i=1}^{k} Y_{a,i} - kb(\theta_a)\right) h_a^0(\theta_a). \tag{1.16}$$

This distribution can be parameterized by two sufficient statistics, $\left(k, s = \sum_{i=1}^{k} Y_{a,k}\right)$: the number of observations and the sum of observations. The current posterior distribution on $\boldsymbol{\theta}$ could therefore be described by a state $S = \{(k^a, s^a)\}_{a=1}^K \in (\mathbb{N}\times\mathbb{R})^K$.

Distributions that form an exponential family are often more simply parameterized by their means than by their natural parameter θ. In many examples of practical interest, it is possible to choose a prior distribution on the mean (and no longer on the natural parameter, as it is the case above) that belongs to a family of conjugate priors. A family of conjugate priors is such that if the prior distribution belongs to that family, the same goes for all the associated posterior distributions. Table 1.4 gives examples of conjugate priors on the mean for several examples of distributions in an exponential families. The posterior distribution, that depends on the number of observations k and on the sum of observations s is also computed. Most of the densities of the different distributions involved are defined in Table 1.1. The inverse Gamma distribution Inv$\Gamma(c,d)$ is the distribution of the random variable $1/X$ if $X \sim \Gamma(c,d)$.

In the Bernoulli case, a natural prior distribution on the mean is a Beta distribution, denoted by Beta(a,b). This distributions has a bounded support in $[0,1]$ and a density

$$f_{(a,b)}(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}\mathbb{1}_{[0,1]}(x).$$

In a Bernoulli bandit model with a Beta(a,b) prior on each mean, the posterior distribution on the mean of arm a at the end of round t is π_a^t = Beta$(S_a(t)+a, N_a(t)+b)$. One often considers the uniform prior $\mathcal{U}([0,1])$ which correspond to Beta$(1,1)$. The history of a Bernoulli bandit game with Beta prior is summarized by a matrix of Beta posteriors, $S = \{(A^a, B^a)\}_{a=1}^K$, that evolves as a state in the following Markov Decision Process:

- state $S = \{(A^a, B^a)\}_{a=1}^K \in (\mathbb{N}\times\mathbb{N})^K$
- actions $\{1...K\}$

– if the current state is $S = \{(A^a, B^a)\}_{a=1}^K$, and action $A_t = a$ is chosen, a binary reward $X_t = X_{a,t}$ is drawn from $\mathcal{B}(\mu_a)$ and the state (i.e. the posterior distribution) is updated in the following way:

$$\begin{aligned} A^a &\leftarrow A^a + X_t \\ B^a &\leftarrow B^a + (1 - X_t). \end{aligned}$$

Figure 1.2 gives an illustration of a transition in this MDP when there are three arms.

$$\begin{pmatrix} 1 & 2 \\ 5 & 1 \\ 0 & 2 \end{pmatrix} \xrightarrow{A_t=2} \begin{cases} \begin{pmatrix} 1 & 2 \\ 6 & 1 \\ 0 & 2 \end{pmatrix} & \text{if } X_{t,2} = 1 \\ \begin{pmatrix} 1 & 2 \\ 5 & 2 \\ 0 & 2 \end{pmatrix} & \text{if } X_{t,2} = 0 \end{cases}$$

Figure 1.2: Transition in the MDP associated to a Bernoulli bandit model with Beta prior

When there are two arms, the current posterior distribution can be written in a 4-tuple (a, b, c, d), with (a, b) (resp. (c, d)) the parameters of the current Beta posterior on the mean of arm 1 (resp. arm 2). The dynamic programming equation (1.15) rewrites in this particular case

$$\begin{aligned} V^*((a,b,c,d),n) \;=\; \max\Big\{ & \frac{a}{a+b} + \frac{a}{a+b} V^*((a+1,b,c,d),n-1) + \frac{b}{a+b} V^*((a,b+1,c,d),n-1); \\ & \frac{c}{c+d} + \frac{c}{c+d} V^*((a,b,c+1,d),n-1) + \frac{d}{c+d} V^*((a,b,c,d+1),n-1) \Big\} \end{aligned} \quad (1.17)$$

and the optimal policy in state (a, b, c, d), with remaining time to play n, chooses arm 1 if the first argument is the maximum, arm 2 otherwise. For T not too large, the decisions of the optimal policy, $g^*((a, b, c, d), n) \in \{1, 2\}$ can be computed by induction and stored, for elements $((a, b, c, d), n)$ such that $a + b + c + d \leq n + \alpha + \beta$ and $n \leq T$ (α and β being the parameters of the prior). This permits to implement the optimal policy for small values of T. When there are more arms, the exact solution becomes even less tractable.

1.3.2 Discounted and Finite-Horizon Gittins indices

Gittins' theorem, stated in the seminal paper [Gittins, 1979], says that an index policy, based on the (later) so-called Gittins indices, is a solution to several sequential allocation problems, when the objective is to maximize the expected sum of discounted rewards. This larger class of problems can be interpreted as planning problems in Markov Decision Processes with particular structure and is presented in the book [Gittins et al., 2011]. Here we introduce Gittins' indices only in the context of (Bayesian) multi-armed bandit models with independent arms.

A possible definition of the Gittins' indices (like the one given by [Berry and Fristedt, 1985]) relies on the introduction of a calibration problem for each arm, that we call $\mathcal{C}_\lambda$. This calibration problem is sometimes called 'one-armed bandit problem'.

Let $\theta \sim \pi$ and, conditionally to θ, let (X_t) be an i.i.d. sequence with distribution ν_θ, that is a one-armed bandit. For $\lambda \in \mathbb{R}$, one considers the following game, denoted by $\mathcal{C}_\lambda$. A each time t, an

agent can choose between receiving a (known) reward λ or drawing the unknown arm and receiving a random reward drawn from ν_θ. His goal is to maximize his rewards with respect to one of the criteria previously considered: either the sum of rewards up to horizon T, or the sum of discounted rewards with a discount factor α. This game can be naturally expressed as a planning problem in a MDP, and writing the associated dynamic programming equation, one can show that the optimal policy is a *stopping policy*: the unknown arm is played until some stopping time τ after which the reward λ is chosen until the end of the game. When the value of λ gets larger, the player has less incentive to play the unknown arm. There exists a critical value λ^* such that for larger value of λ, the optimal strategy in $\mathcal{C}_\lambda$ never draws the unknown arm (and always chooses λ). This critical value, which represents the price worth paying for playing the arm, is the Gittins index.

In the discounted one-armed bandit, the optimal policy plays the unknown arm as long as the current posterior distribution π on the parameter θ is such that $G_\alpha(\pi) > \lambda$, with $G_\alpha(\pi)$ the *Gittins index*, defined in the following way.

Definition 1.11. *The (discounted) Gittins index for the current posterior distribution π is*

$$G_\alpha(\pi) = \inf\left\{\lambda \in \mathbb{R} : \sup_{\tau \geq 0} \mathbb{E}_\pi\left[\sum_{t=1}^{\tau} \alpha^{t-1} X_t + \frac{\alpha^\tau \lambda}{1-\alpha}\right] = \frac{\lambda}{1-\alpha}\right\},$$

where the supremum is taken over the set of stopping times τ, with the convention $\sum_{t=1}^{0} = 0$.

For Bernoulli bandit models, the calibration problem $\mathcal{C}_\lambda$ was already solved by [Bellman, 1956] in the discounted case. In the case of a finite horizon T, one can also generalize the solution proposed by [Bradt et al., 1956] in the binary case. In the finite-horizon one-armed bandit, the optimal policy plays the unknown arm as long as the current posterior distribution π and the remaining time to play n are such that $G(\pi, n) > \lambda$, where $G(\pi, n)$ is the *Finite-Horizon Gittins index*, defined in the following way.

Definition 1.12. *The Finite-Horizon Gittins index for a current posterior π and remaining time n is*

$$G(\pi, n) = \inf\left\{\lambda \in \mathbb{R} : \sup_{0 \leq \tau \leq n} \mathbb{E}_\pi\left[\sum_{t=1}^{\tau} X_t + \lambda(n - \tau)\right] = n\lambda\right\}$$

where the supremum is taken over the set of stopping times τ smaller than n almost surely, with the convention $\sum_{t=1}^{0} = 0$.

These *Dynamic Allocation Indices*, later called Gittins indices, were defined differently in Gittins' original paper (in the discounted case only). It is not difficult to show (see e.g. [Gittins et al., 2011]) that the following two equalities hold, showing as a by product that Gittins' definition coincides with Definition 1.11:

$$G_\alpha(\pi) = \sup_{\tau > 0} \frac{\mathbb{E}_\pi\left[\sum_{t=1}^{\tau} \alpha^{\tau-1} X_t\right]}{\mathbb{E}_\pi\left[\sum_{t=1}^{\tau} \alpha^{\tau-1}\right]} \quad \text{and} \quad G(\pi, n) = \sup_{0 < \tau \leq n} \frac{\mathbb{E}_\pi\left[\sum_{t=1}^{\tau} X_t\right]}{\mathbb{E}_\pi[\tau]}. \tag{1.18}$$

In a multi-armed bandit problem, the Gittins index of an arm whose current posterior is π can be seen as the price worth paying to play this arm (the critical value of λ in the calibration problem for this arm), or the mean reward per unit of time it yields.

We mention here some properties of the FH-Gittins indices that easily follow from their definition and will be useful for the implementation of the associated index policy. First, from the expression

(1.18), the index is lower bounded by $\mathbb{E}_\pi\left[\sum_{t=1}^\tau X_t\right]/\mathbb{E}_\pi[\tau]$ for any stopping time τ. For $\tau = 1$ p.s., one obtains

$$G(\pi, n) \geq \mathbb{E}_{\substack{X\sim\nu_\theta \\ \theta\sim\pi}}[X], \tag{1.19}$$

which allows to interpret the Gittins index as some upper confidence bound on the (posterior) mean, providing an analogy with UCB-like algorithms presented in Section 1.2. Moreover, one can show that, for all $n \geq 2$,

$$G(\pi, n) \geq G(\pi, n-1). \tag{1.20}$$

To prove this, assume that there exists y such that $G(\pi, n) < y < G(\pi, n-1)$ and consider the calibration problem $\mathcal{C}_y$ with horizon n. As $G(\pi, n) < y$, the optimal policy starts by choosing y. But as $y < G(\pi, n-1)$, the next optimal action is to draw the unknown arm, which contradicts the fact that the optimal policy is a stopping policy.

Computation of the Gittins indices. Chapter 8 of [Gittins et al., 2011] addresses the computation of discounted Gittins indices in some examples of exponential family bandit models, whereas [Nino-Mora, 2011] discusses the computation of Finite-Horizon Gittins indices. Among the methods reviewed by the latter, the calibration method seems to perform well. This approaches solves the calibration problem $\mathcal{C}_\lambda$ for a grid of values of λ, and identify an approximation of the critical value λ^* above which the optimal policy in $\mathcal{C}_{\lambda^*}$ never plays the unknown arm.

We now discuss our implementation of the calibration method for computing the indices $G((a,b),n)$ in a Bernoulli bandit model with Beta prior ((a,b) are the parameters of the Beta posterior). One has

$$G((a,b),n) = \inf\{\lambda \in \mathbb{R} : V_\lambda^*((a,b),b) = n\lambda\} \quad \text{with} \quad V_\lambda^*((a,b),n) = \sup_{0\leq\tau\leq n} \mathbb{E}_{(a,b)}\left[\sum_{t=1}^\tau X_t + \lambda(n-\tau)\right].$$

As already noted, the optimal policy in the calibration problem $\mathcal{C}_\lambda$ is a stopping policies, and thus $V^*((a,b),n)$ is the optimal value function in this calibration problem. It satisfies the following dynamic programming equation

$$\begin{aligned} V_\lambda^*((a,b),0) &= 0 \quad \text{for all } (a,b) \\ V_\lambda^*((a,b),n) &= \max\left(\lambda n; \frac{a}{a+b} + \frac{a}{a+b}V_\lambda^*((a+1,b),n-1) + \frac{b}{a+b}V_\lambda^*((a,b+1),n-1)\right). \end{aligned}$$

$V_\lambda^*((a,b),n)$ can thus be computed by induction in $O(n^2)$ arithmetic operations. Rather than using a grid of values of λ or a dichotomic search to obtain an approximation of $G((a,b),n)$, we suggest to apply the secant method to find the first zero on $[0,1]$ of the convex function

$$Z(\lambda; a, b, n) := V_\lambda^*((a,b),n) - n\lambda = \sup_{0\leq\tau\leq n} \mathbb{E}_{(a,b)}\left[\sum_{t=1}^\tau (X_t - \lambda)\right].$$

An illustration is proposed in Figure 1.3. To initialize the algorithm, one needs two lower bounds on the Gittins index: from (1.19), the posterior mean $a/(a+b)$ is a lower bound on $G((a,b),n)$, and a second lower bound can be found using a dichotomic search.

Besides, for $n = 1, 2$, it is easy to obtain a close form expression for $V^*((a,b),n)$ and thus an explicit expression of the Gittins indices. The results are the following (and where already given by [Bradt et al., 1956]):

$$G((a,b),1) = \frac{a}{a+b} \quad \text{and} \quad G((a,b),2) = \frac{a}{a+b} \times \frac{1 + \frac{a+1}{a+b+1}}{1 + \frac{a}{a+b}}. \tag{1.21}$$

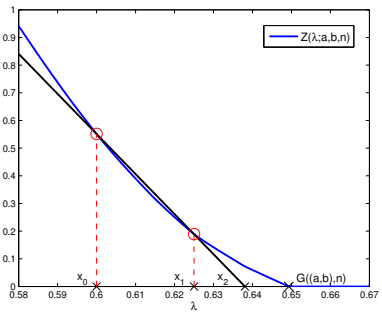

Figure 1.3: Illustration for $a = 15$, $b = 10$ and $n = 30$. The secant (in black) intersects the X axis at x_2, which is a new lower bound x_2 on the Gittins index $G((a,b),n)$.

These expressions illustrate the fact mentioned above that the Finite Horizon Gittins indices are upper bounds on the mean of the posterior distribution, $a/(a+b)$. Moreover, they suggest that these indices take the form

$$G((a,b),k) = \frac{a}{a+b} + B(a,b,k),$$

where the confidence bonus $B(a,b,k)$ seems to decrease when remaining time k decreases, unlike what happens with UCB indices. We will see in Section 1.3.4 that approximations of the Gittins indices indeed feature a different exploration rate than that used by UCB-like algorithms.

1.3.3 Index policies using Gittins indices

Gittins' theorem was presented for the first time by [Gittins and Jones, 1974], but was popularized by the paper [Gittins, 1979]. This result, particularized to bandit problems, is given here as the first statement of Theorem 1.13. The second statement of this theorem highlights the fact that in the undiscounted case, the index policy using the Finite-Horizon Gittins indices is no longer optimal. This is known since [Berry and Fristedt, 1985], who prove that a geometric discounting sequence is necessary for Gittins' theorem to hold (see Chapter 6 therein).

Theorem 1.13. *Let $\alpha > 1$. With an independent prior distribution Π_0, the strategy choosing at round t*

$$A_t = \underset{a=1\dots K}{\text{argmax}}\, G_\alpha(\pi_a^{t-1})$$

maximizes $\mathbb{E}_{\Pi_0}\left[\sum_{t=1}^{\infty} \alpha^{t-1} X_t\right]$.

Let $T \in \mathbb{N}^*$. There exists bandit models and independent prior distributions such that the strategy choosing at time t

$$A_t = \underset{a=1\dots K}{\text{argmax}}\, G(\pi_a^{t-1}, T-t+1)$$

does not maximize $\mathbb{E}_{\Pi_0}\left[\sum_{t=1}^{T} X_t\right]$.

Several proofs of Gittins' theorem have been proposed thereafter by different authors. Our own proof, provided in Section 1.5.2, relies on the prevailing charge argument introduced by [Weber, 1992]. More precisely, it is inspired by the version of this proof presented by [Frostig and Weiss, 1999]. After

proving the first statement of Theorem 1.13, we highlight which parts of the proof cannot be adapted to the finite-horizon setting. To prove the second statement of the theorem, we show that for some choices of independent Beta prior on two-armed Bernoulli bandits, for the horizon $T = 2$, the dynamic programming solution and Gittins' policy, that can both be computed in this simple case, do not coincide. To prove that the FH-Gittins index policy is not optimal, [Berry and Fristedt, 1985] also exhibit a counterexample in the class of Bernoulli bandit models, but with a prior distribution on each mean that takes two values.

From the second statement of Theorem 1.13, the Finite Horizon Gittins algorithm (FH-Gittins), that we define as the index policy using the Finite-Horizon Gittins indices, is not optimal. However, we conjecture that this algorithm is a good approximation of the Bayesian solution of the bandit problem with a finite horizon. To support this claim, we start by presenting a numerical comparison of FH-Gittins and the optimal solution, on a two-armed Bernoulli bandit problem, and for a small horizon $T = 70$ for which the optimal Dynamic Programming (DP) solution can indeed be computed.

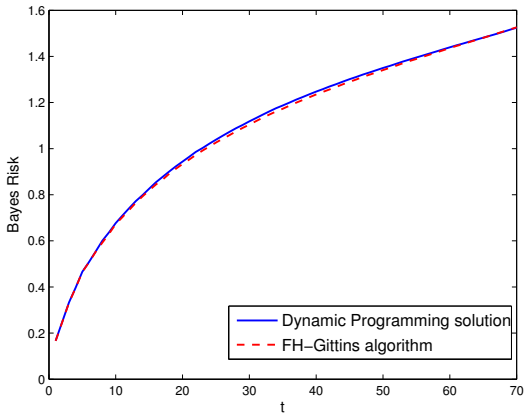

Figure 1.4: Bayes risk of the optimal strategy versus Bayes risk of FH-Gittins

For the FH-Gittins algorithm and the Dynamic Programming solution designed for the horizon $T = 70$ and a uniform prior Π_0, we present in Figure 1.4 the Bayes risk $\mathrm{BR}_{\Pi_0}(t, \mathcal{A})$ as a function of time t. To estimate the Bayes risk, we average over $N = 10^6$ draws of a bandit model from the prior distribution Π_0 on which each algorithm is played up to horizon $T = 70$. We see that indeed the Bayes-risk at horizon T is such that $\mathrm{BR}_{\Pi_0}(T, \mathrm{DP}) \leq \mathrm{BR}_{\Pi_0}(T, \text{FH-Gittins})$, but the difference is really small. Moreover, the Bayes risk of FH-Gittins is even smaller at round $t < T$. [Ginebra and Clayton, 1999] also propose a comparison of different algorithms with the Bayesian optimal solution for small horizon on which it can be computed. In particular, the authors similarly notice that the Bayes-risk at time T of the Finite-Horizon Gittins algorithm (that they call Λ-strategy) is very close to the optimal value, for various choices of prior and horizons.

Due to the costly implementation of the Bayesian optimal policy, it is important to develop good approximations of this strategy that are also efficient. FH-Gittins seems to perform well (based on numerical experiments) and might appears as an appealing alternative from a computational perspective, since the computation of each index require to solve (several) dynamic programming equations but on a much reduced state space. However, for large horizons, these repeated computations of indices also take time. To circumvent this issue, a first idea is to replace the FH-Gittins indices used in the algorithm by good approximations. In the next section, we present some approximations, that follow from the work

of [Chang and Lai, 1987] and [Burnetas and Katehakis, 2003], but only hold in an asymptotic regime in which the remaining time to play is very large. Interestingly, the Finite Horizon Gittins indices happen to be connected to the indices used by the KL-UCB algorithm, the (frequentist) asymptotically optimal algorithm discussed in Section 1.2.3. In Section 1.3.5 we will show that variants of KL-UCB, that are easier to implement than FH-Gittins, are also good approximations of the Bayesian optimal solution in an asymptotic sense.

1.3.4 Approximation of the FH-Gittins indices

[Chang and Lai, 1987] develop approximations of the classical (discounted) Gittins indices for Gaussian bandit models with a Gaussian prior distribution, when the discount factor α is close to 1. To do so, they approximate the solution of each calibration problem $\mathcal{C}_\lambda$ by the solution of an optimal stopping problem for a Brownian motion with drift, with some Gaussian prior distribution on the drift. They mention that a similar approach can be adopted to approximate the Finite-Horizon Gittins indices in the Gaussian case, that we explicit here.

In a Gaussian bandit model with known variance σ^2, with independent Gaussian prior $\mathcal{N}\left(0, \kappa^2\right)$ on each mean, the posterior distribution on the mean of arm a after t observations is

$$\mathcal{N}\left(\frac{S_a(t)}{N_a(t)+\sigma^2/\kappa^2}, \frac{\sigma^2}{N_a(t)+\sigma^2/\kappa^2}\right).$$

Thus each arm is characterized by the vector (u_a, v_a) of mean and variance of its current posterior distribution. The Gittins indices can be considered as functions of these two parameters. One has

$$\begin{aligned} G((u,v),n) &= \inf\left\{\lambda\in\mathbb{R} : \sup_{0\le\tau\le n} \mathbb{E}_{(u,v)}\left[\sum_{t=1}^{\tau}(X_t-\lambda)\right]=0\right\} \\ &= \inf\left\{\lambda\in\mathbb{R} : \sup_{0\le\tau\le n} \mathbb{E}_{(u-\lambda,v)}\left[\frac{\sum_{t=1}^{\tau}X_t}{\sigma\sqrt{n}}\right]=0\right\}, \end{aligned}$$

where the expectation $\mathbb{E}_{(u,v)}$ is taken under the model in which the X_t are i.i.d. with distribution $\mathcal{N}\left(\theta,\sigma^2\right)$ conditionally to θ and $\theta\sim\mathcal{N}(u,v)$. In order to introduce the discretization of a continuous optimal stopping problem, one introduces the notation

$$t'=\frac{t}{n},\quad u'_\lambda=\frac{(u-\lambda)\sqrt{n}}{\sigma},\quad v'=\frac{vn}{\sigma^2},\quad w_n(t')=\frac{\sum_{s=1}^{t}X_t}{\sigma\sqrt{n}},\quad \tau'=\frac{\tau}{n}\quad\text{and}\quad \mu=\frac{\theta\sqrt{n}}{\sigma}.$$

One has

$$w_n(t') \mid \mu\sim\mathcal{N}\left(\mu t', t'\right)\quad\text{and}\quad \mu\sim\mathcal{N}\left(u'_\lambda, v'\right) \tag{1.22}$$

thus , letting $\mathbb{E}'_{(u',v')}$ be the expectation under the model (1.22),

$$\sup_{0\le\tau\le n}\mathbb{E}_{(u-\lambda,v)}\left[\frac{\sum_{t=1}^{\tau}X_t}{\sqrt{n}}\right]=\sup_{\tau'\in\{0,\frac{1}{n},\ldots,\frac{n-1}{n},1\}}\mathbb{E}'_{(u'_\lambda,v')}\left[w_n(\tau')\right]$$

Letting B_t be a Brownian motion, and considering stopping times T with respect to B_t, the Finite-Horizon Gittins index can be approximated, for large values of n, as

$$G((u,v),n)\simeq\inf\left\{\lambda\in\mathbb{R} : \sup_{0\le T\le 1}\mathbb{E}_{\mu\sim\mathcal{N}(u'_\lambda,v')}\left[\mu T+B_T\right]=0\right\}.$$

[Chang and Lai, 1987] propose an asymptotic approximation of the solution of the continuous stopping problem involved that yields, when $v' \to \infty$,

$$T_\lambda^* \simeq \left\{ \inf t \in [0,T] : w(t) \leq -u'_\lambda v'^{-1} - \sqrt{2(t+v'^{-1}) \log\left(\frac{1}{t+v'^{-1}}\right)} \right\}.$$

Approximating the Gittins index by the value of λ such that $T_\lambda^* = 0$ and returning to the original values (u, v) yields

$$G((u,v),n) \simeq u + \sqrt{2v(\log(vn) + o(\log(vn))}.$$

In a bandit game, the index computed by the FH-Gittins algorithm for arm a at round $t+1$ can therefore be approximated by

$$\frac{S_a(t)}{N_a(t)+\sigma^2/\kappa^2} + \sqrt{\frac{2\sigma^2 \log\left(\frac{T-t}{N_a(t)+\sigma^2/\kappa^2}\right)}{N_a(t)+\sigma^2/\kappa^2}}. \tag{1.23}$$

This approximation holds when the remaining time $T-t$ is large and when the ratio $(T-t)/(N_a(t)+\kappa^{-2})$ is large too. [Chang and Lai, 1987] also propose approximations of the discounted Gittins indices in exponential families.

For one-parameter canonical exponential families, [Burnetas and Katehakis, 2003] propose a different approach to approximate the solution of the calibration problems $\mathcal{C}_\lambda$ with finite horizon, that leads to an approximation of the Finite Horizon Gittins indices. In an exponential bandit model, the distribution of each arm can be parameterized by a natural parameter θ_a such that, for every $\theta \in \Theta$, ν_θ has a density

$$f(x|\theta) = A(x)\exp(x\theta - b(\theta)), \quad \theta \subseteq \Theta \subseteq \mathbb{R}.$$

As explained in Section 1.3.1, given a product prior distribution Π_0, the density of the posterior distribution on the parameter of each arm has a simple form (1.16) and can be parameterized by the number of observations k from the arm and the sum of these observations, s. We denote it $H_{(k,s)}$. The FH-Gittins index of this arm also depends on these two parameters and can be written

$$G((k,s),n) = \inf\{\lambda \in T : V^*((k,s),n) = \lambda n\}, \text{ with } V^*((k,s),n) = \sup_{0\leq\tau\leq n} \mathbb{E}_{(k,s)}\left[\sum_{t=1}^{\tau} X_t + (n-\tau)\lambda\right].$$

The expectation $\mathbb{E}_{(k,s)}$ is taken under the model in which the X_t are i.i.d. with distribution ν_θ conditionally to θ and $\theta \sim H_{(k,s)}$. $V^*((k,s),n)$ is the value function in the MDP associated to $\mathcal{C}_\lambda$. The work of [Burnetas and Katehakis, 2003] relies on the following assumption on the prior distribution Π_0.

Assumption 1. Π_0 is supported on A^K, where $A =]\theta^-, \theta^+[\subseteq \Theta$ is such that there exists $m, M > 0$:

$$\forall\, \theta \in A \quad m \leq \ddot{b}(\theta) \leq M. \tag{1.24}$$

Under this assumption we let $]\mu^-, \mu^+[$ be the corresponding interval on the means: $]\mu^-, \mu^+[= \dot{b}(A)$.

The approximation of the solution of $\mathcal{C}_\lambda$ given by [Burnetas and Katehakis, 2003] relies on a careful rewriting of the dynamic programming equation for $V^*((k,s),n)$ (see 1.15). They exhibit, for the 'continuation set' (on which the optimal action in $\mathcal{C}_\lambda$ is to draw the unknown arm)

$$\mathcal{S}_n^\lambda = \{(k,s) : V^*((k,s),n) > n\lambda\},$$

two sets $\underline{\mathcal{S}_n^\lambda}$ and $\overline{\mathcal{S}_n^\lambda}$ such that $\underline{\mathcal{S}_n^\lambda} \subseteq \mathcal{S}_n^\lambda \subseteq \overline{\mathcal{S}_n^\lambda}$. The definition of these two sets involve integral expressions for which one can obtain asymptotic equivalents when the remaining time n goes to infinity. A close examination of Theorem 4.2 of [Burnetas and Katehakis, 2003] shows that, when n is large,

$$(V^*((k,s),n) > n\lambda) \quad \tilde{\Leftrightarrow} \quad \lambda \leq \max\left\{q \in [\mu^-,\mu^+] : q \geq \frac{s}{k}, k\tilde{d}_A\left(\frac{s}{k},q\right) \leq \log\frac{n}{k}\right\}$$

with $\tilde{d}_A(x,y)$ the function defined below. It depends on the divergence $d(x,y)$ associated to the exponential family and is defined on $(\dot{b}(\Theta))^2$ by

$$\tilde{d}_A(x,y) = \begin{cases} d(x,y) & \text{if } x > \mu^- \\ \frac{f(x|\theta^-)}{f(x|\dot{b}^{-1}(y))} & \text{if } x < \mu^- \end{cases}.$$

Thus, for large values of n, the Gittins index can be approximated as

$$G((k,s),n) \simeq \max\left\{q \in [\mu^-,\mu^+] : q \geq \frac{s}{k}, k\tilde{d}_A\left(\frac{s}{k},q\right) \leq \log\frac{n}{k}\right\}$$

and the index computed by the FH-Gittins algorithm for arm a at time $t+1$ is approximately, when $T-t$ is large,

$$\max\left\{q \in [\mu^-,\mu^+] : q \geq \frac{S_a(t)}{N_a(t)}, N_a(t)\tilde{d}_A\left(\frac{S_a(t)}{N_a(t)},q\right) \leq \log\left(\frac{T-t}{N_a(t)}\right)\right\}. \tag{1.25}$$

This approximation is consistent with the one obtained in the Gaussian case (1.23), for which the divergence function is $d(x,y) = (x-y)^2/(2\sigma^2)$. Both approximations of the index used by the FH-Gittins algorithm only hold when the remaining horizon is large. The indices (1.23) and (1.25) are reminiscent of the index used by the KL-UCB algorithm, but with a slightly different exploration rate: $\log t$ is replaced by $\log((T-t)/N_a(t))$. A natural question is therefore: does this modified exploration rate lead to improvements? As elements of answer, a quite similar modified version of KL-UCB will be proved in the next section to be a good approximation of the Bayesian solution, at least asymptotically. Besides, we will see in the numerical experiments of Section 1.4 that the factor $1/N_a(t)$ indeed leads to improvements.

1.3.5 Asymptotically optimal algorithms with respect to the Bayes risk

We saw that in the Bayesian framework, for a given prior distribution, the bandit problem with finite horizon T has an exact solution, that we denote by $\mathcal{A}_{\text{DP}}$ (since it is solution of a Dynamic Programming equation). However, we did not obtain an expression of the Bayes risk of this optimal solution, $\text{BR}_{\Pi_0}(T,\mathcal{A}_{\text{DP}})$. In the particular case of exponential bandit models, [Lai, 1987] provides a prior-dependent, asymptotic lower bound on the Bayes-risk of any bandit algorithm as well as an algorithm matching this bound. Hence, the Bayes risk of the Bayesian optimal strategy in such bandit models must grow at the rate $\log(T)^2$ specified by the lower bound of [Lai, 1987]. In this section, we see that the Bayes risk of the KL-UCB algorithm, discussed in Section 1.2.3 almost matches this lower bound (Proposition 15), and we discuss possible variants of this algorithm, one of them, KL-UCB-H$^+$, being asymptotically optimal with respect to the Bayes risk.

Theorem 1.14 below is a rewriting of Theorem 3 of [Lai, 1987] in the particular case of a product prior. It holds under an extra assumption on the prior distribution: Assumption 1, already used in the previous section. The prior distribution on each arm must be supported on some interval $A =]\theta_-,\theta^+[$,

on which the variance of the arms, $\ddot{b}(\theta)$, is bounded (see (1.24)). Recall that $]\mu^-, \mu^+[= \dot{b}(]\theta^-, \theta^+[)$. For Gaussian distributions with known variance, one can choose $A =]\mu^-, \mu^+[= \mathbb{R}$, whereas for Bernoulli distributions, this assumption is equivalent to considering that all the means belong to an interval of the form $]p, 1-p[$.

Theorem 1.14 ([Lai, 1987], Theorem 3)**.** *Let Π_0 be a distribution on A^K with a density of the form $h(\theta_1)\dots h(\theta_K)$. If h is such that $\int_A |\theta| q(\theta) d\theta < \infty$ and there exists $\rho > 0$ such that if h_{K-1} is the density of the random variable $\max_{1\le i\le K-1}\theta_i$,*

$$\int_A \sup_{\lambda\in]\theta-\rho;\theta]\cap A} h(\lambda) \times h_{K-1}(\theta) d\theta < \infty. \tag{1.26}$$

Let H be the cumulative distribution function (c.d.f) associated to h. For any bandit algorithm $\mathcal{A}$,

$$\liminf_{T\to\infty} \frac{\mathrm{BR}_{\Pi_0}(T,\mathcal{A})}{\log^2(T)} \ge C(\Pi_0, K)$$

with

$$\begin{aligned} C(\Pi_0, K) &= \frac{K}{2}\int_{A^{K-1}} h\left(\max_{1\le i\le K-1}\theta_i\right) h(\theta_1)\dots h(\theta_{K-1}) d\theta_1 \dots d\theta_{K-1} \\ &= \frac{K(K-1)}{2}\int_A h^2(\theta)(H(\theta))^{K-2} d\theta \end{aligned}$$

[Lai, 1987] also proposes an algorithm whose Bayes risk matches the lower bound of Theorem 1.14. In the sequel, we qualify such algorithms as *Bayesian asymptotically optimal* (or asymptotically optimal with respect to the Bayes risk). The algorithm proposed shares strong similarities with the KL-UCB algorithm presented in Section 1.2.3. Let d denote the divergence associated to the exponential family: $d(x,y) = \mathrm{KL}\left(\nu_{\dot{b}^{-1}(x)}, \nu_{\dot{b}^{-1}(y)}\right)$. One introduce the application $\overline{d}_A$ defined on $(\dot{b}(\Theta))^2$ by, for all y,

$$\overline{d}_A(x,y) = \begin{cases} d(x,y) & \text{if } x \in A \\ d(\mu^-, y) & \text{if } x \le \mu^- \\ d(\mu^+, y) & \text{if } x \ge \mu^+ \end{cases}.$$

After an initialization phase in which each arm is drawn once, the algorithm presented by [Lai, 1987] chooses at time $t+1$ the arm maximizing the index

$$U_a(t) = \sup\left\{q \in [\mu^-, \mu^+] : q \ge \frac{S_a(t)}{N_a(t)}, N_a(t)\overline{d}_A\left(\frac{S_a(t)}{N_a(t)}, q\right) \le g\left(\frac{T}{N_a(t)}\right)\right\},$$

for some function g satisfying $g(t) \sim \log t$ when $t \to \infty$ and $g(t) \ge \log t + \xi \log\log t$ for some $\xi > -3/2$. Apart from the fact that KL-UCB directly uses the divergence $d(x,y)$ in place of $\overline{d}_A(x,y)$, another difference between these two index policies is that the algorithm proposed by Lai uses an exploration rate $\log(T/N_a(t))$ in place of the $\log(t)$ used by KL-UCB.

This alternative exploration rate is also reminiscent of the MOSS algorithm of [Audibert and Bubeck, 2010], which is the index policy associated to

$$U_a(t) = \frac{S_a(t)}{N_a(t)} + \sqrt{\frac{\log\left(\frac{T}{KN_a(t)}\right)}{N_a(t)}}.$$

In the same spirit, [Garivier and Cappé, 2011] propose the use of a variant of KL-UCB, called KL-UCB$^+$, in which the exploration rate $\log(t)$ is replaced by $\log(t/N_a(t))$. Another alternative that has been considered in the literature is the UCB-H algorithm (see [Audibert et al., 2009]) in which the $\log(t)$ in UCB1 is replaced by $\log(T)$. Inspired by all these different exploration rates, Definition 1.15 below gathers the corresponding variants of KL-UCB that can be considered, namely KL-UCB$^+$, KL-UCB-H and KL-UCB-H$^+$.

Definition 1.15. *The KL-UCB, KL-UCB$^+$, KL-UCB-H and KL-UCB-H$^+$ algorithms are index policies that choose at time $t+1$ the arm with highest index*

$$u_a(t) = \sup\left\{q > \frac{S_a(t)}{N_a(t)} : d\left(\frac{S_a(t)}{N_a(t)}, q\right) \leq \beta(t)\right\},$$

where the exploration rate $\beta(t)$ is given by

Algorithm	*KL-UCB*	*KL-UCB$^+$*	*KL-UCB-H*	*KL-UCB-H$^+$*
$\beta(t) =$	$f(t)$	$f(t/N_a(t))$	$f(T)$	$f(T/N_a(t))$

with $f(t) = \log(t) + c\log\log(t)$, for some parameter c.

The algorithm proposed by [Lai, 1987] is very similar to KL-UCB-H$^+$ and is asymptotically optimal with respect to the Bayes risk. From (1.23) and (1.25), the indices used by the Finite-Horizon Gittins algorithm can be approximated in some asymptotic regime by indices close to those of KL-UCB-H$^+$. This gives heuristic arguments in favor of the Bayesian asymptotic optimality of FH-Gittins.

Interestingly, to prove the Bayesian asymptotic optimality of his algorithm, [Lai, 1987] starts by showing that it is asymptotically optimal in a frequentist sense, that is with respect to Lai and Robbins' lower bound on the regret (1.10). Then he integrates the asymptotic upper bound obtained over the prior distribution. Quite similarly, one can use the non-asymptotic upper bound on $\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)]$ for KL-UCB given in Theorem 1.8 to obtain the following Bayes-risk bound for KL-UCB. The proof of this new result is given in Section 1.5.3.

Theorem 1.16. *Let Π_0 be a product prior distribution satisfying the assumptions of Theorem 1.14. Then the Bayes risk of the KL-UCB satisfies*

$$\limsup_{T\to\infty} \frac{\mathrm{BR}_{\Pi_0}(T, \text{KL-UCB})}{\log(T)^2} \leq 2C(\Pi_0, K),$$

with $C(\Pi_0, K)$ the constant defined in Theorem 1.14.

Theorem 1.16 shows that the KL-UCB algorithm is almost asymptotically optimal with respect to the Bayes risk, up to a multiplicative factor 2. The fact that KL-UCB, which is optimal in a frequentist sense, might not be optimal in a Bayesian sense will be illustrated on numerical experiments below and in Chapter 3. Conversely, KL-UCB-H$^+$ is asymptotically optimal in both settings. Its asymptotic optimality with respect to the regret is already established by Lai, and can also be established using elements from our analysis of Bayes-UCB that will be presented in Chapter 2. Indeed, Proposition 2.4 will give the (frequentist) asymptotic optimality of both KL-UCB-H$^+$ and KL-UCB$^+$.

The notion of Bayesian asymptotic optimality introduced above, based on Theorem 1.14, might seem not very satisfying, since it applies to prior distributions that can be quite specific. For example, this notion is not defined in a Bernoulli bandit model with a uniform prior on the mean (whereas it is

defined if the prior is uniform on $[p, 1-p]$ for example). Even if the right constant is still to be identified in this context, the Bayes risk of the optimal strategy must also grow as $K\log(T)^2$. Indeed, an easy integration (see the precise computation in Section 1.5.3) shows that any algorithm (like UCB1) that satisfies, for every suboptimal arm a,

$$\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)] \leq \frac{C_1}{(\mu^* - \mu_a)^2}\log(T) + C_2$$

for some constants C_1 and C_2, has its Bayes risk under the uniform prior on $[0,1]^K$ upper bounded by

$$\frac{C_1}{2}K(\log(T))^2 + C_1K(\log(T))^{3/2} + \left(C_2 + \frac{1}{2}\right)K. \tag{1.27}$$

The prior-independent worst-case lower bound on the Bayes risk given in (1.7) is therefore pessimistic in the case of Bernoulli bandit with independent prior distributions. Indeed, while there exists a prior distribution such that the Bayes risk is lower bounded by $\sqrt{KT}/20$, for Bernoulli bandits with independent, uniform prior on the means, the Bayes-risk is rather of order $K\log(T)^2$.

1.4 Numerical study and conclusions

In both the frequentist and Bayesian frameworks, state-of-the-art algorithms presented in the previous sections are mostly index policies. Our presentation was focused on exponential bandit models, for which, in the frequentist setting, the KL-UCB algorithm of [Cappé et al., 2013] is asymptotically optimal with respect to the regret. In the Bayesian framework, with independent arms, the Finite-Horizon Gittins algorithms (inspired by the first index policy proposed by [Gittins, 1979] for discounted rewards) is empirically close to the Bayesian optimal solution, albeit not optimal as in the discounted case. For large horizons, this index policy is however difficult to implement and a variant of KL-UCB, KL-UCB-H$^+$ has been shown by [Lai, 1987] to be asymptotically optimal with respect to the Bayes risk.

We start by illustrating numerically the performance of these different index policies in the Bayesian framework, in order to motivate the use of the FH-Gittins algorithm when the horizon is not too large. In Figure 1.5, we compare FH-Gittins with KL-UCB, KL-UCB$^+$ and KL-UCB-H$^+$ for the horizon $T = 500$. For each strategy, the cumulated Bayes risk is averaged over $N = 5000$ bandit games played up to horizon T. We see that FH-Gittins dramatically outperforms KL-UCB-H$^+$, which is supposed to be a good approximation of the optimal strategy only for large values of T. Besides, KL-UCB-H$^+$ slightly outperform KL-UCB. While KL-UCB-H$^+$ requires the knowledge of the horizon T, KL-UCB$^+$ seems to be a good *anytime* (i.e. that does not use the horizon T) approximation of this algorithm. We carry out similar experiments, reported in Figure 1.6 and in which the Bayes risk is averaged over $N = 10000$ bandit games, for an horizon $T = 1000$. The same trends can be observed, with an increased gap between KL-UCB and KL-UCB-H$^+$.

By running numerical experiments, one can observe differences between the behavior of FH-Gittins and frequentist optimistic algorithms. FH-Gittins seems to explore much less than its frequentist counterpart: it ends up by playing always the same arm. From (1.19), Gittins indices can be seen as upper confidence bounds on the mean, but the 'confidence bonus' shrinks when getting closer to the horizon T. Besides, from (1.20), the FH-Gittins indices of arms that have not been played decrease, giving less incentive to exploration. Conversely, the KL-UCB indices of arms that have not been played increase in order to favor exploration. This property of FH-Gittins can be used in the implementation of the algorithm: if G denotes the value of the FH-Gittins index when a new arm starts to be played, the algorithm

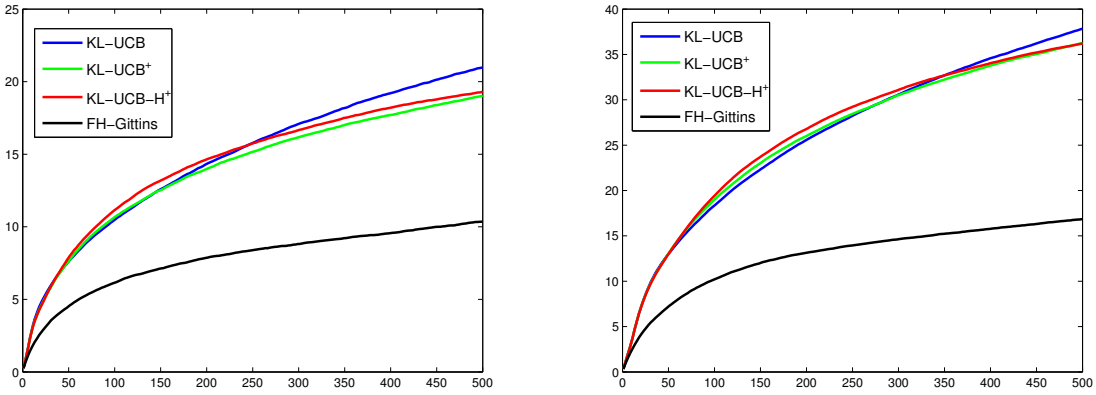

Figure 1.5: Bayes risk of the different strategies under a uniform prior distribution on the means, for K=5 arms (left) and K=10 arms (right)

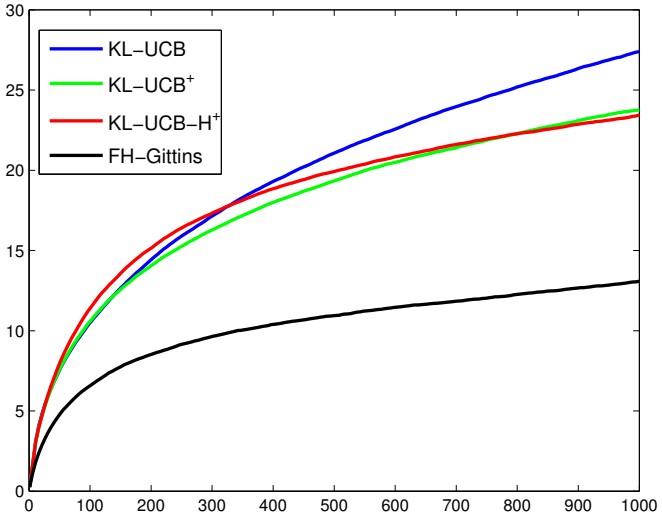

Figure 1.6: Bayes risk of the different strategies for a larger horizon $T = 1000$, with a uniform prior on the means of the $K = 5$ arms

will keep playing this same arm as long as its current FH-Gittins index is larger than G. Thus one can save computations of the indices of other arms during this time.

One of the objectives of this thesis is to present and analyse new efficient bandit algorithms, based on Bayesian ideas but whose performance is measured with (frequentist) regret. A first idea would be to investigate the performance of the Finite-Horizon Gittins algorithm, that is believed to be almost Bayesian optimal, in a frequentist setting. [Ginebra and Clayton, 1994] present a first empirical study for two-armed Bernoulli bandits in which two Bayesian strategies are evaluated with respect to regret, for fixed values of the means: the myopic strategy (that chooses the arm with highest posterior mean) and FH-Gittins (under the name Λ-strategy). The authors show that, for small horizons, there exist values of the means such that the regret of the Bayesian optimal solution is larger than that of these strategies. However, they do not compare FH-Gittins to algorithms that are known to be asymptotically optimal with respect to the regret. In Chapter 2 and Chapter 3, we will propose a numerical comparison of FH-Gittins with KL-UCB, Bayes-UCB and Thompson Sampling on some fixed Bernoulli bandit models. For the fixed bandit models chosen in these experiments, FH-Gittins seems to perform quite well on the (rather) small horizons on which it can be implemented, but its asymptotic optimality with respect to the regret is still to be investigated.

Due to the hardness of implementation of FH-Gittins, we will rather focus in the next two chapters on two Bayesian algorithms that are easy to implement: Bayes-UCB and Thompson Sampling. Moreover, additionally to their good practical performance, we will be able to prove that these Bayesian algorithms are asymptotically optimal with respect to the regret.

1.5 Elements of proof

This section gathers the proofs of some important results presented in this chapter.

1.5.1 Changes of distribution: proof of Lemma 1.3

As all the arms in $\nu = (\nu_1, \dots, \nu_K)$ and $\nu' = (\nu'_1, \dots, \nu'_K)$ are mutually absolutely continuous, there exists a common measure λ such that for all a, ν_a has a density f_a with respect to λ and ν'_a has a density f'_a with respect to λ. One can introduce the log-likelihood ratio of the observations up to time t under a bandit algorithm algorithm $\mathcal{A}$:

$$L_t := \sum_{a=1}^{K} \sum_{s=1}^{t} \mathbb{1}_{(A_s=a)} \log\left(\frac{f_a(X_s)}{f'_a(X_s)}\right).$$

The key element in a change of distribution is the following classical lemma that relates the probabilities of an event under $\mathbb{P}_\nu$ and $\mathbb{P}_{\nu'}$ through the log-likelihood ratio of the observations. Such a result is often used in the bandit literature for ν and ν' that differ just from one arm, for which the expression of the log-likelihood ratio is simpler. In the proof of Theorem 1.2, we indeed use this kind of change of distribution. However, we will consider in Chapter 5 changes of distributions where several arms are modified. A full proof of Lemma 1.17 in this more general setup can be found in the paper [Kaufmann et al., 2014b].

Lemma 1.17. *Let σ be any stopping time with respect to $\mathcal{F}_t$. For every event $A \in \mathcal{F}_\sigma$ (i.e. A such that $A \cap (\sigma = t) \in \mathcal{F}_t$),*

$$\mathbb{P}_{\nu'}(A) = \mathbb{E}_\nu[\mathbb{1}_A \exp(-L_\sigma)]$$

Let σ be a stopping time with respect to $(\mathcal{F}_t)$. We start by showing that for all $A \in \mathcal{F}_\sigma$, $\mathbb{P}_\nu(A) = 0$ if and only if $\mathbb{P}_{\nu'}(A) = 0$. Thus, if $0 < \mathbb{P}_\nu(A) < 1$ one also has $0 < \mathbb{P}_{\nu'}(A) < 1$ and the quantity $d(\mathbb{P}_\nu(A), \mathbb{P}_{\nu'}(A))$ in Lemma 1.3 is well defined. Let $A \in \mathcal{F}_\sigma$. Lemma 1.17 yields $\mathbb{P}_{\nu'}(A) = \mathbb{E}_\nu[\mathbb{1}_A \exp(-L_\sigma)]$. Thus $\mathbb{P}_{\nu'}(A) = 0$ implies $\mathbb{1}_A \exp(-L_\sigma) = 0$ $\mathbb{P}_\nu - a.s.$ As $\mathbb{P}_\nu(\sigma < +\infty) = 1$, $\mathbb{P}_\nu(\exp(L_\sigma) > 0) = 1$ and $\mathbb{P}_{\nu'}(A) = 0 \Rightarrow \mathbb{P}_\nu(A) = 0$. A similar reasoning yields $\mathbb{P}_\nu(A) = 0 \Rightarrow \mathbb{P}_{\nu'}(A) = 0$.

Let $A \in \mathcal{F}_\sigma$ be such that $0 < \mathbb{P}_\nu(A) < 1$ (then $0 < \mathbb{P}_{\nu'}(A) < 1$). Lemma 1.17 and the conditional Jensen inequality lead to

$$\begin{aligned}\mathbb{P}_{\nu'}(A) &= \mathbb{E}_\nu[\exp(-L_\sigma)\mathbb{1}_A] = \mathbb{E}_\nu[\mathbb{E}_\nu[\exp(-L_\sigma)|\mathbb{1}_A]\mathbb{1}_A] \\ &\geq \mathbb{E}_\nu[\exp\left(-\mathbb{E}_\nu[L_\sigma|\mathbb{1}_A]\right)\mathbb{1}_A] = \mathbb{E}_\nu[\exp\left(-\mathbb{E}_\nu[L_\sigma|A]\right)\mathbb{1}_A] \\ &= \exp\left(-\mathbb{E}_\nu[L_\sigma|A]\right)\mathbb{P}_\nu(A),\end{aligned}$$

Writing the same for the event $\overline{A}$ yields $\mathbb{P}_{\nu'}(\overline{A}) \geq \exp\left(-\mathbb{E}_\nu[L_\sigma|\overline{A}]\right)\mathbb{P}_\nu(\overline{A})$, hence

$$\mathbb{E}_\nu[L_\sigma|A] \geq \log\frac{\mathbb{P}_\nu(A)}{\mathbb{P}_{\nu'}(A)} \quad \text{and} \quad \mathbb{E}_\nu[L_\sigma|\overline{A}] \geq \log\frac{\mathbb{P}_\nu(\overline{A})}{\mathbb{P}_{\nu'}(\overline{A})}. \tag{1.28}$$

Therefore one can write

$$\begin{aligned}\mathbb{E}_\nu[L_\sigma] &= \mathbb{E}_\nu[L_\sigma|A]\mathbb{P}_\nu(A) + \mathbb{E}_\nu[L_\sigma|\overline{A}]\mathbb{P}_\nu(\overline{A}) \\ &\geq \mathbb{P}_\nu(A)\log\frac{\mathbb{P}_\nu(A)}{\mathbb{P}_{\nu'}(A)} + \mathbb{P}_\nu(\overline{A})\log\frac{\mathbb{P}_\nu(\overline{A})}{\mathbb{P}_{\nu'}(\overline{A})} = d(\mathbb{P}_\nu(A), \mathbb{P}_{\nu'}(A)).\end{aligned} \tag{1.29}$$

Introducing $(Y_{a,t})$, the sequence of i.i.d. samples successively observed from arm a, the log-likelihood ratio L_t can be rewritten

$$L_t = \sum_{a=1}^{K}\sum_{t=1}^{N_a(t)}\log\left(\frac{f_a(Y_{a,t})}{f'_a(Y_{a,t})}\right); \quad \text{and} \quad \mathbb{E}_\nu\left[\log\left(\frac{f_a(Y_{a,t})}{f'_a(Y_{a,t})}\right)\right] = \mathrm{KL}(\nu_a, \nu'_a).$$

Applying Wald's Lemma (see e.g. [Siegmund, 1985]) to $L_\sigma = \sum_{a=1}^{K}\sum_{t=1}^{N_a(\sigma)}\log\left(\frac{f_a(Y_{a,t})}{f'_a(Y_{a,t})}\right)$ yields

$$\mathbb{E}_\nu[L_\sigma] = \sum_{a=1}^{K}\mathbb{E}_\nu[N_a(\sigma)]\mathrm{KL}(\nu_a, \nu'_a). \tag{1.30}$$

Combining this equality with inequality (1.29) gives Lemma 1.3.

1.5.2 On Gittins' theorem: proof of Theorem 1.13

Proof of the first statement (Gittins' theorem). Let $\alpha \in]0,1[$ be the discount factor. For any bandit algorithm, one can introduce at round t, for each arm a

- the *fair charge* $g_a(t) = G_\alpha(\pi_a^t)$ as the Gittins index of arm a at the end round t (the denomination comes from its interpretation as the highest 'price' worth paying to play the arm),
- the *prevailing charge*, $\underline{g}_a(t) = \min_{1\leq v\leq t} g_a(v)$.

The proof goes as follows: we upper bound the expected sum of discounted rewards for any bandit algorithm by a quantity that involves the prevailing charge process of each arm, and that *does not depend on the algorithm*. Then we show that for Gittins' policy, this inequality is actually an equality.

Lemma 1.18. *For any bandit algorithm $\mathcal{A}$,*

$$\mathbb{E}_{\Pi_0}\left[\sum_{t=1}^{\infty}\alpha^{t-1}X_t\right] \leq \mathbb{E}_{\Pi_0}\left[\sum_{t=1}^{\infty}\alpha^{t-1}\underline{g}_{A_t}(t-1)\right]$$

and this inequality is an equality for Gittins policy.

Proof of Lemma 1.18. The result easily follows by summation if we prove that, for each arm a,

$$\mathbb{E}_{\Pi_0}\left[\sum_{t=1}^{\infty}\alpha^{t-1}X_{a,t}\mathbb{1}_{(A_t=a)}\right] \leq \mathbb{E}_{\Pi_0}\left[\sum_{t=1}^{\infty}\alpha^{t-1}\underline{g}_a(t-1)\mathbb{1}_{(A_t=a)}\right]. \tag{1.31}$$

To show (1.31) holds, we fix an arm a and introduce a sequence of stopping times $\tau_0 = 0$ and

$$\tau_{k+1} \quad = \quad \inf\left\{t > \tau_k \mid \underline{g}_a(t-1) > \underline{g}_a(t)\right\}$$

For $t \in [\tau_k + 1; \tau_{k+1}]$, the prevailing charge $\underline{g}_a(t-1)$ is equal to the Gittins index $g_a(\tau_k)$. Conditionally to $\mathcal{F}_{\tau_k}$, one can consider the calibration problem $\mathcal{C}_{g_a(\tau_k)}$, in which the optimal policy is either to play arm a until τ_{k+1} (i.e. as long as its Gittins index is larger that $g_a(\tau_k)$) or not to play at all and receive reward $g_a(\tau_k)$ at every time step: thus the expected cumulated reward is $g_a(\tau_k)/(1-\alpha)$. The inequality we write below follow from the fact that the policy that plays arm a when $(A_t = a)$ and $t \in [\tau_k + 1; \tau_{k+1}]$, and receives rewards $g_a(\tau_k)$ otherwise cannot be better than the optimal policy:

$$\begin{aligned}
\mathbb{E}_{\Pi_0}\left[\sum_{t=\tau_k+1}^{\tau_{k+1}}\alpha^{t-\tau_k-1}\left(X_{a,t}\mathbb{1}_{(A_t=a)} + g_a(\tau_k)\mathbb{1}_{(A_t\neq a)}\right) + \sum_{t=\tau_{k+1}+1}^{\infty} g_a(\tau_k)\alpha^{t-\tau_k-1}\middle|\mathcal{F}_{\tau_k}\right] &\leq \frac{g_a(\tau_k)}{1-\alpha}\\
\mathbb{E}_{\Pi_0}\left[\sum_{t=\tau_k+1}^{\tau_{k+1}}\alpha^{t-1}\left(X_{a,t}\mathbb{1}_{(A_t=a)} - g_a(\tau_k)\mathbb{1}_{(A_t=a)}\right) + \sum_{t=\tau_k+1}^{\infty} g_a(\tau_k)\alpha^{t-1}\middle|\mathcal{F}_{\tau_k}\right] &\leq \frac{\alpha^{\tau_k}g_a(\tau_k)}{1-\alpha}\\
\mathbb{E}_{\Pi_0}\left[\sum_{t=\tau_k+1}^{\tau_{k+1}}\alpha^{t-1}X_{a,t}\mathbb{1}_{(A_t=a)}\middle|\mathcal{F}_{\tau_k}\right] - \mathbb{E}_{\Pi_0}\left[\sum_{t=\tau_k+1}^{\tau_{k+1}}\alpha^{t-1}\underline{g}_a(t-1)\mathbb{1}_{(A_t=a)}\middle|\mathcal{F}_{\tau_k}\right] &\leq 0
\end{aligned}$$

Summing over k and conditioning gives inequality (1.31). For every arm a, if the bandit algorithm we consider is Gittins strategy, (1.31) is an equality. Indeed, on every interval $[\tau_k + 1, \tau_k]$, if Gittins' policy plays arm a at time $\tau_k + 1$, all the Gittins' indices from the other arms are smaller than $g_a(\tau_k)$, and arm a will therefore be played up to time τ_{k+1}, which coincides with an optimal policy in the calibration problem $\mathcal{C}_{g_a(\tau_k)}$.

□

From Lemma 1.18, one can write, introducing the sequence $\underline{\underline{g}}(t) = \underline{g}_{A_t}(t-1)$,

$$\mathbb{E}_{\Pi_0}\left[\sum_{t=1}^{\infty}\alpha^{t-1}X_t\right] \leq \mathbb{E}_{\Pi_0}\left[\sum_{t=1}^{\infty}\alpha^{t-1}\underline{\underline{g}}(t)\right] \leq \mathbb{E}_{\Pi_0}\left[\sum_{t=1}^{\infty}\alpha^{t-1}\underline{\underline{g}}^*(t)\right],$$

where $(\underline{\underline{g}}^*(t))$ is a nonincreasing rearrangement of $(\underline{\underline{g}}(t))$. Whereas the sequence $(\underline{\underline{g}}(t))_{t\in\mathbb{N}^*}$ depends on the algorithm, as explained below $(\underline{\underline{g}}^*(t))$ only depends on the sequences of successive rewards

obtained from each arm, and not on the algorithm itself any more. As Gittins policy is such that the sequence $(\underline{\underline{g}}(t))$ itself is already nonincreasing (since the arm chosen at time t maximizes $\underline{g}_a(t)$), one has

$$\mathbb{E}_{\Pi_0}\left[\sum_{t=1}^{\infty}\alpha^{t-1}X_t^{\text{Git.}}\right] = \mathbb{E}_{\Pi_0}\left[\sum_{t=1}^{\infty}\alpha^{t-1}\underline{\underline{g}}^*(t)\right],$$

which show that Gittins' policy is optimal.

To understand why $(\underline{\underline{g}}^*(t))$ does not depend on the policy, we introduce the following notation. For every t, $\underline{g}_a(t)$ only depends on the $N_a(t)$ first observations gathered from arm a, $Y_{a,1},\dots,Y_{a,N_a(t)}$. In other words, there exists a sequence of deterministic functions (h_s), with $h_s : \{0,1\}^s \to \mathbb{R}^+$, such that $\underline{g}_a(t) = h_{N_a(t)}(Y_{a,1},\dots,Y_{a,N_a(t)})$. For each arm a, and each $s \in \mathbb{N}^*$, we define the stopping time $\tau_{a,s}$ as the instant at which the s-th draw of arm a occurs (with the convention that $\tau_{a,s} = +\infty$ if arm a has been drawn less than s times). Rearranging the sum, one obtains

$$\sum_{t=1}^{\infty}\alpha^{t-1}\underline{\underline{g}}(t) = \sum_{a=1}^{K}\sum_{s=1}^{\infty}\alpha^{\tau_{a,s}-1}\underbrace{h_s(Y_{a,1},\dots,Y_{a,s-1})}_{:=h_{a,s}}$$

For each a, $(h_{a,s})$ is a nonincreasing sequence that does not depend on the algorithm (A_t). Building a nonincreasing rearrangement of $(\underline{\underline{g}}(t))$ is equivalent to sorting in nonincreasing order the $(h_{a,s})_{a=1,\dots,K,s\in\mathbb{N}^*}$. Letting $\tilde{\tau}_{a,s}$ be the ranking of (a,s) in this sorted list, one has

$$\sum_{t=1}^{\infty}\alpha^{t-1}\underline{\underline{g}}(t) = \sum_{a=1}^{K}\sum_{s=1}^{\infty}\alpha^{\tau_{a,s}-1}\underbrace{h_s(Y_{a,1},\dots,Y_{a,s-1})}_{:=h_{a,s}} \le \sum_{a=1}^{K}\sum_{s=1}^{\infty}\alpha^{\tilde{\tau}_{a,s}-1}h_{a,s} = \sum_{t=1}^{\infty}\alpha^{t-1}\underline{\underline{g}}^*(t).$$

As the sequence of stopping times $\tilde{\tau}_{a,s}$ only depends on the $(h_{a,s})$, and not on (A_t), the values of $\underline{\underline{g}}^*(t)$ are independent of the algorithm.

Why does this proof not work for Finite-Horizon Gittins indices? It is possible to generalize some of the arguments, by introducing

- $g_a(t) = G(\pi_a^t, T-t)$ as the *fair charge* at time t
- $\underline{g}_a(t) = \min_{1\le v\le t} g_a(v)$ as the *prevailing charge* at time t,

this time with finite-horizon Gittins indices. Referring to the same calibration problem as in the proof of Lemme 1.18 (but this time with a finite horizon), one can show an equivalent result:

$$\mathbb{E}_{\Pi_0}\left[\sum_{t=1}^{T}X_t\right] \le \mathbb{E}_{\Pi_0}\left[\sum_{t=1}^{T}\underline{g}_{A_t}(t-1)\right], \tag{1.32}$$

with equality for the FH-Gittins policy. However, both the fact the the horizon is finite and that $\underline{g}_a(t)$ depends on $Y_{a,1},\dots,Y_{a,N_a(t)}$ AND t do not permit to upper bound the quantity in (1.32) by something independent of the policy using some rearrangement argument.

Proof of the second statement. For Bernoulli bandit models with a Beta prior, the Finite-Horizon Gittins indices involved in the algorithm depend on the parameters (a,b) of the current posterior and the remaining time n. They are denotes by $G((a,b),n)$ as in Section 1.3.2.

If the remaining time to play is $T = 2$, and the prior distribution is $\mathrm{Beta}(5,8) \otimes \mathrm{Beta}(3,5)$, the explicit formula for Gittins indices available in this case (1.21) gives

$$G((5,8),2) = \frac{5}{13} \times \frac{1+\frac{6}{14}}{1+\frac{5}{13}} = 0.3968 \quad \text{and} \quad G((3,5),2) = \frac{3}{8} \times \frac{1+\frac{4}{9}}{1+\frac{3}{8}} = 0.3939,$$

thus the FH-Gittins algorithm chooses arm 1 (with highest FH-Gittins index).

With the same prior distribution, the dynamic programming equations (1.17) yield the following result for the optimal value function V^* (which depends on (a,b,c,d), the 4-tuple representing the posterior distribution on each arm, and the remaining time n)

$$\begin{aligned} V^*((5,8,3,5),2) &= \max\{V_1, V_2\} \text{ with} \\ V_1 &= \frac{5}{13} + \frac{5}{13}\max\left(\frac{6}{14}, \frac{3}{8}\right) + \frac{8}{13}\max\left(\frac{5}{14}, \frac{3}{8}\right) = 0.7802 \\ V_2 &= \frac{3}{8} + \frac{3}{8}\max\left(\frac{5}{13}, \frac{4}{9}\right) + \frac{5}{8}\max\left(\frac{5}{13}, \frac{3}{9}\right) = 0.7821 \end{aligned}$$

and as $V_1 < V_2$, the optimal policy (with value function V^*) chooses arm 2.

1.5.3 Proofs of Bayes risk bounds

PROOF OF THEOREM 1.16

The constant $C(\Pi_0, K)$ defined in Theorem 1.14 is such that

$$C(\Pi_0, K) = \frac{K}{2}\mathcal{I}_K \quad \text{with} \quad \mathcal{I}_K := \int_{A^{K-1}} h(\max_{i=2\dots K} \theta_i) h(\theta_2) \dots h(\theta_K) d\theta_2 \dots d\theta_K$$

The following decomposition of the Bayes risk can be deduced from the regret decomposition (1.1):

$$\mathrm{BR}_{\Pi_0}(T, \mathrm{KLUCB}) = \sum_{a=1}^{K} \mathbb{E}_{\Pi_0}\left[(\mu^* - \mu_a)\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)]\right] = K\mathbb{E}_{\Pi_0}\left[(\mu^* - \mu_1)\mathbb{E}_{\boldsymbol{\theta}}[N_1(T)]\right].$$

The last equality follow from the fact that the prior distribution is invariant under permutations of the arms. To compute the expectation $\mathbb{E}_{\Pi_0}\left[(\mu^* - \mu_1)\mathbb{E}_{\boldsymbol{\theta}}[N_1(T)]\right]$, we integrate $(\mu^* - \mu_1)\mathbb{E}_{\boldsymbol{\theta}}[N_1(T)]$ on different regions, on which it can either be upper bounded trivially by $(\mu^* - \mu_1)T$ or, if arm 1 is sub-optimal, by the upper bound given of Theorem 1.8, that be rewritten as a function of the natural parameters in the following form:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}[N_1(T)] \leq\ & \frac{1}{\mathrm{K}(\theta_1, \theta^*)}\log(T) + 2\sqrt{2\pi}\sqrt{M}\frac{(\theta^* - \theta_1)}{\mathrm{K}(\theta_1, \theta^*)^{3/2}}\sqrt{\log(T) + 3\log\log(T)} \\ & + \left(4e + \frac{3}{\mathrm{K}(\theta_1, \theta^*)}\right)\log\log(T) + 8\pi M\frac{(\theta^* - \theta_1)^2}{\mathrm{K}(\theta_1, \theta^*)^2} + 6, \end{aligned} \tag{1.33}$$

using additionally that the variance of the arms are bounded by $\sup \ddot{b}(\theta) \leq M$. We recall that $\mathrm{K}(\theta, \theta') = \mathrm{KL}(\nu_\theta, \nu_{\theta'})$ where the distributions are parameterized by their natural parameter. We also gather here some useful properties of exponential families. First, one has the two Taylor expansions

$$\mathrm{K}(\lambda, \theta) = \frac{\ddot{b}(\theta)}{2}(\theta - \lambda)^2 + o((\theta - \lambda)^2) \tag{1.34}$$

$$\mu(\theta) - \mu(\lambda) = \ddot{b}(\theta)(\theta - \lambda) + o(\theta - \lambda) \tag{1.35}$$

Using Taylor-Lagrange formula together with the assumption that on A, $m \leq \ddot{b}(\theta) \leq M$ also yields the inequalities

$$|\mu(\theta) - \mu(\lambda)| \leq M|\theta - \lambda| \quad \text{and} \quad K(\lambda, \theta) \geq \frac{m}{2}(\lambda - \theta)^2 \tag{1.36}$$

Decomposition of an integral. Let α_T, β_T be two nonincreasing sequences of real numbers, to be chosen later, such that $0 < \alpha_T < \beta_T$ and β_T tends to zero as T tends to infinity. One has

$$\begin{aligned}
\mathbb{E}_{\Pi_0}\left[(\mu^* - \mu_1)\mathbb{E}_{\boldsymbol{\theta}}[N_1(T)]\right] &= \int_{0<\theta^*-\theta_1<\alpha_T} T(\mu(\theta^*) - \mu(\theta_1))d\Pi_0(\boldsymbol{\theta}) && \text{(term A)} \\
&+ \int_{\alpha_T<\theta^*-\theta_1<\beta_T} (\mu(\theta^*) - \mu(\theta_1))\mathbb{E}_{\boldsymbol{\theta}}[N_1(T)]d\Pi_0(\boldsymbol{\theta}) && \text{(term B)} \\
&+ \int_{\beta_T<\theta^*-\theta_1} (\mu(\theta^*) - \mu(\theta_1))\mathbb{E}_{\boldsymbol{\theta}}[N_1(T)]d\Pi_0(\boldsymbol{\theta}) && \text{(term C)}
\end{aligned}$$

Upper bound on term A. Let $\theta_1^* = \max_{i\neq 1} \theta_i$. Term A is upper bounded by

$$\begin{aligned}
& MT \int_{0<\theta^*-\theta_1<\alpha_T} (\theta^* - \theta_1)h(\theta_1)\dots h(\theta_K)d\theta_1 \dots d\theta_K \\
= \quad & MT \int_{A^{K-1}} \left(\int_{\theta_1^*-\alpha_T}^{\theta_1^*} (\theta_1^* - \theta_1)h(\theta_1)d\theta_1 \right) h(\theta_2)\dots h(\theta_K)d\theta_2 \dots d\theta_K \\
\leq \quad & \frac{MT\alpha_T^2}{2} \int_{A^{K-1}} \sup_{\lambda\in[\theta_1^*-\alpha_T,\theta_1^*]} h(\lambda)h(\theta_2)\dots h(\theta_K)d\theta_2 \dots d\theta_K \\
\underset{T\to\infty}{\sim} \quad & \frac{MT\alpha_T^2}{2} \int_{A^{K-1}} h(\max_{i=2\dots K} \theta_i)h(\theta_2)\dots h(\theta_K)d\theta_2 \dots d\theta_K = \frac{MT\alpha_T^2}{2}\mathcal{I}_K.
\end{aligned}$$

The last equivalent follows from Assumption (1.26) - which allows to apply the dominated convergence theorem.

Upper bound on term B. To upper bound Term B, we use inequality (1.33). It then boils down to controlling the three integrals

$$\begin{aligned}
I_1(T) &= \int_{\alpha_T<\theta^*-\theta_1<\beta_T} \frac{\mu(\theta^*) - \mu(\theta_1)}{\mathrm{K}(\theta_1, \theta^*)} d\Pi_0(\boldsymbol{\theta}), \\
I_2(T) &= \int_{\alpha_T<\theta^*-\theta_1<\beta_T} \frac{(\mu(\theta^*) - \mu(\theta_1))(\theta^* - \theta_1)}{\mathrm{K}(\theta_1, \theta^*)^{3/2}} d\Pi_0(\boldsymbol{\theta}), \\
I_3(T) &= \int_{\alpha_T<\theta^*-\theta_1<\beta_T} \frac{(\mu(\theta^*) - \mu(\theta_1))^2(\theta^* - \theta_1)}{\mathrm{K}(\theta_1, \theta^*)^2} d\Pi_0(\boldsymbol{\theta}).
\end{aligned}$$

For $I_1(T)$, we use as [Lai, 1987], that the following equivalent holds, uniformly in θ,

$$\frac{\mu(\theta) - \mu(\theta_1)}{\mathrm{K}(\theta_1, \theta)} \underset{\theta_1\to\theta}{\sim} \frac{2}{\theta - \theta_1}.$$

The equivalent can be obtained using the Taylor expansions (1.34) and (1.35)). Therefore one can write

$$\begin{aligned}
I_1(T) &= \int_{A^{K-1}} \left(\int_{\theta_1^*-\beta_T}^{\theta_1^*-\alpha_T} \frac{\mu(\theta_1^*)-\mu(\theta_1)}{\mathrm{K}(\theta_1,\theta_1^*)} h(\theta_1) d\theta_1 \right) h(\theta_2)\dots h(\theta_K) d\theta_2 \dots d\theta_K \\
&\underset{T\to\infty}{\sim} \int_{A^{K-1}} \left(\int_{\theta_1^*-\beta_T}^{\theta_1^*-\alpha_T} \frac{2h(\theta_1)}{\theta_1^*-\theta_1} d\theta_1 \right) h(\theta_2)\dots h(\theta_K) d\theta_2 \dots d\theta_K \\
&\leq \int_{A^{K-1}} \left(\int_{\theta_1^*-\beta_T}^{\theta_1^*-\alpha_T} \frac{2}{\theta_1^*-\theta_1} d\theta_1 \right) \sup_{\lambda\in[\theta_1^*-\beta_T,\theta_1^*-\alpha_T]} h(\lambda) h(\theta_2)\dots h(\theta_K) d\theta_2 \dots d\theta_K \\
&= 2\log\left(\frac{\beta_T}{\alpha_T}\right) \int_{A^{K-1}} \sup_{\lambda\in[\theta_1^*-\beta_T,\theta_1^*-\alpha_T]} h(\lambda) h(\theta_2)\dots h(\theta_K) d\theta_2 \dots d\theta_K \\
&\underset{T\to\infty}{\sim} 2\log\left(\frac{\beta_T}{\alpha_T}\right) \mathcal{I}_K
\end{aligned}$$

Using inequalities (1.36), one can also show that

$$\begin{aligned}
I_2(T) &\leq \frac{2^{2/3}M}{m^{2/3}} \int_{\alpha_T<\theta^*-\theta_1<\beta_T} \frac{1}{\theta_1^*-\theta_1} d\Pi_0(\theta) \underset{T\to\infty}{\sim} \frac{2^{2/3}M}{m^{2/3}} \log\left(\frac{\beta_T}{\alpha_T}\right) \mathcal{I}_K, \\
I_3(T) &\leq \frac{4M^2}{m^2} \int_{\alpha_T<\theta^*-\theta_1<\beta_T} \frac{1}{\theta_1^*-\theta_1} d\Pi_0(\theta) \underset{T\to\infty}{\sim} \frac{4M^2}{m^2} \log\left(\frac{\beta_T}{\alpha_T}\right) \mathcal{I}_K.
\end{aligned}$$

These estimations of $I_1(T), I_2(T)$ and $I_3(T)$ together with the upper bound on $\mathbb{E}_{\boldsymbol{\theta}}[N_1(T)]$ given by (1.33) leads to

$$\int_{\alpha_T<\theta^*-\theta_1<\beta_T} (\mu(\theta^*)-\mu(\theta_1)) \mathbb{E}_{\boldsymbol{\theta}}[N_1(T)] d\Pi_0(\boldsymbol{\theta}) = 2\log\left(\frac{\beta_T}{\alpha_T}\right)\log(T)\mathcal{I}_K + 0\left(\log\left(\frac{\beta_T}{\alpha_T}\right)\sqrt{\log(T)}\right)$$

Upper bound on term C. To upper bound Term C, using again inequality (1.33), we would need to upper bound the same three integrals but on a different region:

$$\begin{aligned}
I_1(T) &= \int_{\theta^*-\theta_1>\beta_T} \frac{\mu(\theta^*)-\mu(\theta_1)}{\mathrm{K}(\theta_1,\theta^*)} d\Pi_0(\boldsymbol{\theta}), \\
I_2(T) &= \int_{\theta^*-\theta_1>\beta_T} \frac{(\mu(\theta^*)-\mu(\theta_1))(\theta^*-\theta_1)}{\mathrm{K}(\theta_1,\theta^*)^{3/2}} d\Pi_0(\boldsymbol{\theta}), \\
I_3(T) &= \int_{\theta^*-\theta_1>\beta_T} \frac{(\mu(\theta^*)-\mu(\theta_1))^2(\theta^*-\theta_1)}{\mathrm{K}(\theta_1,\theta^*)^2} d\Pi_0(\boldsymbol{\theta})
\end{aligned}$$

For $I_1(T)$, one can write, using inequality (1.36),

$$I_1(T) \leq \frac{M}{m} \int_{A^{K-1}} \left(\int_{\theta_1^*-\beta_T}^{\theta_1^*-\alpha_T} \frac{2h(\theta_1)}{\theta_1^*-\theta_1} d\theta_1 \right) h(\theta_2)\dots h(\theta_K) d\theta_2 \dots d\theta_K \leq \frac{2M}{m}\frac{1}{\beta_T}$$

Similarly, one obtains

$$I_2(T) \leq \frac{2^{3/2}M}{m^{2/3}}\frac{1}{\beta_T} \quad \text{and} \quad I_3(T) \leq \frac{4M^2}{m^2}\frac{1}{\beta_T}.$$

This shows that C is of order $O\left(\frac{\log(T)}{\beta_T}\right)$.

Choice of α_T and β_T and conclusion. The following tabular summarizes what we proved above.

term A	term B	term C
$O(T\alpha_T^2)$	$2\log\left(\frac{\beta_T}{\alpha_T}\right)\log(T)\mathcal{I}_K + o\left(\log\left(\frac{\beta_T}{\alpha_T}\right)\log(T)\right)$	$O\left(\frac{\log(T)}{\beta_T}\right)$

Choosing $\alpha_T = \frac{1}{\sqrt{T}\sqrt{\log(T)}}$ and $\beta_T = \frac{1}{\sqrt{\log(T)}}$, term B is the leading term and one obtains

$$\mathbb{E}_{\Pi_0}\left[(\mu^* - \mu_a)\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)]\right] = \log(T)^2\mathcal{I}_K + o(\log(T)^2).$$

Thus

$$\mathrm{BR}_{\Pi_0}(T, \mathrm{KLUCB}) = K\mathcal{I}_K\log(T)^2 + o(\log(T)^2) = 2C(\Pi_0, K)\log(T)^2 + o(\log(T)^2),$$

which concludes the proof.

PROOF IF THE BAYES RISK BOUND (1.27)

Let $\mathcal{A}$ be an algorithm that satisfies, on every Bernoulli bandit model parameterized by $\boldsymbol{\theta}$, for every sub-optimal arm a,

$$\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)] \leq \frac{C_1}{(\mu^* - \mu_a)^2}\log(T) + C_2.$$

Let Π_0 be the uniform prior on the means. One has

$$\mathrm{BR}_{\Pi_0}(T, \mathcal{A}) = \sum_{a=1}^{K}\mathbb{E}_{\Pi_0}\left[(\mu^* - \mu_a)\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)]\right]$$

and for every a, letting $\mu_a^* = \max_{i\neq a}\mu_i$, one can write

$$\begin{aligned}
\mathbb{E}_{\Pi_0}\left[(\mu^* - \mu_a)\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)]\right] &= \int_{0<\mu^*-\mu_a<\frac{1}{\sqrt{T}}}(\mu^* - \mu_a)T d\mu_1 \,.\, d\mu_K \\
&\quad + \int_{\frac{1}{\sqrt{T}}<\mu^*-\mu_a<\frac{1}{\sqrt{\log(T)}}}\frac{C_1\log(T)}{(\mu^* - \mu_a)}d\mu_1 \,.\, d\mu_K \\
&\quad + \int_{\mu^*-\mu_a>\frac{1}{\sqrt{\log(T)}}}\frac{C_1\log(T)}{(\mu^* - \mu_a)}d\mu_1 \,.\, d\mu_K + C_2 \\
&= T\int_{[0,1]^{K-1}}\left(\int_{\mu_a^*-\frac{1}{\sqrt{T}}}^{\mu_a^*}(\mu_a^* - \mu_a)d\mu_a\right)d\mu_1 \,.\, d\mu_{a-1}d\mu_{a+1} \,.\, d\mu_K \\
&\quad + \int_{[0,1]^{K-1}}\left(\int_{\mu_a^*-\frac{1}{\sqrt{\log(T)}}}^{\mu_a^*-\frac{1}{\sqrt{T}}}\frac{C_1\log(T)}{\mu_a^* - \mu_a}d\mu_a\right)d\mu_1 \,.\, d\mu_{a-1}d\mu_{a+1} \,.\, d\mu_K \\
&\quad + C_1\sqrt{\log(T)}\log(T) + C_2 \\
&= T\frac{1}{2}\left(\frac{1}{\sqrt{T}}\right)^2 + C_1\log(T)\log\left(\frac{\sqrt{T}}{\sqrt{\log(T)}}\right) + C_1\sqrt{\log(T)}\log(T) + C_2 \\
&\leq \frac{C_1}{2}(\log(T))^2 + C_1(\log(T))^{3/2} + C_2 + \frac{1}{2}.
\end{aligned}$$

□

Chapter 2

Bayes-UCB

In this chapter, we introduce and analyse the Bayes-UCB algorithm. We show that this algorithm can be applied in many contexts and shares strong similarities with existing frequentist algorithms. For Bernoulli bandits models, we propose a finite-time analysis proving its asymptotic optimality with respect to Lai and Robbins' lower bound on the regret. Bayes-UCB has been the object of a paper with Olivier Cappé and Aurélien Garivier for the conference AISTATS in 2012 ([Kaufmann et al., 2012a]) of which the content of this chapter is largely inspired.

Compared to the original paper, the statement of the Bayes-UCB algorithm does no longer depend on the horizon T and the finite-time analysis has been slightly improved: following the last refinements of [Cappé et al., 2013] for KL-UCB, we are now able to give a fully explicit upper bound on the expected number of draws of each suboptimal arm. Moreover, we point out that our finite-time analysis of Bayes-UCB can be adapted to show that two variants of KL-UCB presented in Chapter 1, KL-UCB$^+$ and KL-UCB-H$^+$, are asymptotically optimal.

Contents

2.1 Introduction

As discussed in the previous chapter, the literature on stochastic multi-armed bandit problems is separated in two distinct approaches. In the frequentist view, the expected mean rewards associated with each arm is considered as unknown deterministic quantity and the goal of the algorithm is to achieve the best parameter-dependent performance. In contrast, in the Bayesian view, each arm is characterized by a parameter which is endowed with a prior distribution. The Bayesian performance is then defined as the average performance over all possible problem instances weighted by the prior on the parameters. Here we argue that algorithms derived from the Bayesian perspective also prove efficient when evaluated using frequentist measures of performance.

Recall that in the classical (frequentist) parametric stochastic multi-armed bandit model, an agent faces K independent arms which depend on unknown parameters $\theta_1, \dots, \theta_K \in \Theta$. The draw of arm a at time t results in a reward X_t that is extracted from an i.i.d sequence $(X_{a,t})_{t\geq 1}$ marginally distributed under ν_{θ_a}, whose expectation is denoted by μ_a. The agent sequentially draws the arms according to a strategy $\mathcal{A} = (A_t)_{t\geq 1}$, where A_t denotes the arm chosen at round t, based on previous rewards $X_s = X_{A_s,s}$ for $1 \leq s \leq t-1$. The agent's goal is to find a strategy that maximizes the expected cumulated reward until time T, or equivalently minimizes the cumulated regret

$$\mathrm{R}_\theta(T, \mathcal{A}) = \mathbb{E}_\theta\left[\sum_{t=1}^{T} \mu^* - \mu_{A_t}\right] = \sum_{a=1}^{K} (\mu^* - \mu_a)\mathbb{E}_\theta[N_a(T)]\,, \tag{2.1}$$

where $\mu^* = \max\{\mu_a : 1 \leq a \leq K\}$ and $N_a(t)$ denotes the number of draws of arm a up to time t.

[Lai and Robbins, 1985], followed by [Burnetas and Katehakis, 1996], have provided lower bounds on the number of suboptimal draws under any uniformly efficient strategy: for any arm a such that $\mu_a < \mu^*$,

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\theta[N_a(T)]}{\log(T)} \geq \frac{1}{\inf_{\theta\in\Theta:\mu(\theta)>\mu^*} \mathrm{KL}(\nu_{\theta_j}, \nu_\theta)}. \tag{2.2}$$

For important classes of distributions, recent contributions have provided finite-time analysis of strategies that are asymptotically optimal in so far that they reach this lower bound. Among them, the KL-UCB algorithm of [Cappé et al., 2013], that uses confidence intervals based on Kullback Leibler divergence, has been proved optimal in one-parameter exponential family bandit models.

When considering the multi-armed bandit model from a Bayesian point of view, one assumes that the parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)$ is drawn from a prior distribution Π^0. More precisely, we assume in the following that the parameters $(\theta_a)_{1\leq a\leq K}$ are drawn independently from prior distributions $(\pi_a^0)_{1\leq a\leq K}$ (usually chosen to be all equal), and that conditionally on $(\theta_a)_{1\leq a\leq K}$, the sequences $(X_{1,t})_{t\geq 1}, \dots, (X_{K,t})_{t\geq 1}$ are jointly independent and i.i.d. with marginal distributions $\nu_{\theta_1}, \dots, \nu_{\theta_K}$.

In this Bayesian setting, the goal is to maximize $\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$, where the expectation is relative to the entire probabilistic model, including the randomization over θ. Bayesian optimality can equivalently be measured considering the Bayes risk, $\mathrm{BR}_{\Pi^0}(T, \mathcal{A}) = \mathbb{E}[\mathrm{R}_\theta(T, \mathcal{A})]$, that averages the regret over the parameters. A major appeal of the Bayesian framework is the fact that a strategy with minimal Bayesian regret can be described: it appears as the solution of a planning problem in an associated Markov Decision Process.

To define a Bayesian strategy, let Π^t denote the posterior distribution of θ after t rounds of game. Due to our choice of independent priors on $(\theta_a)_{1\leq a\leq K}$, Π^t is a product distribution which is equivalently

defined by the marginal posterior distributions $\pi_1^t, ..., \pi_K^t$ on $\theta_1, \ldots, \theta_K$, after t rounds. If at round t one chooses arm $A_t = a$ and consequently observes $X_t = X_{a,t}$, the Bayesian update for arm a is

$$\pi_a^t(\theta_a) \propto \nu_{\theta_a}(X_t)\, \pi_a^{t-1}(\theta_a)\,, \tag{2.3}$$

whereas for $i \neq a$, $\pi_i^t = \pi_i^{t-1}$. A Bayesian algorithm is allowed to exploit the knowledge of the posterior Π^t to determine the next action A_{t+1}.

We already presented in Chapter 1 two Bayesian algorithms: the index policies associated to the Gittins indices and Finite-Horizon Gittins indices. Recall that [Gittins, 1979] considers the infinite-horizon discounted problem in which one tries to maximize $\mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^t X_t\right]$, where $0 < \alpha < 1$ is a real discount parameter. He shows that the index policy associated to the Gittins indices (that depend on α) is optimal with respect to this alternative performance criterion. However, the model reduction argument he uses no longer holds when the horizon T is known and there is no discount. Therefore, the FH-Gittins algorithm does not coincide with the Bayesian optimal policy. This being said, we gave in Chapter 1 some arguments indicating that FH-Gittins should be very close to the Bayesian optimal solution. Here we go further and report in Section 2.4 some experiments on Bernoulli bandits that illustrate our finding that FH-Gittins outperforms its frequentist UCB-like competitors on their own ground, that is, when evaluated using the parameter-dependent (frequentist) regret.

[Lai, 1987] shows that a variant of the frequentist optimal KL-UCB algorithm, KL-UCB-H$^+$, is also asymptotically optimal with respect to the Bayes risk (see Theorem 1.14). Conversely, our finding that a (believed) close-to-optimal Bayesian strategy also achieves remarkable parameter-dependent performance for most (all?) value of the parameter θ is currently not supported by mathematical arguments. Furthermore, computing the finite-horizon variant of the Gittins indices is only feasible for moderate horizons due to the need to repeatedly perform (and store the results of) dynamic programming recursions on reduced models. Even for small horizons, the associated computational load and memory footprint are orders of magnitude larger than those of the UCB-like algorithms.

Our objective is thus to propose a generic bandit algorithm, termed Bayes-UCB, that is inspired by the Bayesian interpretation of the problem but retains the simplicity of UCB-like algorithms. Our hope is that this algorithm is simple enough to be effectively implemented and yet is able to reach the asymptotic lower bound of (2.2), including in cases that are currently not handled by UCB-like algorithms. In addition to promising simulation results reported in Section 2.4, we provide several significant elements that support our hopes. First, it is shown in Section 2.2 that instantiating the generic Bayes-UCB algorithm in different specific cases (one-parameter exponential families rewards, Gaussian-armed bandit with unknown means and variances, linear bandits, Gaussian process optimization) yields algorithms that share striking similarities with methods previously proposed in the literature. In the case of Bernoulli rewards, we provide in Section 2.3 a finite-time analysis of the Bayes-UCB algorithm that implies that it reaches the lower bound (2.2). The proof of this result is based on an interesting connection between Bayes-UCB and variants of the KL-UCB algorithm.

2.2 The Bayes-UCB algorithm

We start by presenting the rationale for the proposed algorithm before stating it more formally. First, being inspired by the Bayesian modeling of the bandit problem, the Bayes-UCB strategy is a function of the posteriors $(\pi_a^t)_{1\leq a\leq K}$. Due to the nature of our performance measure, the relevant aspect of θ_a is the expectation μ_a. Hence, denoting by λ_a^t, for $1 \leq a \leq K$, the posterior distribution of the mean μ_a induced by π_a^t, the proposed strategy is a function of $(\lambda_a^t)_{1\leq a\leq K}$ only.

Algorithm 1 Bayes-UCB

Require: Π^0 (initial prior on θ) c (parameters of the quantile)

1: **for** $t = 0$ **to** $T-1$ **do**
2: **for** each arm $a = 1, \dots, K$ **do**
3: compute

$$q_a(t) = Q\left(1 - \frac{1}{t(\log t)^c}, \lambda_a^t\right)$$

(with the convention $q_a(0) = 1, q_a(1) = 1$)
4: **end for**
5: draw an arm $A_{t+1} \in \arg\max_{a=1\dots K} q_a(t)$
6: get reward $X_{t+1} = X_{A_{t+1},t+1}$ and update Π^{t+1} according to (2.3)
7: **end for**

The use of fixed-level quantiles of $(\lambda_a^t)_{1\le a\le K}$ as confidence indices appears in [Pavlidis et al., 2008] as a special case of the Interval Estimation method. To be more specific, denote by $Q(t,\rho)$ the quantile function associated to the distribution ρ, such that $\mathbb{P}_\rho(X \le Q(t,\rho)) = t$. [Pavlidis et al., 2008] use indices of the form $Q(1-\alpha, \lambda_a^t)$ for $1 \le a \le K$, with α chosen to be equal to a few percents. In Bayes-UCB, we acknowledge the strong similarity between these posterior indices based on quantiles and the upper confidence bounds used in UCB and its variants: we consider indices of the form $Q(1-\alpha_t, \lambda_a^t)$, where α_t is of order $1/t$. As will be shown in Section 2.3 below for the case of binary rewards, this $1/t$ rate is deeply connected with the form of the upper confidence bounds used in variants of UCB that are known to reach the bound in (2.2). It is conjectured that no other rate can provide an algorithm that reaches the bound in (2.2) and that, furthermore, choices of the form $1/t^\beta$ with $\beta < 1$ do not even guarantee a finite-time logarithmic control of the regret. As a more pragmatic comment, we also observed in experiments not reported here that, in the case of binary rewards, the empirical performance of the method were superior when using $\alpha_t \equiv 1/t$. We are now ready to state the generic version of the Bayes-UCB algorithm.

In Algorithm 1, the horizon-dependent term $(\log t)^c$ is an artifact of the theoretical analysis that enables us, for $c \ge 5$, to both guarantee finite-time logarithmic regret bounds and achieve asymptotic optimality with respect to (2.2) in Bernoulli bandit models. But in simulations, the choice $c = 0$ actually proved to be the most satisfying. In cases where the prior Π^0 is chosen to correspond to an improper prior (see, e.g., the Gaussian models below), $q_a(t)$ is not defined when $t = 1$. In those cases it suffices, as is commonly done in most bandit algorithms, to make sure that initially one gathers a sufficient number of observations to guarantee that the posterior Π^t indeed becomes proper, for instance by drawing each arm a few times.

As such, Algorithm 1 corresponds to a general principle that does not even require the prior Π^0 to be chosen as a product distribution: Bayes-UCB can still compute one index for each arm, based on its marginal posterior distribution. In this case, the posterior update no longer reduces to (2.3) but is a global update on the joint distribution Π^t. In fact, the GP-UCB algorithm for Gaussian processes [Srinivas et al., 2010] can be seen as a variant of Bayes-UCB in which dependencies, in contrast, are of fundamental importance. This point will be later emphasized in Chapter 4, but here we mostly consider cases where the coordinates of $\boldsymbol{\theta}$ are independent. Implementing Algorithm 1 may require additional tools from the Bayesian computational toolbox to perform (or approximate) the Bayesian update of Π^t and/or to compute (or, again, approximate) the quantiles $q_a(t)$. We first discuss several important models for which Algorithm 1 corresponds to a a procedure that can be implemented exactly without the need to resort

to numerical approximation (an example of the opposite situation will be considered in Section 2.4.3 below).

Bayes-UCB for one-parameter exponential family bandits. In the case where the reward distributions belong to a one-parameter canonical exponential family, that is $\nu_{\theta_a}(x) = A(x)\exp(\theta_a x - b(\theta_a))$, with $\theta_a \in \mathbb{R}$, as recalled in Section 1.3.1, it is well known that the prior distribution π_a^0 can be chosen to belong to the conjugate family so that the posteriors π_a^t are all members of the same conjugate family, indexed by their sufficient statistics. As shown in Figure 1.4 of Chapter 1, in many cases of interest, the posterior distribution on the mean (λ_a^t) then belong to a well-known parametric family of distribution, for which the quantile $q_a(t)$ is easy to compute. For Bernoulli rewards, for instance, using the prior $\mathrm{Beta}(a,b)$ for the probability of observing a non-zero reward, we have $\pi_a^t = \mathrm{Beta}(a + S_a(t), b + N_a(t) - S_a(t))$, where $S_a(t) = \sum_{i=1}^t \mathbb{1}\{A_t = j\}X_t$ is the sum of rewards gathered up to time t. Likewise, for exponential rewards with a $\Gamma(c,d)$ prior on the parameter, $\pi_a^t = \Gamma(c + N_a(t), d + S_a(t))$ and $\lambda_a^t = \mathrm{Inv}\Gamma(c + N_a(t), d + S_a(t))$.

As will be proved below for binary rewards, the Bayes-UCB algorithm in that case is surprisingly related to the KL-UCB algorithm.

Bayes-UCB for Gaussian bandits with unknown means and variances. In general exponential family models, the Bayesian update is usually still computable explicitly (at least when using conjugate priors) but the relationship between the parameter θ_a and the expectation μ_a is less direct. A significant case where Bayes-UCB corresponds to a simple and efficient algorithm is when the rewards are assumed to be Gaussian, with both unknown mean μ_a and unknown variance σ_a^2. For simplicity, we consider improper non-informative priors on each arm, that is, $\pi_a^0(\mu_a, \sigma_a) = 1/\sigma_a^2$. It is well known that the marginal posterior distribution of μ_a at time t is then such that

$$\left.\frac{\mu_a - S_a(t)/N_a(t)}{\sqrt{S_a^{(2)}(t)/N_a(t)}}\right| X_1, ..., X_t \sim \mathcal{T}(N_a(t) - 1)\,,$$

where

$$S_a^{(2)}(t) = \frac{\left(\sum_{s=1}^t \mathbb{1}\{A_s = a\}X_s^2\right) - S_a^2(t)/N_a(t)}{N_a(t) - 1}\,,$$

and $\mathcal{T}(k)$ denote the Student-t distribution with k degrees of freedom. Therefore Bayes-UCB is the index policy associated to upper confidence bound

$$q_a(t) = \frac{S_a(t)}{N_a(t)} + \sqrt{\frac{S_a^{(2)}(t)}{N_a(t)}}Q\left(1 - \frac{1}{t}, \mathcal{T}(N_a(t) - 1)\right)\,,$$

omitting the $(\log t)^c$ factor for clarity. The Bayes-UCB index above is related to the index used in the UCB1-norm algorithm of [Auer et al., 2002a], where the quantile is replaced by $\sqrt{16\log(t-1)}$, which is obtained as an upper bound of $Q(1 - 1/t^4, \mathcal{T}(N_a(t) - 1))$. The practical performances of these two variants (Bayes-UCB and UCB1-norm) will be illustrated in Section 2.4 below.

Bayes-UCB for linear bandits. We end this section with the more elaborate case of linear bandits in which the arms can be numerous but share a strong common structure. We will consider the case of

Gaussian rewards with a multivariate Gaussian prior for the parameter $\theta \in \mathbb{R}^d$ that defines the model. The arms are fixed vectors $U_1, ..., U_K \in \mathbb{R}^d$. In this model, the choice of arm $A_t = a$ at time t results in the reward

$$y_t = U_a^T \theta + \sigma^2 \epsilon_t,$$

where ϵ_t is some centered noise and M^T denotes the transpose of a matrix (or vector) M. In the sequel, we assume that $\epsilon_t \sim \mathcal{N}(0,1)$. Following [Rusmevichientong and Tsitsiklis, 2010], our goal is to find algorithms $\mathcal{A}$ that minimize the frequentist regret

$$\mathrm{R}_\theta(T, \mathcal{A}) = \mathbb{E}_\theta \left[\sum_{t=1}^{T} \left(\max_{1 \le a \le K} (U_a^T \theta) - U_{A_t}^T \theta \right) \right].$$

Denoting by $Y_t = [y_1, ..., y_t]^T \in \mathbb{R}^d$ the vector of rewards and $X_t = [U_{A_1} ... U_{A_t}]^T \in \mathcal{M}_{t,d}(\mathbb{R})$ the design matrix, the model rewrites:

$$Y_t = X_t \theta + \sigma^2 E_t \ , \ \text{where} \ \ E_t \sim \mathcal{N}(0, \sigma^2 \mathrm{Id}_t) \, .$$

The Bayesian modeling here consists in a Gaussian $\mathcal{N}\left(0, \kappa^2 \mathrm{Id}_d\right)$ prior on θ, assuming the noise parameter σ^2 to be known. The posterior is

$$\theta | X_t, Y_t \ \sim \mathcal{N}\left(M_t, \Sigma_t\right) \ ,$$

where

$$M_t = (X_t^T X_t + (\sigma/\kappa)^2 \mathrm{I}_d)^{-1} X_t^T Y_t \quad \text{and} \quad \Sigma_t = \sigma^2 (X_t^T X_t + (\sigma/\kappa)^2 \mathrm{I}_d)^{-1} \, .$$

The posterior distribution λ_a^t on $\mu_a = U_a' \theta$ is therefore $\mathcal{N}(U_a^T M_t, U_a^T \Sigma_t U_a)$. Hence, using the notation where $\|x\|_A := \sqrt{x^T A x}$, Bayes-UCB selects an arm that maximizes the index:

$$q_a(t) = U_a^T M_t + \|U_a\|_{\Sigma_t} Q\left(1 - \frac{1}{t}, \mathcal{N}(0,1)\right).$$

[Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010, Abbasi-Yadkori et al., 2011] propose an optimistic approach for this problem based on a confidence ellipsoid located around the least-square estimate $\hat{\theta}_t$. This method is equivalent to choosing arm a such that $U_a \hat{\theta}_t + \rho(t) \|U_a\|_{(X_t^T X_t)^{-1}}$ is maximal. For an improper prior ($\kappa = \infty$), we have $M_t = \hat{\theta}_t$ and $\Sigma_t = \sigma^2 (X_t^T X_t)^{-1}$. Thus, this approach can again be interpreted as a particular case of Bayes-UCB. In Section 2.4.3, we consider the case where θ is a sparse vector. It is not obvious how to design an UCB algorithm for this case. Yet, we show that Bayes-UCB can be implemented, using for example Gibbs sampling.

In Chapter 4, we will consider contextual linear bandits, a more general framework that encompasses the (static) linear bandit introduced here. We will provide some theoretical guarantees for Bayes-UCB in this more general framework.

2.3 Analysis of the Bayes-UCB algorithm for binary rewards

In this section, we focus on the case where the rewards have a Bernoulli distribution, and when the prior distribution on each arm is the Beta$(1,1)$; that is the uniform distribution $\mathcal{U}([0,1])$. In this case, recall the Bayes-UCB algorithm is the index policy associated to

$$q_a(t) = Q\left(1 - \frac{1}{t(\log(t))^c}; \mathrm{Beta}(S_a(t) + 1, N_a(t) - S_a(t) + 1\right).$$

We show below that the Bayes-UCB algorithm is optimal, in the sense that it reaches the lower-bound (2.2) of Lai and Robbins.

2.3.1 Asymptotic optimality and links with frequentist algorithms

Theorem 2.1. *The Bayes-UCB algorithm with a uniform prior on the means and with the parameter $c = 5$ satisfies, for every $\epsilon > 0$ and for T such that*

$$\log T + 5\log\log T \geq \frac{d(\mu_2,\mu_1)}{1+\epsilon}\exp\left(\frac{8}{(\mu_1(1-\mu_1))^2}\frac{(1+\epsilon)^2}{\epsilon^2 d(\mu_2,\mu_1)^2}\right),$$

$$\mathbb{E}_{\boldsymbol{\theta}}[N_a(T)] \leq \frac{1+\epsilon}{d(\mu_a,\mu_1)}\log(T) + \sqrt{\log T + 5\log\log T}\sqrt{\frac{2\pi(1+\epsilon)^3 d'(\mu_a,\mu_1)^2}{d(\mu_a,\mu_1)^3}}$$
$$+\left(\frac{1+\epsilon}{d(\mu_a,\mu_1)} + \frac{2e+3}{1-\mu_1}\right)\log\log T + 27 + 2(1+\epsilon)^2\left(\frac{d'(\mu_a,\mu_1)}{d(\mu_a,\mu_1)}\right)^2.$$

The analysis relies on the following tight bounds we have on the index $q_a(t)$.

Lemma 2.2. *Denoting by $d(x,y)$ the KL divergence between Bernoulli distributions with parameters x and y, the posterior quantile $q_a(t)$ used by the Bayes-UCB algorithm with parameter $c = 5$ satisfies*

$$\tilde{u}_a(t) \leq q_a(t) \leq u_a(t)\ ,$$

where

$$u_a(t) = \operatorname*{argmax}_{x > \frac{S_a(t)}{N_a(t)}}\left\{d\left(\frac{S_a(t)}{N_a(t)}, x\right) \leq \frac{\log(t) + 5\log(\log(t))}{N_a(t)}\right\},$$

$$\tilde{u}_a(t) = \operatorname*{argmax}_{x > \frac{S_a(t)}{N_a(t)+1}}\left\{d\left(\frac{S_a(t)}{N_a(t)+1}, x\right) \leq \frac{\log\left(\frac{t}{N_a(t)+2}\right) + 5\log(\log(t))}{(N_a(t)+1)}\right\}.$$

Surprisingly, the Bayesian quantiles match the upper confidence bounds based on Kullback-Leibler divergence used by variants of the KL-UCB algorithm of [Cappé et al., 2013]. The quantile $q_a(t)$ is exactly upper bounded by the KL-UCB index $u_a(t)$ using the exploration rate $f(t) = \log t + 5\log\log t$, whereas it is upper bounded by a biased version of the index used by the KL-UCB$^+$ variant, that uses $f(t/N_a(t))$ as an exploration rate (see Definition 1.15 in Chapter 1).

Remark 2.3. *If Bayes-UCB were defined more generally depending on some exploration function $f(t)$ as the algorithm choosing at time $t+1$ the arm maximizing the index*

$$q_a(t) = Q\left(1 - e^{-f(t)}, \pi_a^t\right),$$

then following the proof of Lemma 2.2, one would have $q_a(t) \leq u_a(t)$ were $u_a(t)$ is the KL-UCB index with exploration function $f(t)$, i.e.

$$u_a(t) = \sup\{q \geq \hat{\mu}_a(t) : N_a(t)d(\hat{\mu}_a(t), q) \leq f(t)\}.$$

Thanks to this link with the KL-UCB indices, the finite-time analysis we provide share similarities with that of [Cappé et al., 2013]. The main difficulty is to deal with the bias and the alternative exploration rate that appear in the indices $\tilde{u}_a(t)$, since no finite-time analysis of KL-UCB$^+$ existed in the literature. Following the same lines as our analysis of Bayes-UCB, it is now possible to give a finite-time analysis for the two variants KL-UCB$^+$ and KL-UCB-H$^+$, leading to the following result.

Proposition 2.4. *For bandits with rewards in an exponential family with associated divergence $d(x,y) = KL(\nu_{\dot{b}^{-1}(x)}, \nu_{\dot{b}^{-1}(y)})$, the following instances of KL-UCB$^+$ and KL-UCB-H$^+$, defined as the index policies respectively associated to the index $u_a^+(t)$ and $u_a^{H,+}(t)$ given by*

$$u_a^+(t) = \sup\left\{q \geq \hat{\mu}_a(t) : N_a(t)d\left(\hat{\mu}_a(t), q\right) \leq \log\frac{t}{N_a(t)} + 5\log\log\frac{t}{N_a(t)}\right\}$$
$$u_a^{H,+}(t) = \sup\left\{q \geq \hat{\mu}_a(t) : N_a(t)d\left(\hat{\mu}_a(t), q\right) \leq \log\frac{T}{N_a(t)} + 5\log\log\frac{T}{N_a(t)}\right\}$$

are asymptotically optimal with respect to Lai and Robbins' lower bound.

2.3.2 Bayes-UCB beyond Bernoulli distributions

Albeit being designed for Bernoulli distributions, as explained by [Cappé et al., 2013], the KL-UCB algorithm using the divergence associated to Bernoulli rewards can also be used for rewards distributions bounded in $[0,1]$, without any modification and with the same theoretical guarantees.

For Bayes-UCB, a slight modification of the original algorithm, introduced by [Agrawal and Goyal, 2012] for Thompson Sampling (an other Bayesian algorithm studied in Chapter 3), yields a provably efficient algorithm for bandit models with bounded rewards. A bounded bandit model with means $\mu_1, \dots, \mu_K$ can be 'transformed' into a Bernoulli bandit model with same means by introducing for each arm the reward process $\tilde{X}_{a,t} = (U_{a,t} \leq X_{a,t})$ where $(X_{a,t})$ is the original reward process of arm a and $(U_{a,t})_{t\in\mathbb{N}^*}$ is an independent sequence of i.i.d. random variables uniform on $[0,1]$. The modified algorithm proceeds as follows: if arm a is drawn at time t and reward $X_t = X_{a,t}$ is observed, a transformed binary reward $\tilde{X}_t = (U_t \leq X_t)$ is collected, where the (U_t) are i.i.d. uniform random variables. Letting $\tilde{S}_a(t) = \sum_{s=1}^t \mathbb{1}_{(A_s=a)}\tilde{X}_s$, the arm chosen at time $t+1$ maximizes the index

$$\tilde{q}_a(t) = Q\left(1 - \frac{1}{t(\log t)^5}; \text{Beta}\left(\tilde{S}_a(t)+1, N_a(t) - \tilde{S}_a(t) + 1\right)\right). \tag{2.4}$$

Theorem 2.1 also holds for this algorithm, whereas there is no provable guarantee if the original Bayes-UCB algorithm designed for Bernoulli rewards is applied to bounded rewards. Indeed, the sum of 'true' rewards $S_a(t)$ is no longer an integer, and Lemma 2.2 on which our analysis is based would no longer hold (one can only control the tail —and thus the quantiles— of Beta distributions with integer coefficients).

The asymptotic optimality of Bayes-UCB is established only for Bernoulli distributions, but we believe this algorithm is also asymptotically optimal for other distributions in an exponential family. For Gaussian distributions with known variance σ^2, Bayes-UCB with a normal (or improper) prior on the mean can also be proved to be asymptotically optimal, since tight bounds can be obtained for the quantiles of the (Gaussian) posterior distribution (using for example Theorem 1.2.3. of [Durrett, 2010]). For Thompson Sampling, we also propose a finite-time analysis for Bernoulli rewards, and extend it to rewards in an exponential family. This extension notably relies on an upper bound on the tail of each posterior distribution (see Lemma 3.11). Such a result could be used too in an analysis of Bayes-UCB, but one would also need either a lower bound of the tail of the posterior, or (as in the analysis of Thompson Sampling) a result showing separately that the optimal arm has to be drawn a lot. We leave the generalization of Bayes-UCB to exponential families as future work.

We now give the proof of Lemma 2.2 and Theorem 2.1, that hold for Bernoulli bandit models.

2.3.3 Finite-time analysis

We start by giving a sketch of the proof of Lemma 2.2. This proof is detailed in Section 2.5. Lemma 2.2 relies on two ingredients. The first is a connection between Beta and Binomial distributions: for any integers a, b, the distribution $\text{Beta}(a, b)$ is the law of the a-th order statistic among $a+b-1$ uniform random variables, so that

$$\mathbb{P}(X \geq x) = \mathbb{P}(S_{a+b-1,x} \leq a-1) = \mathbb{P}(S_{a+b-1,1-x} \geq b)\,,$$

where $S_{n,x}$ denotes a binomial distribution with parameters n and x. Bounding the beta quantiles then boils down to controlling the binomial tails, which is achieved by our second ingredient, Lemma 2.5 below. The upper bound in inequality (2.5) is easily obtained using Chernoff inequality, whereas the lower bound follows from a careful application of the method of types (adapting Lemma 2.1.9. of [Dembo and Zeitouni, 2010]).

Lemma 2.5. *If $S_{n,x}$ is binomial with parameters n and x and k is an integer such that $k \geq nx$,*

$$\frac{e^{-nd(\frac{k}{n},x)}}{n+1} \leq \mathbb{P}(S_{n,x} \geq k) \leq e^{-nd(\frac{k}{n},x)}. \tag{2.5}$$

We now prove Theorem 2.1, the proof of intermediate lemmas being postponed to Section 2.5. To ease the notation, we denote by $\mathbb{P}$ and $\mathbb{E}$ (in place of $\mathbb{P}_{\boldsymbol{\theta}}$ and $\mathbb{E}_{\boldsymbol{\theta}}$) the probability and expectation for a fixed Bernoulli bandit model parameterized by $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_K)$. We recall some useful notations, already defined in the previous chapter: $\hat{\mu}_a(t) = S_a(t)/N_a(t)$ is the empirical mean of rewards obtained from arm a up to the end of round t (with the convention that it is set to zero when $N_a(t) = 0$). $(Y_{a,k})$ is the sequence of successive rewards obtained from arm a. It is i.i.d. with Bernoulli distribution of mean μ_a and we let $\hat{\mu}_{a,s} = \frac{1}{s}\sum_{k=1}^{s} Y_{a,k}$.

Proof of Theorem 2.1. Assume without loss of generality that arm 1 is optimal and let $a \neq 1$ be a sub-optimal arm. Following the classic regret analysis for index policies discussed in Section 1.2.3 of Chapter 1, one would use the following decomposition of the event $(A_{t+1} = a)$:

$$(A_{t+1} = a) \subseteq (\mu_1 \geq q_1(t)) \bigcup (\mu_1 \leq q_1(t), A_{t+1} = a) \subseteq (\mu_1 \geq q_1(t)) \bigcup (\mu_1 \leq q_a(t), A_{t+1} = a)\,,$$

using that if a is drawn at time $t+1$, $q_a(t) > q_1(t)$. However, for the purpose of our analysis, we introduce here a slightly different decomposition, that depends on the sequence

$$g_t = \sqrt{\frac{2}{\log(t)}}.$$

Using the results and notation of Lemma 2.2, we write

$$\begin{aligned}(A_{t+1} = a) &\subseteq (\mu_1 - g_t \geq q_1(t)) \bigcup (\mu_1 - g_t \leq q_a(t), A_{t+1} = a) \\ &\subseteq (\mu_1 - g_t \geq \tilde{u}_1(t)) \bigcup (\mu_1 - g_t \leq u_a(t), A_{t+1} = a)\,.\end{aligned} \tag{2.6}$$

We analyse the Bayes-UCB algorithm with an initialization phase that draws each arm once, although no initialization is required in principle. This assumption is needed to make sure that when $t \geq K$, all the indices $u_a(t), \tilde{u}_a(t)$ (that do depend on $N_a(t)$, unlike the quantile $q_a(t)$), are well defined. One has

$$\mathbb{E}[N_a(T)] = 1 + \mathbb{E}\left[\sum_{t=K}^{T-1} \mathbb{1}_{(A_{t+1}=a)}\right],$$

which yields, using the decomposition (2.6),

$$\mathbb{E}[N_a(T)] \leq 1 + \underbrace{\sum_{t=K}^{T-1} \mathbb{P}\left(\mu_1 - g_t \geq \tilde{u}_1(t)\right)}_{A} + \underbrace{\sum_{t=K}^{T-1} \mathbb{P}\left(\mu_1 - g_t \leq u_a(t), A_{t+1} = a\right)}_{B}.$$

Term A will be small ($o(\log T)$) since arm 1 is not likely to be 'under-estimated' by its index $\tilde{u}_1(t)$ at each round. To show this, we have to adapt the proof of the self-normalized informational inequality introduced by [Garivier and Cappé, 2011], stated as Lemma 1.9 in the previous chapter, to the alternative exploration rate and the bias of index $\tilde{u}_1(t)$. This is done in Lemma 2.6, and the reason why the extra term g_t is needed to handle the alternative exploration rate appears in its proof.

Lemma 2.6.

$$\sum_{t=K}^{T-1} \mathbb{P}\left(\mu_1 - g_t \geq \tilde{u}_1(t)\right) \quad \leq \quad \frac{2e+3}{1-\mu_1} \log\log T + 26$$

To upper bound term B, we start by using that $g_{N_a(t)} \geq g_t$ and $x \mapsto d^+(\hat{\mu}_a(t), x)$ is non-decreasing, where $d^+(x,y) = d(x,y)\mathbb{1}_{(x<y)}$:

$$\begin{aligned}\sum_{t=K}^{T-1} \mathbb{P}\left(\mu_1 - g_t \leq u_a(t), A_{t+1} = a\right) &\leq \sum_{t=K}^{T-1} \mathbb{P}\left(A_{t+1} = a, N_a(t) d^+(\hat{\mu}_a(t), \mu_1 - g_t) \leq f(T)\right) \\ &\leq \sum_{t=K}^{T-1} \mathbb{P}\left(A_{t+1} = a, N_a(t) d^+(\hat{\mu}_a(t), \mu_1 - g_{N_a(t)}) \leq f(T)\right)\end{aligned}$$

Summing over the possible values of $N_a(t)$ and interverting the sums to get rid of the self-normalized quantities (a technique already described in Lemma 1.10 in Chapter 1), yields

$$(B) \leq \sum_{s=1}^{T} \mathbb{P}\left(s d^+\left(\hat{\mu}_{a,s}, \mu_1 - g_s\right) \leq f(T)\right).$$

Lemma 2.7 below permits to conclude the proof. A similar term is bounded in Appendix A.2. of [Cappé et al., 2013]: we adapt their proof to the presence of the extra term g_s.

Lemma 2.7. *Let*

$$N_a(\epsilon) = \frac{d(\mu_a, \mu_1)}{1+\epsilon} \exp\left(\frac{8}{(\mu_1(1-\mu_1))^2} \frac{(1+\epsilon)^2}{\epsilon^2 d(\mu_a, \mu_1)^2}\right)$$

For T such that $f(T) \geq N_a(\epsilon)$, one has

$$\begin{aligned}\sum_{s=1}^{T} \mathbb{P}\left(s d^+\left(\hat{\mu}_{a,s}, \mu_1 - g_s\right) \leq f(T)\right) \quad \leq \quad & (1+\epsilon)\frac{f(T)}{d(\mu_a, \mu_1)} + \sqrt{f(T)} \sqrt{\frac{2\pi(1+\epsilon)^3 d'(\mu_a, \mu_1)^2}{d(\mu_a, \mu_1)^3}} \\ & + \; 2(1+\epsilon)^2 \left(\frac{d'(\mu_a, \mu_1)}{d(\mu_a, \mu_1)}\right)^2.\end{aligned}$$

□

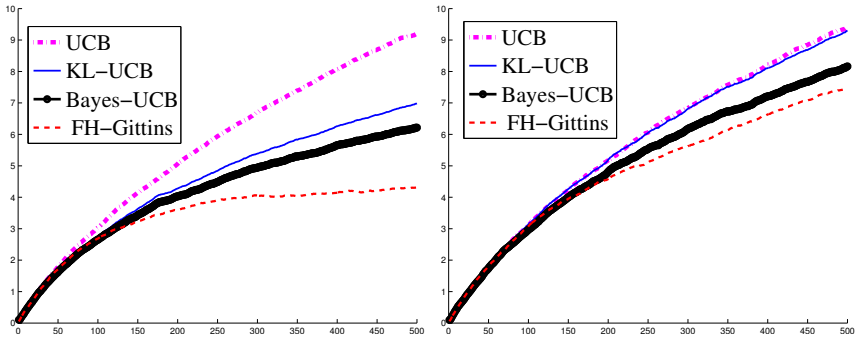

Figure 2.1: Cumulated regret for the two armed-bandit problem with $\mu_1 = 0.1, \mu_2 = 0.2$ (left) and $\mu_1 = 0.45, \mu_2 = 0.55$ (right).

2.4 Numerical experiments

2.4.1 Binary bandits

Numerical experiments have been carried out in a frequentist setting for bandits with Bernoulli rewards: for a fixed parameter θ and an horizon T, N bandit games with Bernoulli rewards are repeated for a given strategy. The main purpose of these numerical experiments is to compare the performance of Bayes-UCB in terms of cumulated regret with those of UCB and KL-UCB. These are presented in Figure 2.1, where the regret is averaged over $N = 5000$ simulations for two different two-armed bandit problems with horizon $T = 500$. We also include in the comparison the Bayesian algorithm based on Finite-Horizon Gittins indices (FH-Gittins). Whereas the performance of FH-Gittins are more striking in the left situation (0.1/0.2) than in the right one (0.45/0.55), Bayes-UCB improves equally over KL-UCB in all scenarios.

The horizon T had to be chosen quite small because of the numerical complexity of the FH-Gittins algorithm. In the next chapter, we will display numerical results in the Bernoulli case for larger horizons, also including Thompson Sampling.

2.4.2 Gaussian rewards with unknown means and variances

For the bandit problem with Gaussian rewards with unknown mean and variance, few algorithms have been proposed. We compare Bayes-UCB with UCB1-norm and UCB-Tuned, two algorithms introduced by [Auer et al., 2002a]. UCB1-norm is designed for Gaussian rewards and a logarithmic upper bound on its regret is given, whereas UCB-Tuned relies on estimates of the variance of each arm and can be used for general distributions, but without any theoretical guarantee. Figure 2.2 presents the regret in a 4-arms problem, on a horizon $T = 10000$, averaged over $N = 1000$ simulations. UCB-Tuned seems unadapted to the problem, whereas UCB1-norm and Bayes-UCB achieve a regret proving that the asymptotic lower bound of Burnetas & Katehakis is pessimistic for such short horizons (see also [Garivier and Cappé, 2011]). Bayes-UCB outperforms UCB1-norm, mostly because of the more appropriate choice of a quantile of order $1 - 1/t$.

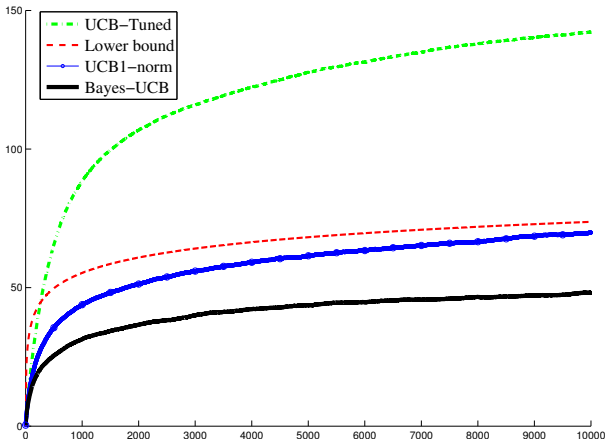

Figure 2.2: Regret in a 4-arms problem with parameters $\mu = [1.8\ 2\ 1.5\ 2.2]$, $\sigma = [0.5\ 0.7\ 0.5\ 0.3]$.

2.4.3 Sparse linear bandits

The linear bandit model presented in Section 2.2 relies on linear regression. Many recent works have highlighted the importance of sparsity issues in this context. We show that Bayes-UCB can address sparse linear bandit problems by using a prior that encourages sparsity of the parameter θ. This 'spike-and-slab' prior is defined as follows: the coordinates of θ are independent, with distribution

$$\theta_a \sim \epsilon\delta_0 + (1-\epsilon)\mathcal{N}(0,\kappa^2)\ .$$

Let C be the random vector in $\mathbb{R}^d$ indicating the non-zero coordinates of θ: $C_a = \mathbb{1}_{(\theta_a \neq 0)}$. If J denotes a set of indices, let $X_{t,J} \in \mathcal{M}_{t,|J|}(\mathbb{R})$ be the submatrix of the design matrix X_t with columns in J only and $\theta_J \in \mathbb{R}^{|J|}$ the subvector with coordinates in J.

Given C and Y_t, denote by J_1 the set of non-zero coordinates in C. The subvector θ_{J_1} is the solution of a Bayesian regression problem with prior $\mathcal{N}(0,\kappa^2 I_{|J_1|})$, hence

$$\theta_{J_1}|C, Y_t \ \sim\ \mathcal{N}\left((X'_{t,J_1}X_{t,J_1} + (\sigma/\kappa)^2\mathrm{I}_{|J_1|})^{-1}X'_{t,J_1}Y_t;\ \sigma^2(X'_{t,J_1}X_{t,J_1} + (\sigma/\kappa)^2\mathrm{I}_{|J_1|})^{-1}\right)\ .$$

The marginal distribution of C given Y is

$$P(C|Y) \propto \epsilon^{|J_0|}(1-\epsilon)^{|J_1|}\mathcal{N}\left(Y_t|0,\kappa^2 X_{t,J_1}X'_{t,J_1} + \sigma^2\mathrm{I}_t\right)\ .$$

The normalization term involves a sum over 2^d possible configurations of C. When d is small, the exact Bayes-UCB indices can be computed, as the dot-product $U'_a\theta$ follows a mixture of Gaussian distributions. For higher dimensions, one can use Gibbs sampling to sample from $C|Y$, and produce samples from $\theta|Y$ that lead to approximated values of $q_a(t)$.

Numerical simulation have been carried out for a sparse problem in dimension $d = 10$ where θ only has two non-zero coordinates. In Figure 2.3 we compare the regret of Bayes-UCB for three different priors: the general multivariate Gaussian prior discussed in Section 2.2, an oracle Gaussian prior on the first two coordinates only (meaning that the sparsity pattern is known) and Bayes-UCB with a sparse prior. The 20 arms of the problem are chosen randomly on the unit sphere and the regret is averaged over $N = 100$ simulations for an horizon $T = 1000$. As expected, the use of a sparsity-inducing prior in this case results in an algorithm with greatly enhanced performance. Such experiments should now be carried out in larger dimension using more sophisticated MCMC algorithms, designed for sampling from sparsity-inducing priors in a regression model, like the STMALA algorithm of [Schreck et al., 2013].

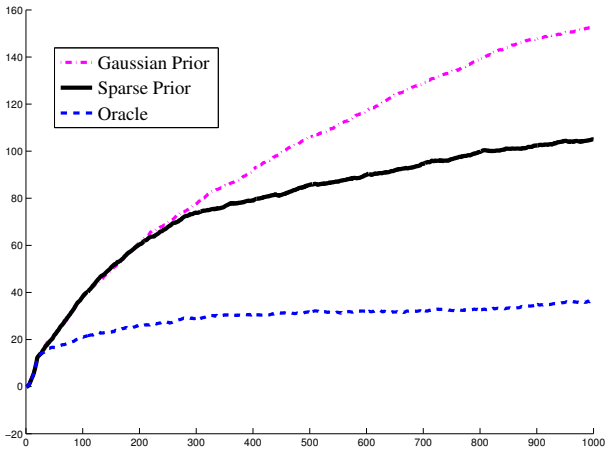

Figure 2.3: Cumulated regret in a 20 arms problem for Bayes-UCB with different prior distributions.

2.5 Elements of proof

2.5.1 Proof of Lemma 2.2

If $X \sim \text{Beta}(a,b)$, using Lemma 2.5 and the link between the tail of Beta and binomial distribution mentioned above, we get, for $x > \frac{a-1}{a+b-1}$,

$$\frac{e^{-(a+b-1)d(\frac{a-1}{a+b-1},x)}}{a+b} \leq \mathbb{P}(X \geq x) \leq e^{-(a+b-1)d(\frac{a-1}{a+b-1},x)}$$

Let $q_{1-\gamma} = Q(1-\gamma, \text{Beta}(a,b))$. Since :

$$(a+b-1)d\left(\frac{a-1}{a+b-1}, x\right) \geq \log(1/\gamma) \quad \Rightarrow \quad x \geq q_{1-\gamma}$$

we have that :

$$\begin{aligned} x_+^* &= \underset{x>\frac{a-1}{a+b-1}}{\text{argmin}}\{(a+b-1)d\left(\frac{a-1}{a+b-1}, x\right) \geq \log(1/\gamma)\} \\ &= \underset{x>\frac{a-1}{a+b-1}}{\text{argmax}}\{(a+b-1)d\left(\frac{a-1}{a+b-1}, x\right) \leq \log(1/\gamma)\} \end{aligned}$$

is still an upper bound for the quantile $q_{1-\gamma}$. The same reasoning shows $q_{1-\gamma}$ is lower-bounded by

$$x_-^* = \underset{x>\frac{a-1}{a+b-1}}{\text{argmax}} \left\{(a+b-1)d\left(\frac{a-1}{a+b-1}, x\right) \leq \log\left(\frac{1}{\gamma(a+b)}\right)\right\}$$

Moreover we can easily show that

$$x_+^* \leq \underset{x>\frac{a-1}{a+b-2}}{\text{argmax}}\{(a+b-2)d\left(\frac{a-1}{a+b-2}, x\right) \leq \log(1/\gamma)\}$$

using mainly the fact that $y \mapsto d(y,x)$ is decreasing for $y < x$.We get the final result using $a = S_a(t) + 1$,$b = N_a(t) - S_a(t) + 1$ and $\gamma = 1/(t\log(t)^c)$.

2.5.2 Proof of Lemma 2.6

To upper bound $(A) = \sum_{t=K}^{T-1} \mathbb{P}\left(\mu_1 - g_t \geq \tilde{u}_1(t)\right)$ we start by splitting the sum according to the number of draws of the optimal arm.

$$(A) \leq \underbrace{\sum_{t=K}^{T-1} \mathbb{P}\left(\mu_1 - g_t > \tilde{u_1}(t)\ ,\ N_1(t) + 2 \leq \log^2(t)\right)}_{A_1} + \underbrace{\sum_{t=K}^{T-1} \mathbb{P}\left(\mu_1 - g_t > \tilde{u}_1(t)\ ,\ N_1(t) + 2 \geq \log^2(t)\right)}_{A_2}\ .$$

To upper bound term A_1, we use that $\log\left(\frac{t}{N_a(t)+2}\right)$ in $\tilde{u_1}(t)$ is lower-bounded by $\log\left(\frac{t}{\log(t)^2}\right)$:

$$\begin{aligned}
&\left(\mu_1 - g_t > \tilde{u}_1(t), N_1(t) + 2 \leq (\log t)^2\right) \\
&= \left((N_1(t)+1)d^+\left(\frac{S_1(t)}{N_1(t)+1}, \mu_1 - g_t\right) \geq \log\frac{t}{N_1(t)+2} + 5\log\log t, N_1(t) + 2 \leq (\log t)^2\right) \\
&\subseteq \left((N_1(t)+1)d^+\left(\frac{S_1(t)}{N_1(t)+1}, \mu_1 - g_t\right) \geq \log\frac{t}{\log^2 t} + 5\log(\log t)\right) \\
&\subseteq \left((N_1(t)+1)d^+\left(\frac{S_1(t)}{N_1(t)+1}, \mu_1\right) \geq \log t + 3\log(\log t)\right).
\end{aligned}$$

Hence,

$$(A_1) \leq \sum_{t=K}^{T-1} \mathbb{P}\left(\exists s \in \{1, \dots t\} : (s+1)d^+\left(\frac{S_{1,s}}{s+1}, \mu_1\right) \geq \log(t) + 3\log\log(t)\right),$$

where $S_{1,s} = \sum_{k=1}^{s} Y_{1,k}$ is the sum of the s first rewards from arm 1. An adaptation of the proof of Lemma 1.9 yields the following self-normalized inequality, whose proof is given below.

Lemma 2.8.

$$\mathbb{P}\left(\exists s \in \{1, \dots, t\} : (s+1)d^+\left(\frac{S_{1,s}}{s+1}, \mu_1\right) \geq \delta\right) \leq \frac{1}{1-\mu_1}\left(\delta\log(t) + 1\right)\exp(-\delta + 1).$$

Lemma 2.8 leads to the upper-bound

$$\begin{aligned}
(A_1) &\leq \frac{e}{1-\mu_1}\sum_{t=K}^{T-1}\frac{\log^2(t) + 3\log(t)\log(\log(t)) + 1}{t(\log(t))^3} \\
&\leq \frac{e}{1-\mu_1}\left(2 + \frac{3}{e}\right)\sum_{t=K}^{T-1}\frac{1}{t\log(t)} \leq \frac{2e+3}{1-\mu_1}\log\log T.
\end{aligned}$$

On the events involved in term A_2, the optimal arm has been sufficiently drawn to be well estimated, so we use that

$$\left(\mu_1 - g_t > \tilde{u}_1(t)\right) \subseteq \left(\mu_1 - g_t > \frac{S_1(t)}{N_1(t)+1}\right) \subseteq \left(\mu_1 - g_t \geq \frac{S_1(t)}{N_1(t)} - \frac{1}{N_1(t)+1}\right).$$

Without the term g_t, the probability of the above event wouldn't be small, but here one can write

$$\begin{aligned}(A_2) &\leq \sum_{t=K}^{T-1} \mathbb{P}\left(\left(\mu_1 - g_t \geq \frac{S_1(t)}{N_1(t)} - \frac{1}{N_1(t)+1}\right) \cap \left(N_1(t) > \log(t)^2 - 2\right)\right) \\ &\leq 5 + \sum_{t=6}^{T-1} \mathbb{P}\left(\exists s \in [\log(t)^2 - 2; t] \ : \ \mu_1 - g_t \geq \frac{\sum_{r=1}^{s} Y_{1,r}}{s} - \frac{1}{s+1}\right) \\ &\leq 5 + \sum_{t=6}^{T-1} \mathbb{P}\left(\exists s \in [\log(t)^2 - 2; t] \ : \ \sum_{r=1}^{s} (\mu_1 - Y_{1,r}) \geq g_t s - 1\right)\end{aligned}$$

We start the sum at $t = 6$ since when $t \geq 6$ one has simultaneously $\log(t)^2 - 2 \geq 1$ and $\forall s \geq \log(t)^2 - 2, g_t s - 1 \geq 0$. For each t, noting $t' = \log(t)^2 - 2$, we use a peeling and split the interval $[t'; t]$ in smaller intervals of the form $[2^k t'; 2^{k+1} t'[$ on which we will use a maximal inequality.

$$\begin{aligned}\mathbb{P}\left(\exists s \in [t'; t] \ : \ \sum_{r=1}^{s} (\mu_1 - Y_{1,r}) \geq g_t s - 1\right) &= \sum_{k=0}^{\frac{\log(t/t')}{\log(2)}} \mathbb{P}\left(\exists s \in [2^k t'; 2^{k+1} t'[\ : \ \sum_{r=1}^{s} (\mu_1 - Y_{1,r}) \geq g_t s - 1\right) \\ &\leq \sum_{k=0}^{\infty} \mathbb{P}\left(\exists s \in [2^k t'; 2^{k+1} t'[\ : \ \sum_{r=1}^{s} (\mu_1 - Y_{1,r}) \geq g_t 2^k t' - 1\right) \\ &\leq \sum_{k=0}^{\infty} \mathbb{P}\left(\exists s \in [1; 2^{k+1} t'[\ : \ \sum_{r=1}^{s} (\mu_1 - Y_{1,r}) \geq g_t 2^k t' - 1\right) \\ &\leq \sum_{k=0}^{\infty} \exp\left(-\frac{2\left(g_t 2^k t' - 1\right)^2}{2^{k+1} t'}\right) \leq \sum_{k=0}^{\infty} \exp\left(-2^k \frac{(g_t t' - \frac{1}{2^k})^2}{t'}\right) \leq \sum_{k=0}^{\infty} e^{-\frac{\left(g_t t' - 1\right)^2}{t'}(k+1)},\end{aligned}$$

where we use that $2^k \geq 1 + k$. And, with the expressions g_t and t',

$$\sum_{k=0}^{\infty} e^{-\frac{\left(g_t t' - 1\right)^2}{t'}(k+1)} = \frac{1}{e^{\frac{(g_t t' - 1)^2}{t'}} - 1} = \frac{1}{\exp\left(2\frac{\left(\log^2 t - \frac{1}{\sqrt{2}}\sqrt{\log t} - 2\right)^2}{\log^3 t - 2\log t}\right) - 1} \underset{t \to \infty}{\sim} \frac{1}{t^2}$$

This function of t is the general term of a convergent series. A numerical evaluation of the sum of this series finally yields $(A_2) \leq 26$, which concludes the proof.

□

Proof of Lemma 2.8. As in the proof of Lemma 1.9 (given in Appendix A.2), we start with a peeling argument, but on the values of $s + 1$. Let $\gamma > 1$.

$$\begin{aligned}&\mathbb{P}\left(\exists s \in \{1, \dots, t\} : (s+1) d^+\left(\frac{S_{1,s}}{s+1}, \mu_1\right) \geq \delta\right) \qquad (2.7) \\ \leq &\sum_{k=1}^{\left\lceil \frac{\log(t)}{\log(\gamma)} \right\rceil} \mathbb{P}\left(\gamma^{k-1} \leq s + 1 < \gamma^k \ , \ (s+1) d^+\left(\frac{S_{1,s}}{s+1}, \mu_1\right) \geq \delta\right) \\ \leq &\sum_{k=1}^{\left\lceil \frac{\log t}{\log(\gamma)} \right\rceil} \mathbb{P}(E_k) \text{ with } E_k := \left(\gamma^{k-1} \leq s + 1 < \gamma^k \ , \ d^+\left(\frac{S_{1,s}}{s+1}, \mu_1\right) \geq \frac{\delta}{\gamma^k}\right)\end{aligned}$$

If k is such that $\frac{\delta}{\gamma^k} \geq d(0,\mu_1)$ clearly $\mathbb{P}(E_k) = 0$. Otherwise, one can introduce a unique z_k such that

$$\frac{S_{1,s}}{s+1} < z_k < \mu_1 \quad \text{and} \quad d(z_k,\mu_1) = \frac{\delta}{\gamma^k}.$$

As Bernoulli distribution belong to an exponential family, we have from Lemma 1.4 that

$$d(z_k,\mu_1) = \max_{\lambda\in\mathbb{R}} \{\lambda z_k - \phi_{\mu_1}(\lambda)\}$$

where $\phi_{\mu_1}(\lambda) = \mathbb{E}_{Y\sim\mathcal{B}(\mu_1)}[e^{\lambda Y}]$. The maximum is obtained for $\lambda_k = \log\frac{z_k}{1-z_k} - \log\frac{\mu_1}{1-\mu_1}$. Thus there exists $\lambda_k < 0$ such that $d(z_k,\mu_1) = \lambda_k z_k - \phi_{\mu_1}(\lambda_k)$ and one can check that $\phi_{\mu_1}(\lambda_k) = \log\frac{1-\mu_1}{1-z_k}$. Therefore,

$$\begin{aligned}
E_k &\subset \left(\gamma^{k-1} \leq s+1, \frac{S_{1,s}}{s+1} \leq z_k, \lambda_k z_k - \phi_{\mu_1}(\lambda_k) = \frac{\delta}{\gamma^k}\right) \\
&\subset \left(\frac{S_{1,s}}{s+1} \leq z_k, \lambda_k z_k - \phi_{\mu_1}(\lambda_k) \geq \frac{\delta}{\gamma(s+1)}\right) \\
&\subset \left(\lambda_k \frac{S_{1,s}}{s+1} - \phi_{\mu_1}(\lambda_k) \geq \frac{\delta}{\gamma(s+1)}\right) \\
&\subset \left(\lambda_k S_{1,s} - (s+1)\phi_{\mu_1}(\lambda_k) \geq \frac{\delta}{\gamma}\right)
\end{aligned}$$

For every $\lambda \in \mathbb{R}$, $W_s^\lambda = \exp\left(\lambda S_{1,s} - (s+1)\phi_{\mu_1}(\lambda)\right)$ is a martingale. Thus,

$$\begin{aligned}
\mathbb{P}(E_k) &\leq \mathbb{P}\left(W_t^{\lambda_k} \geq \exp(\frac{\delta}{\gamma})\right) \underset{\text{Markov}}{\leq} e^{-\frac{\delta}{\gamma}}\mathbb{E}\left[W_t^{\lambda_k}\right] \\
&\underset{\text{super-martingale}}{\leq} e^{-\frac{\delta}{\gamma}}\mathbb{E}[W_0^{\lambda_k}] = e^{-\frac{\delta}{\gamma}}e^{-\phi_{\mu_1}(\lambda_k)} = \frac{1-z_k}{1-\mu_1}e^{-\frac{\delta}{\gamma}} \leq \frac{1}{1-\mu_1}e^{-\frac{\delta}{\gamma}}.
\end{aligned}$$

Summing over k and letting $\gamma = \frac{\delta}{\delta-1} > 1$, (2.7) is upper bounded by

$$\frac{1}{1-\mu_1}\left(\frac{\log(t)}{\log\left(\delta/(\delta-1)\right)} + 1\right)\exp(-\delta+1) \leq \frac{1}{1-\mu_1}\left(\delta\log(t)+1\right)e^{-\delta+1},$$

which concludes the proof.

2.5.3 Proof of Lemma 2.7

The quantity to be upper bounded is

$$(B)' := \sum_{s=1}^{T} \mathbb{P}\left(s d^+\left(\hat{\mu}_{a,s}, \mu_1 - g_s\right) \leq f(T)\right).$$

The function $g(q) = d^+(\hat{\mu}_{a,s}, q)$ is convex and differentiable and $g'(q) = \frac{q-\hat{\mu}_{a,s}}{q(1-q)}\mathbb{1}_{(\hat{\mu}_{a,s}\leq q)}$, thus

$$d^+(\hat{\mu}_{a,s}, \mu_1 - g_s) \geq d^+(\hat{\mu}_{a,s}, \mu_1) - g_s\frac{\mu_1 - \hat{\mu}_{a,s}}{\mu_1(1-\mu_1)} \geq d^+(\hat{\mu}_{a,s}, \mu_1) - g_s\frac{2}{\mu_1(1-\mu_1)}.$$

And

$$(B)' \leq \sum_{s=1}^{T} \mathbb{P}\left(d^+(\hat{\mu}_{a,s}, \mu_1) \leq \frac{f(T)}{s} + \frac{2g_s}{\mu_1(1-\mu_1)} \right).$$

Let $\epsilon > 0$. Introducing

$$K_T := \left\lceil \frac{(1+\epsilon) f(T)}{d(\mu_a, \mu_1)} \right\rceil$$

we can split the sum:

$$(B)' \leq K_T + \sum_{s=K_T+1}^{T} \mathbb{P}\left(d^+(\hat{\mu}_{a,s}, \mu_1) \leq \frac{f(T)}{s} + b_T \right), \quad \text{with} \quad b_T = \frac{2g_{K_T}}{\mu_1(1-\mu_1)}.$$

If $N_a(\epsilon)$ is chosen as in the statement of Lemma 2.7, one has

$$f(T) \geq N_a(\epsilon) \;\Rightarrow\; \frac{f(T)}{s} + b_T \;\leq\; \frac{d(\mu_a, \mu_1)}{1+\epsilon} + b_T \leq d(\mu_a, \mu_1).$$

For such values of T, for each $s \geq K_T + 1$ there exists $\mu^*(s) \in]\mu_a; \mu_1[$ such that $d(\mu^*(s), \mu_1) = \frac{f(T)}{s} + b_T$. Then, using Hoeffding inequality and comparing with an integral,

$$(B)' \leq K_T + \int_{K_T}^{\infty} \exp(-2s(\mu^*(s) - \mu_a)^2) ds \,.$$

Using the convexity of the function $x \mapsto d(x, \mu_1)$, a lower bound on $\mu^*(s) - \mu_a$ can be obtained, as in Appendix 2 of [Cappé et al., 2013]:

$$\mu^*(s) - \mu_a \geq \frac{d(\mu_a, \mu_1) - \left[\frac{f(T)}{s} + b_T \right]}{-d'(\mu_a, \mu_1)}$$

[Cappé et al., 2013] also provide tight upper bound on the resulting integrals, and following their approach allows us to conclude the proof:

$$\begin{aligned}
(B)' &\leq K_T + \int_{K_T}^{\infty} \exp\left(-\frac{2s}{d'(\mu_a, \mu_1)^2} \left(\frac{f(T)}{s} + b_T - d(\mu_a, \mu_1) \right)^2 \right) ds \\
&\leq K_T + f(T) \int_{\frac{1+\epsilon}{d(\mu_a, \mu_1)}}^{\infty} \exp\left(-\frac{2uf(T)}{d'(\mu_a, \mu_1)^2} \left(\frac{1}{u} + b_T - d(\mu_a, \mu_1) \right)^2 \right) du \\
&\leq K_T + f(T) \int_{\frac{1+\epsilon}{d(\mu_a, \mu_1)}}^{\frac{2(1+\epsilon)}{d(\mu_a, \mu_1)}} \exp\left(-\frac{2(1+\epsilon)\left(\frac{1}{u} + b_T - d(\mu_a, \mu_1) \right)^2}{d(\mu_a, \mu_1) d'(\mu_a, \mu_1)^2} f(T) \right) du \\
&\quad + f(T) \int_{\frac{2(1+\epsilon)}{d(\mu_a, \mu_1)}}^{\infty} \exp\left(-\frac{2uf(T)}{d'(\mu_a, \mu_1)^2} \frac{d(\mu_a, \mu_1)^2}{4(1+\epsilon)^2} \right) du \\
&\leq K_T + f(T) \frac{4(1+\epsilon)^2}{d(\mu_a, \mu_1)^2} \int_0^{\infty} \exp\left(-v^2 f(T) \frac{2(1+\epsilon)}{d(\mu_a, \mu_1) d'(\mu_a, \mu_1)^2} \right) dv \\
&\quad + 2(1+\epsilon)^2 \left(\frac{d'(\mu_a, \mu_1)}{d(\mu_a, \mu_1)} \right)^2 \\
&\leq K_T + \sqrt{f(T)} \sqrt{ \frac{2\pi(1+\epsilon)^3 d'(\mu_a, \mu_1)^2}{d(\mu_a, \mu_1)^3} + 2(1+\epsilon)^2 \left(\frac{d'(\mu_a, \mu_1)}{d(\mu_a, \mu_1)} \right)^2 }.
\end{aligned}$$

Chapter 3

Thompson Sampling

In this chapter, we propose a finite-time analysis of Thompson Sampling that proves its asymptotic optimality in the context of regret minimization, when the rewards are Bernoulli distributed and the algorithm uses a uniform prior on the means. This work is a joint work with Nathaniel Korda and Rémi Munos, and was presented to the conference ALT in 2012 ([Kaufmann et al., 2012b]). In a follow-up work ([Korda et al., 2013]), we also proved that Thompson Sampling with a specific prior, the Jeffreys' prior is asymptotically optimal when the rewards belong to an exponential family.

A large part of this chapter is devoted to the presentation of our analysis for Bernoulli bandits (with only minor changes compared to that proposed in the paper [Kaufmann et al., 2012b]). We highlight the links with the Bayes-UCB analysis presented in the previous chapter. Our results relative to Thompson Sampling for exponential families are also be presented, with only a sketch of proof, the whole paper ([Korda et al., 2013]) being provided in Appendix B. This chapter also includes a numerical study that illustrates the performance of Thompson Sampling both in terms of regret and Bayes risk, compared to that of Bayes-UCB and other bandit algorithms discussed in previous chapters.

Contents

3.1 Introduction

In 1933, Thompson proposed the first (Bayesian) bandit algorithm in the context of two-armed Bernoulli bandits, which model clinical trials with two possible treatments. The proposed algorithm chooses the next arm based on P, the posterior probability of arm 1 being better than arm 2: arm 1 is chosen at random with probability $f(P)$, arm 2 with probability $1 - f(P)$, where f is some nondecreasing function. Thompson's work ([Thompson, 1933, Thompson, 1935]) was focused on the computation of the probability P when the posterior distribution on the mean of each arm is a Beta distribution. However, when $f(P) = P$, that is when arms are sampled according to their posterior probabilities of being optimal, Thompson's algorithm can be implemented without computing the probability P. For example, when the arms are independent, drawing at time $t+1$ two independent samples $\theta_1(t)$ and $\theta_2(t)$ from each posterior distribution π_1^t and π_2^t and choosing the arm with highest sample (i.e. $A_{t+1} = \text{argmax}_a \theta_a(t)$) is equivalent to Thompson's algorithm. Indeed, for a prior distribution Π_0,

$$\mathbb{P}_{\Pi_0}(A_{t+1} = 1|\mathcal{F}_t) = \mathbb{P}_{\Pi_0}(\theta_1(t) > \theta_2(t)|\mathcal{F}_t) = \mathbb{P}_{\Pi_0}(\theta_1 > \theta_2|\mathcal{F}_t) = P_t,$$

since conditionally to $\mathcal{F}_t$, θ_1 and θ_2 are distributed according to the posterior distribution π_1^t and π_2^t, just like the samples $\theta_1(t)$ and $\theta_2(t)$.

This simple principle 'draw each arm according to its posterior probability of being optimal' is now referred to as 'Thompson Sampling' and can be easily generalized beyond two-armed Bernoulli bandit models. As explained above for two-armed bandits, one way to implement Thompson Sampling is to draw a model according to our current belief (i.e. posterior distribution) and act optimally in this sampled model. This is how we define Thompson Sampling for parametric bandit models with independent arms in Algorithm 2. Each bandit model depends on a parameter $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$, and the mean of an arm parameterized by θ is given by $\mu(\theta)$. As usual, π_a^t denotes the posterior distribution on θ_a at the end of round t. When the arms are no longer independent, Thompson' heuristic samples a bandit model ν from the current posterior distribution over the *joint* distribution of the arms and chooses the optimal arm in this sampled model. This is how Thompson Sampling is implemented in contextual linear bandit, that will be studied in Chapter 4. Thompson Sampling using the prior distribution Π_0 will be denoted in the sequel by TS_{Π_0}.

Thompson Sampling for independent arms is still an index policy, but the index computed for each arm is longer an optimistic estimate of the mean —it can even be smaller than the posterior mean—, unlike the quantile used by Bayes-UCB. Bayes-UCB may appear as a 'regularized' version of Thompson Sampling, or conversely Thompson Sampling may be seen as a 'noisy' version of Bayes-UCB, in which the quantile is estimated using only one sample. Although Thompson Sampling was historically

Algorithm 2 Thompson Sampling for parametric bandits with independent arms (TS_{Π_0})

Require: $\Pi^0 = (\pi_1^0, \dots, \pi_K^0)$ (initial prior on $\boldsymbol{\theta}$)
1: **for** $t = 1$ **to** T **do**
2: **for** each arm $a = 1, \dots, K$ **do**
3: draw a sample $\theta_a(t-1) \sim \pi_a^{t-1}$
4: **end for**
5: draw an arm $A_t \in \arg\max_{a=1\dots K} \mu\left(\theta_a(t-1)\right)$
6: get reward $X_t = X_{A_t,t}$ and update the posterior distribution Π^t
7: **end for**

introduced first, in this thesis we started by presenting Bayes-UCB, because this algorithm is inspired by the optimistic principle on which asymptotically optimal frequentist algorithms are based. Moreover, the finite-time analysis we propose for Thompson Sampling partly relies on the introduction of well-chosen quantiles of the posterior distribution, and thus theoretical elements from the analysis of Bayes-UCB will be useful in the analysis that we propose here for Thompson Sampling.

Despite its simplicity, Thompson Sampling has been more or less forgotten for decades: it was acknowledged as the first bandit algorithm (for example by [Berry and Fristedt, 1985]), but there has been no attempt to analyse its theoretical properties or empirical performance before the late 2000's. At this period, it was re-discovered (sometimes independently) by several authors, under different names. The Bayesian Learning Automaton proposed by [Granmo, 2010] is nothing but Thompson Sampling for two-armed Bernoulli bandits, and the author gives a first consistency result: the probability that the optimal arm is chosen at time t goes to one as t goes to infinity. [Scott, 2010] uses Thompson Sampling under the name *randomized probability matching* and proposes an empirical evaluation of this method in Bernoulli bandit models and in the generalized linear bandit model (see [Filippi et al., 2010b]). Similarly, [Chapelle and Li, 2011] propose an empirical evaluation of Thompson Sampling in contextual bandit models. [May et al., 2012] study a slightly modified version of Thompson Sampling, called Optimistic Bayesian Sampling (OBS): if the sample $\theta_a(t)$ is such that $\mu(\theta_a(t))$ is smaller than the posterior mean $\mathbb{E}[\mu(\theta_a)|\mathcal{F}_t]$, it is replaced by the posterior mean. They prove that Thompson Sampling and OBS satisfy an 'average reward convergence criterion' for contextual bandits, which rewrites for classical bandits

$$\frac{\sum_{s=1}^{t} \mu_{A_s}}{t\mu^*} \xrightarrow[t\to\infty]{a.s} 1.$$

All these works attracted a lot of interest in Thompson Sampling, because efficient algorithms for contextual bandit models can be used in add prediction systems (see Chapter 4). Besides [Chapelle and Li, 2011] gave the first insight that Thompson Sampling might empirically outperform UCB-like algorithms in classical bandits. However, no theoretical results in terms of regret or Bayes risk could be extracted from these first works. The first logarithmic (finite-time) regret bound on the regret was given by [Agrawal and Goyal, 2012]. More precisely, they prove the following theorem for Bernoulli bandit models. To simplify the presentation, assume that arm 1 is the unique optimal arm and let $\Delta_a = \mu_1 - \mu_a$.

Theorem 3.1 ([Agrawal and Goyal, 2012],Theorem 2)**.** *Thompson Sampling using a uniform prior on the means, denoted by* Π_U*, satisfies*

$$\mathrm{R}_{\boldsymbol{\theta}}(T, \mathrm{TS}_{\Pi_U}) \leq O\left(\left(\sum_{a=2}^{K} \frac{1}{\Delta_a^2}\right)^2 \log(T)\right)$$

This result is optimal in the sense that the regret is logarithmic, as prescribed by Lai and Robbins' lower bound. However, its is not optimal in terms of the distribution-dependent constant multiplying $\log(T)$, since Lai and Robbins lower bound states that the regret of an uniformly efficient algorithm $\mathcal{A}$ satisfies in this particular case

$$\mathrm{R}_{\boldsymbol{\theta}}(T, \mathcal{A}) \geq \left(\sum_{a=2}^{K} \frac{\Delta_a}{d(\mu_a, \mu^*)}\right) \log(T) \tag{3.1}$$

A subgaussian approximation of the distribution-dependent term (using Pinsker's inequality) yields the sum over the suboptimal arms of the quantity $1/\Delta_a$. Whereas Theorem 3.1 exhibits a worse dependency

in the gaps Δ_a, [Agrawal and Goyal, 2012] present a refined result for two-armed bandits showing that the regret is $O(\log(T)/\Delta_2)$, which matches the subgaussian approximation of (3.1) up to a constant factor.

Even in the Bernoulli case, whether Thompson Sampling is asymptotically optimal with respect to Lai and Robbins lower bound was not known, and in the paper [Kaufmann et al., 2012b] we answer this open question positively: we provide the first finite-time analysis showing that Thompson Sampling using a uniform prior is asymptotically optimal. The finite-time analysis for Bernoulli bandit models given in Section 3.2 closely follows the paper [Kaufmann et al., 2012b], with minor modifications leading in particular to more explicit constants. Later, [Agrawal and Goyal, 2013a] proposed a different finite-time analysis of Thompson Sampling that also proves its asymptotic optimality, still for Bernoulli bandits. Using elements from their analysis, we were later able in the paper [Korda et al., 2013] to prove the asymptotic optimality of Thompson Sampling when rewards belong to exponential families. Thompson Sampling for exponential family bandits is discussed in Section 3.3.

[Agrawal and Goyal, 2013a] also present a distribution-independent upper bound, showing that the regret of Thompson Sampling for Bernoulli bandits satisfies $\mathrm{R}_{\boldsymbol{\theta}}(T, \mathrm{TS}_{\Pi_U}) \leq O(\sqrt{KT\log(T)})$, and is thus optimal, up to a logarithmic factor in T, with respect to the distribution-independent worst case lower bound (1.6) given in Chapter 1. In the same spirit, [Russo and Van Roy, 2014] propose prior-independent upper bound on the Bayes risk of Thompson Sampling, for very general bandit models. Their analysis relies on strong connections with UCB-like algorithm and will be discussed further in Chapter 4. Their result for K-armed bandits has been improved by [Bubeck and Liu, 2013] who show that, if the rewards are bounded in $[0,1]$, for any prior distribution Π_0, one has $\mathrm{BR}_{\Pi_0}(T, \mathrm{TS}_{\Pi_0}) \leq 14\sqrt{KT}$. This bound essentially matches the worst-case lower bound (1.7) presented in Chapter 1. However, whether Thompson Sampling matches the prior-dependent lower bound on the regret given by [Lai, 1987] (see Theorem 1.14 in Chapter 1) is not known yet, and we propose some numerical experiments in Section 3.4 that investigate this question. As we prove the asymptotic optimality of Thompson Sampling in a frequentist sense, we also illustrate in this experimental section its good performance in terms of regret, when compared to the Bayesian and frequentist algorithms studied so far.

3.2 Finite-time analysis of Thompson Sampling for binary bandits

In this section, we fix a Bernoulli bandit model, parameterized by $\boldsymbol{\theta} = (\mu_1, \dots, \mu_K)$ where arm a is a Bernoulli distribution with mean μ_a. We assume that there is a unique optimal arm. This is not a restrictive assumption since it can be shown that adding a second optimal arm can only improve the performance of Thompson Sampling (as explained in Appendix A of [Agrawal and Goyal, 2012]). We moreover assume without loss of generality that the arms are ordered such that $\mu_1 > \mu_2 \geq \dots \geq \mu_K$. To ease the notation, we denote by $\mathbb{P}$ and $\mathbb{E}$ (in place of $\mathbb{P}_{\boldsymbol{\theta}}$ and $\mathbb{E}_{\boldsymbol{\theta}}$) the probability and expectation under this bandit model.

As usual, we denote by $S_a(t)$ the number of successes observed from action a at the end of round t, and denote the empirical mean by: $\hat{\mu}_a(t) := S_a(t)/N_a(t)$. With an uniform prior distribution over the means μ_a of the arms, the posterior distribution on μ_a at the end of round t is explicitly

$$\pi_a^t = \mathrm{Beta}\left(S_a(t) + 1, N_a(t) - S_a(t) + 1\right).$$

Let $F_{a,b}^{\mathrm{Beta}}$ denote the cdf of a $\mathrm{Beta}(a,b)$ distribution and $F_{j,\mu}^{\mathrm{B}}$ (resp $f_{j,\mu}^{\mathrm{B}}$) the cdf (resp pdf) of a $\mathrm{Binomial}(j,\mu)$ distribution. We recall an important link between Beta and Binomial distributions already

used in the analysis of Bayes-UCB and in the analysis of Thompson Sampling by [Agrawal and Goyal, 2012]:

$$F^{\text{Beta}}_{a,b}(y) = 1 - F^{B}_{a+b-1,y}(a-1)$$

We use this 'Beta-Binomial trick' at several stages of our analysis.

We denote by $q_a(t)$ a quantile of the posterior distribution that will be useful in the proof and by $u_a(t)$ the associated KL-UCB index. More precisely, letting $Q(\alpha, \pi)$ be the α-quantile of distribution π, these quantities are defined by

$$u_a(t) := \operatorname*{argmax}_{x > \frac{S_a(t)}{N_a(t)}} \left\{ d\left(\frac{S_a(t)}{N_a(t)}, x\right) \leq \frac{\log(T)}{N_a(t)} \right\} \quad \text{and} \quad q_a(t) := Q\left(1 - \frac{1}{T}, \pi^t_a\right).$$

Recall that from Lemma 2.2 in Chapter 2 (see Remark 2.3)

$$q_a(t) < u_a(t).$$

3.2.1 Sketch of Analysis

Like the analysis of Bayes-UCB, our analysis of Thompson Sampling is inspired by standard analysis of frequentist index policies. At round $t+1$, these policies compute an index $U_a(t)$ for each arm a, based on the sequence of observed rewards from this arm up to the end of round t, and choose $A_{t+1} = \operatorname{argmax}_a U_a(t)$. Such an analysis aims to bound the number of draws of a suboptimal arm, a, by considering two possible events that might lead to a play of this arm:

- the optimal arm (arm 1) is under-estimated, i.e. $U_1(t) < \mu_1$;
- the optimal arm is not under-estimated and the suboptimal arm a is drawn at time $t+1$.

Taking these to be a good description of the event $A_{t+1} = a$ leads to the decomposition

$$\mathbb{E}[N_a(T)] \leq \sum_{t=0}^{T-1} \mathbb{P}\left(U_1(t) < \mu_1\right) + \sum_{t=0}^{T-1} \mathbb{P}\left(\left(U_a(t) \geq \mu_1\right) \cap \left(A_{t+1} = a\right)\right).$$

As explained in Section 1.2.3 of Chapter 1, the analysis of an optimistic algorithm then proceeds by showing that the left term (the 'under-estimation' term) is $o\left(\log(T)\right)$ and the right term is of the form $\frac{1}{d(\mu_a,\mu_1)}\log(T) + o\left(\log(T)\right)$ (or at worst $\frac{2}{\Delta_a^2}\log(T) + o\left(\log(T)\right)$ as in the analysis of UCB1). This scheme of proof works for example for the analysis of UCB1 or KL-UCB (see [Cappé et al., 2013]).

However we cannot directly apply this method to analyse Thompson Sampling, as the sample $\theta_a(t)$ is not an optimistic estimate of μ_a. Indeed, even when π^t_1 is well concentrated and therefore close to a Gaussian distribution centered in μ_1, $\mathbb{P}\left(\theta_1(t) < \mu_1\right)$ is close to $\frac{1}{2}$ and the under-estimation term will not be small compared to $\log T$. Hence we will not compare in our proof the sample $\theta_1(t)$ to μ_1, but to $\mu_1 - \sqrt{6\log(t)/N_1(t)}$ (if $N_1(t) > 0$) which is the lower bound of an UCB interval. We set the convention that if $N_1(t) = 0$, $\sqrt{6\log(t)/N_1(t)} = \infty$. Similarly, when $N_a(t) = 0$, the indices $q_a(t)$ and $u_a(t)$ previously defined are set to $q_a(t) = u_a(t) = 1$.

As observed by [Agrawal and Goyal, 2012] the main difficulty in a regret analysis for Thompson Sampling is to control the number of draws of the optimal arm. We provide this control in the form of Proposition 3.2 whose proof, given in Section 3.2.3, explores in depth the randomized nature of Thompson Sampling.

Proposition 3.2. *There exists constants $b = b(\mu_1, \mu_2) \in (0,1)$ and $C_b < \infty$ such that*

$$\sum_{t=1}^{\infty} \mathbb{P}\left(N_1(t) \leq t^b\right) \leq C_b.$$

Remark 3.3. *In general, a result on the regret like* $\mathbb{E}[N_1(t)] \geq t - K\log(t)$ *does not imply a deviation inequality for* $N_1(t)$ *(see [Salomon and Audibert, 2011]). Proposition 3.2 is therefore a strong result, that enables us to adapt the standard analysis mentioned above.*

We can then reduce to analyzing the behavior of the algorithm once it has seen a reasonable number of draws from arm 1, and thus the posterior distribution is well concentrated. Using Proposition 3.2 and the new decomposition yields:

Theorem 3.4. *Consider* $\epsilon > 0$ *and* b *and* C_b *as in Proposition 3.2. For every suboptimal arm* a*, there exists constants* $N(b)$ *and* $N(\epsilon, \mu_1, \mu_a)$ *such that for* $T \geq N(\epsilon, \mu_1, \mu_a)$*,*

$$\mathbb{E}[N_a(T)] \leq (1+\epsilon)\frac{\log T}{d(\mu_a,\mu_1)} + \sqrt{\log(T)}\sqrt{\frac{2\pi(1+\epsilon)^3 d'(\mu_a,\mu_1)^2}{d(\mu_a,\mu_1)^3}} + 2(1+\epsilon)^2\left(\frac{d'(\mu_a,\mu_1)}{d(\mu_a,\mu_1)}\right)^2 + 5 + 2C_b + N(b).$$

The constants are made more explicit in the proofs of Proposition 3.2 and Theorem 3.4. The fact that Theorem 3.4 holds for every $\epsilon > 0$ gives the asymptotic optimality of Thompson Sampling.

3.2.2 Proof of Theorem 3.4

Step 1: Decomposition. First we recall the modified decomposition mentioned above:

$$\begin{aligned}\mathbb{E}[N_a(T)] &\leq \sum_{t=0}^{T-1}\mathbb{P}\left(\theta_1(t) \leq \mu_1 - \sqrt{\frac{6\log t}{N_1(t)}}\right) + \sum_{t=0}^{T-1}\mathbb{P}\left(\theta_a(t) > \mu_1 - \sqrt{\frac{6\log t}{N_1(t)}}, A_{t+1} = a\right)\\ &\leq \sum_{t=0}^{T-1}\mathbb{P}\left(\theta_1(t) \leq \mu_1 - \sqrt{\frac{6\log t}{N_1(t)}}\right)\\ &\quad + \sum_{t=0}^{T-1}\mathbb{P}\left(\theta_a(t) > \mu_1 - \sqrt{\frac{6\log t}{N_1(t)}}, A_{t+1} = a, \theta_a(t) < q_a(t)\right) + \sum_{t=0}^{T-1}\mathbb{P}\left(\theta_a(t) > q_a(t)\right).\end{aligned}$$

The sample $\theta_a(t)$ from the posterior is not very likely to exceed the quantile $q_a(t)$ introduced above:

$$\sum_{t=0}^{T-1}\mathbb{P}\left(\theta_a(t) > q_a(t)\right) \leq \sum_{t=0}^{T-1}\mathbb{E}\left[\mathbb{P}\left(\theta_a(t) > q_a(t) | \mathcal{F}_t\right)\right] \leq \sum_{t=0}^{T-1}\frac{1}{T} = 1.$$

Finally, using that $u_a(t) \geq q_a(t)$ yields

$$\mathbb{E}[N_a(T)] \leq \underbrace{\sum_{t=0}^{T-1}\mathbb{P}\left(\theta_1(t) \leq \mu_1 - \sqrt{\frac{6\log t}{N_1(t)}}\right)}_{A} + \underbrace{\sum_{t=0}^{T-1}\mathbb{P}\left(u_a(t) > \mu_1 - \sqrt{\frac{6\log t}{N_1(t)}}, A_{t+1} = a\right)}_{B} + 1. \quad (3.2)$$

Step 2: Bounding term A. Let b and C_b be defined in Proposition 3.2.

To deal with term A we show a new self-normalized deviation inequality adapted to the randomization occurring at each round of Thompson Sampling.

Lemma 3.5. *Let b and C_b be defined as in Proposition 3.2 and define*

$$N_0(b) = \inf\left\{t \in \mathbb{N} : \log(t)t^b \geq (\sqrt{6}-\sqrt{5})^{-2}\right\}.$$

One has

$$\sum_{t=1}^{\infty} \mathbb{P}\left(\theta_1(t) \leq \mu_1 - \sqrt{\frac{6\log t}{N_1(t)}}\right) \leq N_0(b) + 3 + C_b < \infty.$$

Proof Let (U_t) denote a sequence of i.i.d. uniform random variables on $[0,1]$, and let $S_{1,s} = \sum_{k=1}^{s} Y_{1,s}$ be the sum of the first s rewards from arm 1. In the following, we make the first use of the link between Beta and Binomial distributions:

$$\begin{aligned}
&\mathbb{P}\left(\theta_1(t) \leq \mu_1 - \sqrt{\frac{6\log t}{N_1(t)}}\right) = \mathbb{P}\left(U_t \leq F^{\text{Beta}}_{S_1(t)+1, N_1(t)-S_1(t)+1}\left(\mu_1 - \sqrt{\frac{6\log t}{N_1(t)}}\right)\right) \\
&= \mathbb{P}\left(\left(U_t \leq 1 - F^{\text{B}}_{N_1(t)+1, \mu_1 - \sqrt{\frac{6\log t}{N_1(t)}}}(S_1(t))\right) \cap \left(N_1(t) \geq t^b\right)\right) + \mathbb{P}\left(N_1(t) \leq t^b\right) \\
&= \mathbb{P}\left(\left(F^{\text{B}}_{N_1(t)+1, \mu_1 - \sqrt{\frac{6\log t}{N_1(t)}}}(S_1(t)) \leq U_t\right) \cap \left(N_1(t) \geq t^b\right)\right) + \mathbb{P}\left(N_1(t) \leq t^b\right) \\
&\leq \mathbb{P}\left(\exists s \in \{t^b \ldots t\} : F^{\text{B}}_{s+1, \mu_1 - \sqrt{\frac{6\log t}{s}}}(S_{1,s}) \leq U_t\right) + \mathbb{P}\left(N_1(t) \leq t^b\right) \\
&= \sum_{s=\lceil t^b \rceil}^{t} \mathbb{P}\left(S_{1,s} \leq (F^{\text{B}})^{-1}_{s+1, \mu_1 - \sqrt{\frac{6\log t}{s}}}(U_t)\right) + \mathbb{P}\left(N_1(t) \leq t^b\right)
\end{aligned}$$

The first term in the final line of this display now deals only with Binomial random variables with large numbers of trials (greater than t^b), and so we can draw on standard concentration techniques to bound this term. Proposition 3.2 takes care of the second term.

Let s be fixed. The random variable

$$(F^{\text{B}})^{-1}_{s+1, \mu_1 - \sqrt{6\log t/s}}(U_t) \sim \text{Bin}\left(s+1, \mu_1 - \sqrt{6\log t/s}\right)$$

is independent from $S_{1,s} \sim \text{Bin}(s, \mu_1)$. One can write $S_{1,s} = \sum_{l=1}^{s} Y_{1,s}$, where $(Y_{1,l})$ is an i.i.d sequence of Bernoulli random variables with mean μ_1. Introducing $(\tilde{Y}_{1,l})_l$, a second i.i.d. sequence of Bernoulli random variables with mean $\mu_1 - \sqrt{6\log t/s}$ that is independent from $(Y_{1,l})$, one has

$$\begin{aligned}
\mathbb{P}\left(S_{1,s} \leq (F^{\text{B}})^{-1}_{s+1, \mu_1 - \sqrt{\frac{6\log t}{s}}}(U_t)\right) &= \mathbb{P}\left(\sum_{l=1}^{s} Y_{l,s} \leq \sum_{l=1}^{s+1} \tilde{Y}_{l,s}\right) \\
&\leq \mathbb{P}\left(\sum_{l=1}^{s}\left(Y_{1,l} - \tilde{Y}_{1,l} - \sqrt{\frac{6\log t}{s}}\right) \leq -\left(\sqrt{6s\log t} - 1\right)\right).
\end{aligned}$$

Letting $Z_l := \mu_1 - \sqrt{6\log t/s}$, (Z_l) is a sequence of i.i.d centered random variable with range 2, and Hoeffding's inequality can be used to bound the last sum. Moreover, for $t \geq N_0(b)$ where

$$N_0(b) := \inf\left\{t \in \mathbb{N} : \log(t)t^b \geq (\sqrt{6}-\sqrt{5})^{-2}\right\}$$

one has $\sqrt{6s\log t} - 1 > \sqrt{5s\log t}$, for $s \geq t^b$, and one can write

$$\mathbb{P}\left(S_{1,s} < (F^{\mathrm{B}})^{-1}_{s+1,\mu_1-\sqrt{\frac{6\log t}{s}}}(U_t)\right) \leq \exp\left(-2\frac{(\sqrt{5s\log t})^2}{4s}\right) = e^{-\frac{5}{2}\log t} = \frac{1}{t^{\frac{5}{2}}}.$$

We conclude that

$$\sum_{t=1}^{\infty}\mathbb{P}\left(\theta_1(t) < \mu_1 - \sqrt{\frac{6\log t}{N_1(t)}}\right) \leq N_0(b) + \sum_{t=1}^{\infty}\frac{1}{t^{\frac{3}{2}}} + C_b \leq N_0(b) + 3 + C_b.$$

□

Step 3: Bounding Term B. We specifically show that

Lemma 3.6. *For T such that*

$$\log(T) \geq \frac{d(\mu_a,\mu_1)}{1+\epsilon}\exp\left(\frac{8}{(\mu_1(1-\mu_1))^2}\frac{(1+\epsilon)^2}{\epsilon^2 d(\mu_a,\mu_1)^2}\right),$$

one has

$$\begin{aligned}(B) \quad \leq \quad & (1+\epsilon)\frac{\log(T)}{d(\mu_a,\mu_1)} + \sqrt{\log(T)}\sqrt{\frac{2\pi(1+\epsilon)^3 d'(\mu_a,\mu_1)^2}{d(\mu_a,\mu_1)^3}} + 2(1+\epsilon)^2\left(\frac{d'(\mu_a,\mu_1)}{d(\mu_a,\mu_1)}\right)^2 \\ & +1 + N_1(b) + C_b.\end{aligned}$$

with $N_1(b) := \inf\left\{t \geq e^{2/b} : 3(\log t)^2 \leq t^b\right\}$.

Proof Using Proposition 3.2, term B can be rewritten

$$\begin{aligned}(B) &\leq \sum_{t=0}^{T-1}\mathbb{P}\left(u_a(t) > \mu_1 - \sqrt{\frac{6\log t}{N_1(t)}}, A_{t+1} = a, N_1(t) \geq t^b\right) + \sum_{t=0}^{T-1}\mathbb{P}\left(N_1(t) \leq t^b\right) \\ &\leq \sum_{t=0}^{T-1}\mathbb{P}\left(u_a(t) > \mu_1 - \sqrt{\frac{6\log t}{t^b}}, A_{t+1} = a\right) + C_b\end{aligned}$$

Let $N_1(b)$ be defined in the statement of Lemma 3.6. For $t \geq N_1(b)$, one has $\sqrt{6\log t/t^b} \leq \sqrt{2/\log(t)}$. Letting $g_t := \sqrt{\frac{2}{\log t}}$ as in the finite-time analysis proposed for Bayes-UCB in the previous chapter, one can write, in a very similar way,

$$\begin{aligned}(B) &\leq \sum_{t=N_1(b)+1}^{T-1}\mathbb{P}\left(u_a(t) > \mu_1 - g_t, A_{t+1} = a\right) + N_1(b) + C_b \\ &= \mathbb{E}\left[\sum_{t=N_1(b)+1}^{T-1}\left(\mathbb{1}_{(N_a(t)=0,A_{t+1}=a)} + \sum_{s=1}^{t}\mathbb{1}_{(N_a(t)=s,A_{t+1}=a)}\mathbb{1}_{(sd^+(\mu_{a,s},\mu_1-g_t)\leq\log(T))}\right)\right] + N_1(b) + C_b \\ &\leq 1 + \mathbb{E}\left[\sum_{t=N_1(b)+1}^{T-1}\sum_{s=1}^{t}\mathbb{1}_{(N_a(t)=s,A_{t+1}=a)}\mathbb{1}_{(sd^+(\mu_{a,s},\mu_1-g_s)\leq\log(T))}\right] + N_1(b) + C_b \\ &\leq \sum_{s=1}^{T-1}\mathbb{P}\left(sd^+(\mu_{a,s},\mu_1-g_s) \leq \log(T))\right) + 1 + N_1(b) + C_b.\end{aligned}$$

We used as in the previous chapter that for all s, $\sum_{t\in\mathbb{N}} \mathbb{1}_{(N_a(t)=s)}\mathbb{1}_{(A_{t+1}=a)} \leq 1$ and that when $s \leq t$, $g_s \leq g_t$. The sum in the right hand side can now exactly be upper bounded using Lemma 2.7 in Chapter 2 with $f(T) = \log(T)$, which concludes the proof.

□

Conclusion: Theorem 3.4 follows from Lemmas 3.5, 3.6 and inequality (3.2), letting $N(b) := N_0(b) + N_1(b)$.

3.2.3 Proof of Proposition 3.2: Exploiting the randomized nature of Thompson Sampling.

Since we focus on the number of draws of the optimal arm, let τ_j be the occurrence of the j^{th} play of the optimal arm (with $\tau_0 := 0$). Let $\xi_j := (\tau_{j+1} - 1) - \tau_j$: this random variable measures the number of time steps between the j^{th} and the $(j+1)^{th}$ play of the optimal arm, and so $\sum_{a=2}^{K} N_a(t) = \sum_{j=0}^{N_1(t)} \xi_j$. For each suboptimal arm, a relevant quantity is

$$C_a = \frac{32}{(\mu_1 - \mu_a)^2}.$$

We let $C = \max_{a\neq 1} C_a = 32/(\mu_1 - \mu_2)^2$ and introduce $\delta_a = (\mu_1 - \mu_a)/2$ and $\delta = \delta_2$.

Step 1: Initial Decomposition of Summands. First we use a union bound on the summands to extract the tails of the random variables ξ_j:

$$\begin{aligned}
\mathbb{P}(N_1(t) \leq t^b) &= \mathbb{P}\left(\sum_{a=2}^{K} N_a(t) \geq t - t^b\right) \\
&\leq \mathbb{P}\left(\exists j \in \left\{0, .., \lfloor t^b \rfloor\right\} : \xi_j \geq t^{1-b} - 1\right) \\
&\leq \sum_{j=0}^{\lfloor t^b \rfloor} \mathbb{P}(\xi_j \geq t^{1-b} - 1) \qquad (3.3)
\end{aligned}$$

This means that there exists a time range of length $t^{1-b}-1$ during which only suboptimal arms are played. In the case of two arms this implies that the (unique) suboptimal arm is played $\lceil \frac{t^{1-b}-1}{2} \rceil$ times during the first half of this time range. Thus its posterior becomes well concentrated around its mean with high probability, and we can use this fact to show that the probability that the suboptimal action is chosen a further $\lceil \frac{t^{1-b}-1}{2} \rceil$ times in a row is very small. In Figure 3.2.3 below, one indeed sees that the posterior of arm 2 —in blue— becomes concentrated, and so the samples $\theta_2(s)$ used by Thompson Sampling on the second half of the time range are very likely to always fall below $\mu_2 + \delta$. Thus, as arm 1 is not drawn on this time range, the samples $\theta_1(s)$ (that are i.i.d. samples from the current posterior on arm 1 —in red—, that does not change on the time range) have to always fall in the shaded region, which is not very likely.

To precise this heuristic argument and generalize it to more arms, we introduce a notion of a *saturated*, suboptimal action:

Definition 3.7. *Let t be fixed. For any $a \neq 1$, an action a is said to be* saturated *at time s if it has been chosen at least $C_a \ln(t)$ times. That is $N_a(s) \geq C_a \ln(t)$. We shall say that it is* unsaturated *otherwise. Furthermore at any time we call a choice of an unsaturated, suboptimal action an* interruption.

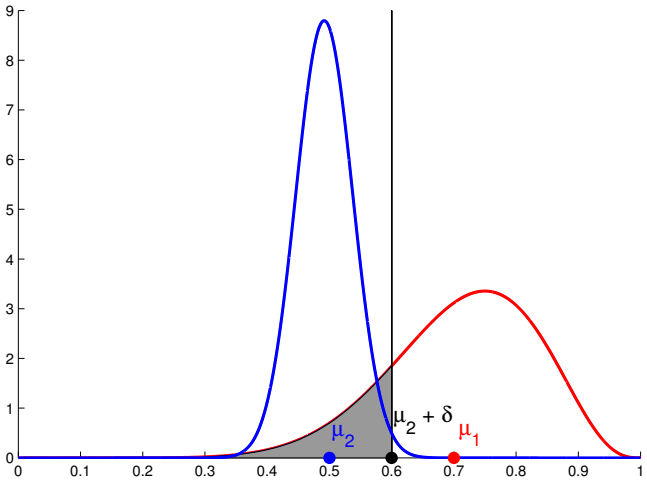

Figure 3.1: Thompson Sampling draws a lot the optimal arm: an heuristic explanation for two-armed bandits

We want to study the event $E_j = \{\xi_j \geq t^{1-b} - 1\}$. We introduce the interval $\mathcal{I}_j = \{\tau_j, \tau_j + \lceil t^{1-b} - 1 \rceil\}$ (included in $\{\tau_j, \tau_{j+1}\}$ on E_j) and begin by decomposing it into K subintervals:

$$\mathcal{I}_{j,l} := \left\{ \tau_j + \left\lceil \frac{(l-1)(t^{1-b}-1)}{K} \right\rceil, \tau_j + \left\lceil \frac{l(t^{1-b}-1)}{K} \right\rceil \right\}, \; l = 1, \ldots, K.$$

Now for each interval $\mathcal{I}_{j,l}$, we introduce:

- $F_{j,l}$: the event that by the end of the interval $\mathcal{I}_{j,l}$ at least l suboptimal actions are saturated;
- $n_{j,l}$: the number of interruptions during this interval.

We use the following decomposition to bound the probability of the event E_j:

$$\mathbb{P}(E_j) = \mathbb{P}(E_j \cap F_{j,K-1}) + \mathbb{P}(E_j \cap F^c_{j,K-1}) \tag{3.4}$$

To bound both probabilities, we will need the fact, stated in Lemma 3.8, that the probability of $\theta_1(s)$ being smaller than $\mu_2 + \delta$ during a long subinterval of $\mathcal{I}_j$ is small. This follows from the fact that the posterior on the optimal arm is always $\text{Beta}(S_1(\tau_j) + 1, j - S_1(\tau_j) + 1)$ on $\mathcal{I}_j$: hence, when conditioned on $S_1(\tau_j)$, $\theta_1(s)$ is an i.i.d. sequence with non-zero support above $\mu_2 + \delta$, and thus is unlikely to remain below $\mu_2 + \delta$ for a long time period. This idea is also important in the analysis of Thompson Sampling by [Agrawal and Goyal, 2012].

Lemma 3.8. *$\exists \lambda_0 = \lambda_0(\mu_1, \mu_2) > 1$ such that for $\lambda \in]1, \lambda_0[$, for every τ_j-measurable interval $\mathcal{J}$, such that $|\mathcal{J}| \geq f(t)$ for some positive function f, one has*

$$\mathbb{P}\left(\mathcal{J} \subseteq [\tau_j, \tau_{j+1}[, \; \forall s \in \mathcal{J} \; \theta_1(s) \leq \mu_2 + \delta\right) \leq (\alpha_{\mu_1,\mu_2})^{f(t)} + C_{\lambda,\mu_1,\mu_2} \frac{1}{f(t)^\lambda} e^{-j d_{\lambda,\mu_1,\mu_2}}$$

where $C_{\lambda,\mu_1,\mu_2}, d_{\lambda,\mu_1,\mu_2} > 0$ and $\alpha_{\mu_1,\mu_2} = (1/2)^{1-\mu_2-\delta}$.

The proof of this important lemma will be postponed to Section 3.5.1 and all the constants are explicitly defined there. Another key point in the proof is the fact that a sample from a saturated suboptimal arm cannot fall too far from its true mean. The following lemma is very close to Lemma 7 of [Agrawal and Goyal, 2012]. We propose a proof in Section 3.5.2.

Lemma 3.9.

$$\mathbb{P}\left(\exists s \leq t, \exists a \neq 1 : \theta_a(s) > \mu_a + \delta_a, N_a(s) > C_a \ln(t)\right) \leq \frac{2(K-1)}{t^2}.$$

Step 2: Bounding $\mathbb{P}(E_j \cap F_{j,K-1})$. On the event $E_j \cap F_{j,K-1}$, only saturated suboptimal arms are drawn on the interval $\mathcal{I}_{j,K}$. Using the concentration results for samples of these arms in Lemma 3.9, we get

$$\begin{aligned}
\mathbb{P}(E_j \cap F_{j,K-1}) \leq & \mathbb{P}(\{\exists s \in \mathcal{I}_{j,K}, a \neq 1 : \theta_a(s) > \mu_a + \delta\} \cap E_j \cap F_{j,K-1}) \\
& + \mathbb{P}(\{\forall s \in \mathcal{I}_{j,K}, a \neq 1 : \theta_a(s) \leq \mu_a + \delta_a\} \cap E_j \cap F_{j,K-1}) \\
\leq & \mathbb{P}(\exists s \leq t, a \neq 1 : \theta_a(s) > \mu_a + \delta_a, N_a(t) > C_a \ln(t)) \\
& + \mathbb{P}(\{\forall s \in \mathcal{I}_{j,K}, a \neq 1 : \theta_a(s) \leq \mu_2 + \delta\} \cap E_j \cap F_{j,K-1}) \\
\leq & \frac{2(K-1)}{t^2} + \mathbb{P}(\mathcal{I}_{j,K} \subseteq [\tau_j, \tau_{j+1}[,\ \forall s \in \mathcal{I}_{j,K}\ \theta_1(s) \leq \mu_2 + \delta).
\end{aligned}$$

The last inequality comes from the fact that if arm 1 is not drawn, the sample $\theta_1(s)$ must be smaller than some sample $\theta_a(s)$ and therefore smaller than $\mu_2 + \delta$. Since $\mathcal{I}_{j,K}$ is a τ_j-measurable interval of size $\left\lceil \frac{t^{1-b}-1}{K} \right\rceil$ we get using Lemma 3.8, for some fixed $\lambda \in]1, \lambda_0[$,

$$\begin{aligned}
&\mathbb{P}(\mathcal{I}_{j,K} \subseteq [\tau_j, \tau_{j+1}[,\ \forall s \in \mathcal{I}_{j,K}\ \theta_1(s) \leq \mu_2 + \delta) \\
&\leq (\alpha_{\mu_1,\mu_2})^{\frac{t^{1-b}-1}{K}} + C_{\lambda,\mu_1,\mu_2} \frac{1}{\left(\frac{t^{1-b}-1}{K}\right)^{\lambda}} e^{-j d_{\lambda,\mu_1,\mu_2}} =: g(\mu_1, \mu_2, b, j, t). \qquad (3.5)
\end{aligned}$$

Hence we have show that

$$\mathbb{P}(E_j \cap F_{j,K-1}) \leq \frac{2(K-1)}{t^2} + g(\mu_1, \mu_2, b, j, t), \qquad (3.6)$$

and choosing b such that $b < 1 - \frac{1}{\lambda}$, the following hypothesis on g holds:

$$\sum_{t \geq 1} \sum_{j \leq t^b} g(\mu_1, \mu_2, b, j, t) < +\infty.$$

Step 3: Bounding $\mathbb{P}(E_j \cap F^c_{j,K-1})$. We show through an induction that for all $2 \leq l \leq K$, if t is larger than some deterministic constant $N_{\mu_1,\mu_2,b}$ specified in the base case,

$$\mathbb{P}(E_j \cap F^c_{j,l-1}) \leq (l-2)\left(\frac{2(K-1)}{t^2} + f(\mu_1, \mu_2, b, j, t)\right)$$

for some function f such that $\sum_{t \geq 1} \sum_{1 \leq j \leq t^b} f(\mu_1, \mu_2, b, j, t) < \infty$. For $l = K$ we get

$$\mathbb{P}(E_j \cap F^c_{j,K-1}) \leq (K-2)\left(\frac{2(K-1)}{t^2} + f(\mu_1, \mu_2, b, j, t)\right). \qquad (3.7)$$

Step 4: The Base Case of the induction. Note that on the event E_j only suboptimal arms are played during $\mathcal{I}_{j,1}$. Hence at least one suboptimal arm must be played $\lceil \frac{t^{1-b}-1}{K^2} \rceil$ times.

There exists some deterministic constant $N_{\mu_1,\mu_2,b}$ such that for $t \geq N_{\mu_1,\mu_2,b}$, $\lceil \frac{t^{1-b}-1}{K^2} \rceil \geq C \ln(t)$ (the constant depends only on μ_1 and μ_2 because $C = C_2$). So when $t \geq N_{\mu_1,\mu_2,b}$, at least one suboptimal arm must be saturated by the end of $\mathcal{I}_{j,1}$. Hence, for $t \geq N_{\mu_1,\mu_2,b}$

$$\mathbb{P}(E_j \cap F_{j,1}^c) = 0.$$

This concludes the base case.

Step 5: The Induction. As an inductive hypothesis we assume that for some $2 \leq l \leq K-1$ if $t \geq N_{\mu_1,\mu_2,b}$ then

$$\mathbb{P}(E_j \cap F_{j,l-1}^c) \leq (l-2)\left(\frac{2(K-1)}{t^2} + f(\mu_1,\mu_2,b,j,t)\right).$$

Then, making use of the inductive hypothesis,

$$\begin{aligned}\mathbb{P}(E_j \cap F_{j,l}^c) &\leq \mathbb{P}(E_j \cap F_{j,l-1}^c) + \mathbb{P}(E_j \cap F_{j,l}^c \cap F_{j,l-1}) \\ &\leq (l-2)\left(\frac{2(K-1)}{t^2} + f(\mu_1,\mu_2,b,j,t)\right) + \mathbb{P}(E_j \cap F_{j,l}^c \cap F_{j,l-1}).\end{aligned}$$

To complete the induction we therefore need to show that:

$$\mathbb{P}(E_j \cap F_{j,l}^c \cap F_{j,l-1}) \leq \frac{2(K-1)}{t^2} + f(\mu_1,\mu_2,b,j,t). \tag{3.8}$$

On the event $(E_j \cap F_{j,l}^c \cap F_{j,l-1})$, there are exactly $l-1$ saturated arms at the beginning of interval $\mathcal{I}_{j,l}$ and no new arm is saturated during this interval. As a result there cannot be more than $KC\ln(t)$ interruptions during this interval, and so we have

$$\mathbb{P}(E_j \cap F_{j,l}^c \cap F_{j,l-1}) \leq \mathbb{P}(E_j \cap F_{j,l-1} \cap \{n_{j,l} \leq KC\ln(t)\}).$$

Let $\mathcal{S}_l$ denote the set of saturated arms at the end of $\mathcal{I}_{j,l}$ and introduce the following decomposition:

$$\begin{aligned}&\mathbb{P}(E_j \cap F_{j,l-1} \cap \{n_{j,l} \leq KC\ln(t)\}) \\ &\quad\leq \underbrace{\mathbb{P}(\{\exists s \in \mathcal{I}_{j,l}, a \in \mathcal{S}_{l-1} : \theta_a(s) > \mu_a + \delta_a\} \cap E_j \cap F_{j,l-1})}_{A} \\ &\quad+ \underbrace{\mathbb{P}(\{\forall s \in \mathcal{I}_{j,l}, a \in \mathcal{S}_{l-1} : \theta_a(s) \leq \mu_a + \delta_a\} \cap E_j \cap F_{j,l-1} \cap \{n_{j,l} \leq KC\ln(t)\})}_{B}.\end{aligned}$$

Clearly, using Lemma 3.9:

$$(A) \leq \mathbb{P}\left(\exists s \leq t, \exists a \neq 1 : \theta_a(s) > \mu_a + \delta_a, N_a(s) > C_a \ln(t)\right) \leq \frac{2(K-1)}{t^2}.$$

To deal with term (B), we introduce for k in $\{0,\ldots,n_{j,l}-1\}$ the random intervals $\mathcal{J}_k$ as the time range between the k^{th} and $(k+1)^{st}$ interruption in $\mathcal{I}_{j,l}$. For $k \geq n_{j,l}$ we set $\mathcal{J}_k = \varnothing$. Note that on the event in the probability (B) there is a subinterval of $\mathcal{I}_{j,l}$ of length $\left\lfloor \frac{t^{1-b}-1}{CK^2\ln(t)} \right\rfloor$ during which there are no interruptions.

Moreover on this subinterval of $\mathcal{I}_{j,l}$, for all $a \neq 1$, $\theta_a(s) \leq \mu_2 + \delta_2$. (This holds for unsaturated arms as well as for saturated arms since their samples are smaller than the maximum sample of a saturated arm.) Therefore,

$$\begin{aligned}(B) &\leq \mathbb{P}\Big(\big\{\exists k \in \{0,...,n_{j,l}\} : |\mathcal{J}_k| \geq (t^{1-b}-1)/(CK^2\ln(t))\big\} \\ &\qquad \cap \big\{\forall s \in \mathcal{I}_{j,l}, a \in \mathcal{S}_{l-1} : \theta_a(s) \leq \mu_2 + \delta\big\} \cap E_j \cap F_{j,l-1}\Big) \\ &\leq \sum_{k=1}^{KC\ln(t)} \mathbb{P}\left(\left\{|\mathcal{J}_k| \geq \frac{t^{1-b}-1}{CK^2\ln(t)}\right\} \cap \{\forall s \in \mathcal{J}_k, a \neq 1 : \theta_a(s) \leq \mu_2 + \delta\} \cap E_j\right) \\ &\leq \sum_{k=1}^{KC\ln(t)} \mathbb{P}\left(\left\{|\mathcal{J}_k| \geq \frac{t^{1-b}-1}{CK^2\ln(t)}\right\} \cap \{\forall s \in \mathcal{J}_k,\ \theta_1(s) \leq \mu_2 + \delta\}\right) \end{aligned} \tag{3.9}$$

Now, we have to bound the probability that $\theta_1(s) \leq \mu_2 + \delta$ for all s in an interval of size $\frac{t^{1-b}-1}{CK^2\ln(t)}$ in $\mathcal{I}_j$. So we apply Lemma 3.8 to get:

$$(B) \leq CK\ln(t)(\alpha_{\mu_1,\mu_2})^{\frac{t^{1-b}-1}{CK^2\ln(t)}} + C_{\lambda,\mu_1,\mu_2}\frac{CK\ln(t)}{\left(\frac{t^{1-b}-1}{CK^2\ln(t)}\right)^{\lambda}} e^{-jd_{\lambda,\mu_1,\mu_2}} := f(\mu_1,\mu_2,b,j,t).$$

Choosing the same b as in (3.5), we get that $\sum_{t\geq 1}\sum_{1\leq j\leq t^b} f(\mu_1,\mu_2,b,j,t) < +\infty$. It follows that for this value of b, (3.8) holds and the induction is complete.

Step 8: Conclusion. Let b be the constant chosen in Step 2. From the decomposition (3.4) and the two upper bounds (3.6) and (3.7), we get, for $t \geq N_{\mu_1,\mu_2,b}$:

$$\mathbb{P}(E_j) \leq (K-2)\left(\frac{2(K-1)}{t^2} + f(\mu_1,\mu_2,b,j,t))\right) + \frac{2(K-1)}{t^2} + g(\mu_1,\mu_2,b,j,t).$$

Recalling (3.3), summing over the possible values of j and t we obtain:

$$\begin{aligned}\sum_{t\geq 1}\mathbb{P}(N_1(t) \leq t^b) \leq & N_{\mu_1,\mu_2,b} + 2(K-1)^2\sum_{t\geq 1}\frac{1}{t^{2-b}} \\ &+ \sum_{t\geq 1}\sum_{j=1}^{t^b}[Kf(\mu_1,\mu_2,b,j,t) + g(\mu_1,\mu_2,b,j,t)] < C_{\mu_1,\mu_2,b}\end{aligned}$$

for some constant $C_{\mu_1,\mu_2,b} < \infty$.

3.3 Thompson Sampling for Exponential families

Just like the analysis of Bayes-UCB, the finite-time analysis we gave in the previous section for Thompson Sampling strongly relies on the specific properties of Beta posterior with integer coefficients, and can therefore only be applied to Bernoulli rewards and uniform prior. In the paper [Korda et al., 2013] we propose a different analysis, that is suited for exponential family bandit models, with the use of the Jeffreys' prior. While providing the full paper in Appendix B, we present in this Section the outline of this new analysis, trying to highlight the similarities and differences with our previous work. The notation used in this Section is consistent with the notation used in all the thesis, whereas that used in Appendix B are slightly different.

3.3.1 Thompson Sampling with Jeffreys' prior for general one-parameter canonical exponential families

In this work, we consider bandit models in which each arm belong to one-parameter canonical exponential family, for which the density of ν_θ is given by

$$f(x|\theta) = A(x)\exp(T(x)\theta - b(\theta)). \tag{3.10}$$

This is the general definition of one-parameter canonical exponential families given for example by [Bickel and Doksum, 2001], even if so far we only considered particular cases in which the *sufficient statistic* T is such that $T(x) = x$. Most of the examples of practical interest actually belong to this sub-class, but in Appendix B we give two examples with more general sufficient statistics: Pareto and Weibull distributions.

The properties of these exponential families include

$$\dot{b}(\theta) = \mathbb{E}_{X\sim\nu_\theta}[T(X)] \quad \text{and} \quad \ddot{b}(\theta) = \mathrm{Var}_{X\sim\nu_\theta}[T(X)].$$

Whereas these distributions can still be parameterized by their means $\mu(\theta)$, a natural alternative parametrization consists in using the mean of the sufficient statistic, $\tilde{\mu}(\theta) = \mathbb{E}_{X\sim\nu_\theta}[T(X)]$. For a given exponential family, one can introduce the divergence associated as a function of this new parameter:

$$d(\tilde{\mu}, \tilde{\mu}') = \mathrm{KL}\left(\nu_{\dot{b}^{-1}(\tilde{\mu})}, \nu_{\dot{b}^{-1}(\tilde{\mu}')}\right).$$

When $T(x) = x$, $\tilde{\mu}(\theta) = \mu(\theta)$, so this definition of the divergence function d coincides with the definition given in Chapter 1 in this particular case.

For $X \sim \nu_\theta$ with mean of sufficient statistic $\tilde{\mu}$, introducing $\tilde{\phi}_X(\lambda) = \log\mathbb{E}\left[e^{\lambda T(X)}\right]$ and $\tilde{\phi}^*_X(x)$ its Fenchel-Legendre transform (convex conjugate function), one has $d(x, \tilde{\mu}) = \tilde{\phi}^*_X(x)$. This important property is given in Lemma 1.4 of Chapter 1 for exponential families such that $T(x) = x$, for which it allows to build confidence interval based on KL-divergence. Similarly, for more general exponential families, it is possible to build KL-confidence intervals for the mean of the sufficient statistic $\tilde{\mu}$, and to define and analyse the associated KL-UCB algorithm as the index policy associated to

$$u_a(t) = \sup\left\{\tilde{q} \geq \frac{1}{N_a(t)}\sum_{i=1}^{N_a(t)} T(Y_{a,i}) : N_a(t)d\left(\frac{1}{N_a(t)}\sum_{i=1}^{N_a(t)} T(Y_{a,i}), \tilde{q}\right) \leq f(t)\right\}.$$

For exponential families defined by (3.10), it is also easy to implement the Thompson Sampling algorithm, since the posterior distribution still has an explicit form. If the prior distribution on the parameter θ has density $h_0(\theta)$, the posterior distribution after n observations $y_1, \ldots, y_n$ is given by

$$p(\theta|y_1, \ldots, y_n) \propto h_0(\theta)\exp\left(\theta\sum_{i=1}^{n} T(y_i) - nb(\theta)\right).$$

We consider in the sequel the particular implementation of Thompson Sampling that uses the Jeffreys' prior. This prior, introduced by [Jeffreys, 1946] is non-informative in the sense that it is invariant under re-parametrization of the parameter space. It can be shown to be proportional to the square root of the Fischer information $I(\theta)$, which is equal to $\ddot{b}(\theta)$ in our particular case. Thompson Sampling thus draws for each arm a a sample

$$\theta_a(t) \sim \pi_a^t \propto \sqrt{\ddot{b}(\theta)}\exp\left(\theta\sum_{i=1}^{N_a(t)} T(Y_{a,i}) - N_a(t)b(\theta)\right)$$

and chooses at time $t+1$ arm $A_{t+1} = \operatorname{argmax}_a \mu(\theta_a(t))$.

We give in Appendix B examples of implementation for classical distributions in an exponential family with alternative parametrization. For example for Bernoulli bandits, the Jeffreys' prior on the mean is $\mathrm{Beta}(1/2, 1/2)$ and does not coincide with the uniform prior always considered so far. This version of Thompson Sampling thus draws samples from $\mathrm{Beta}(1/2 + S_a(t), 1/2 + N_a(t) - S_a(t))$ and cannot be analysed with our previous tools. In many other cases, Jeffreys' prior turns out to be an improper prior ($\int_\Theta \sqrt{\ddot{b}(\theta)}d\theta = +\infty$), but after one observation, the resulting posterior becomes a probability distribution, thus Thompson Sampling can be implemented with an initialization phase drawing each arm once.

3.3.2 Main result and sketch of the proof

Theorem 3.10 is the main result proved in the paper [Korda et al., 2013]. It implies that Thompson Sampling using the Jeffreys' prior is asymptotically optimal when the rewards distributions belong to an exponential family defined by (3.10). We use the shorthand $\mathrm{K}(\theta, \theta')$ to refer to the Kullback-Leibler divergence between the distributions ν_θ and $\nu_{\theta'}$.

Theorem 3.10. *Assume that $\mu_1 > \mu_a$ for all $a \neq 1$, the prior distribution Π_J is such that for all a $\pi_{a,0}$ is taken to be the Jeffreys' prior over Θ. Then for every $\epsilon > 0$ there exists a constant $C(\epsilon, \mathcal{P})$ depending on ϵ and on the problem $\mathcal{P}$ such that the regret of Thompson Sampling using the Jeffreys' prior satisfies*

$$R_\theta(T, TS_{\pi_J}) \leq \frac{1+\epsilon}{1-\epsilon}\left(\sum_{a=2}^{K} \frac{(\mu_1 - \mu_a)}{\mathrm{K}(\theta_a, \theta_1)}\right)\ln(T) + C(\epsilon, \mathcal{P}).$$

Our analysis relies on three main ingredients: a new decomposition inspired by the one proposed by [Agrawal and Goyal, 2013a], a non-asymptotic upper bound on the tail of the posterior distribution provided that the sufficient statistics are well concentrated (Theorem B.4 in Appendix B) and a result akin to Lemma 3.2 that controls the number of draws of the optimal arm.

When Jeffreys' prior is improper, the statement of the posterior concentration result is quite involved, and requires to introduce events (denoted by $\tilde{E}_{a,t}$ in Appendix B) with complicated expression. For the sake of clarity, we present below the outline of our finite-time analysis in cases where Jeffreys' prior is proper. The general case is dealt with in Appendix B.

A new decomposition. [Agrawal and Goyal, 2013a] propose an alternative finite-time analysis that leads to the asymptotic optimality of Thompson Sampling with a uniform prior in Bernoulli bandit models. The decomposition we use here is close to the one they introduce.

For all $a = 1 \dots K$ let $\delta_a > 0$ be fixed and let $E_a(t) (= E_a(t, \delta_a))$ be the event defined by

$$E_a(t) = \left(N_a(t) \neq 0 \Rightarrow \left|\frac{1}{N_a(t)}\sum_{s=1}^{N_a(t)} T(Y_{a,s}) - b'(\theta_a)\right| \leq \delta_a\right).$$

For $a \neq 1$, let $\Delta_a < \mu_1 - \mu_a$ be fixed and let $E_a^\theta(t) (= E_a^\theta(t, \Delta_a))$ be the event

$$E_a^\theta(t) = \left(\mu(\theta_a(t)) \leq \mu_a + \Delta_a\right).$$

On $E_a(t)$ the empirical sufficient statistic is well concentrated around its mean and on $E_a^\theta(t)$ the sample from π_a^t used in the algorithm does not lead to an over-estimation of the true mean μ_a. The conjunction

of these two events holds with high probability. Now we introduce the following decomposition:

$$\mathbb{E}\left[N_a(T)\right] = \underbrace{\sum_{t=0}^{T-1} \mathbb{P}\left(A_{t+1}=a, E_a(t), E_a^\theta(t)\right)}_{(A)} + \underbrace{\sum_{t=0}^{T-1} \mathbb{P}\left(A_{t+1}=a, E_a(t), (E_a^\theta(t))^c\right)}_{(B)} + \underbrace{\sum_{t=0}^{T-1} \mathbb{P}\left(A_{t+1}=a, (E_a(t))^c\right)}_{(C)}.$$

Term (C) only concerns the concentration of empirical sufficient statistics around their means and standard concentration techniques show that it is bounded. To deal with term (B), one needs to introduce the posterior concentration result below.

Lemma 3.11. *Let $\pi(n,s)$ denote the posterior distribution under Jeffreys' prior and n observations such that the sum of sufficient statistics is $s = \sum_{i=1}^n T(y_i)$. If Jeffreys' prior is proper and*

$$\left(\left|\frac{s}{k} - \dot{b}(\theta_a)\right| \le \delta\right),$$

there exists three constants N_a, $C_{1,a}$ and $C_{2,a} = C_{2,a}(\Delta)$ such that for $k \ge N_a$,

$$\begin{aligned}
\mathbb{P}_{\theta\sim\pi(k,s)}\left(\mu(\theta) > \mu_a + \Delta\right) &\le C_{1,a} k \exp(-(k-1)(1-\delta C_{2,a})\mathrm{K}(\theta_a, \mu^{-1}(\mu_a+\Delta))) \\
\mathbb{P}_{\theta\sim\pi(k,s)}\left(\mu(\theta) < \mu_a - \Delta\right) &\le C_{1,a} k \exp(-(k-1)(1-\delta C_{2,a})\mathrm{K}(\theta_a, \mu^{-1}(\mu_a+\Delta)))
\end{aligned}$$

Using Lemma 3.11, it can be proved (see Section B.8) in Appendix B) that for all $\epsilon > 0$, there exists a constant $C_1(\epsilon, \boldsymbol{\theta}, \delta, \Delta_a)$ such that

$$(B) \le \frac{\log(T)}{(1-\epsilon)(1-\delta_a C_{2,a})\mathrm{K}(\theta_a, \mu^{-1}(\mu_a+\Delta_a))} + C_1(\epsilon, \boldsymbol{\theta}, \delta_a, \Delta_a).$$

Term (A) is the most delicate to control: when the sample from arm a used by the algorithm does not over-estimate the mean μ_a, it should be explained why arm a should not be drawn to much. To do so, we show that this implies some event regarding the optimal arm that hold with small probability since this arm has been drawn a lot. Indeed, one can show a deviation result similar to Proposition 3.2, stated as Proposition 3.12. The proof of this result is essentially the same as that of Proposition 3.2, except that we replace Lemma 3.8 (in which constants are computed explicitly) by an asymptotic argument based on the posterior concentration phenomenon.

Proposition 3.12. *For all $b \in]0,1[$, $\sum_{t=1}^\infty \mathbb{P}(N_1(t) \le t^b) \le +\infty$.*

Term (A) can be upper bounded in the following way. As Proposition 3.12 holds for any value of b we apply it for example for $b = 1/2$. Let $\Delta'_a = \mu_1 - \mu_a - \Delta_a$.

$$\begin{aligned}
(A) &\le \sum_{t=0}^{T-1} \mathbb{P}(A_{t+1}=a, E_a^\theta(t), N_1(t) > \sqrt{t}) + C_{1/2} \le \sum_{t=0}^{T-1} \mathbb{P}(\mu(\theta_1(t)) \le \mu_1 - \Delta'_a, N_1(t) > \sqrt{t}) + C_{1/2} \\
&\le \sum_{t=0}^{T-1} \mathbb{P}(\mu(\theta_1(t)) \le \mu_1 - \Delta'_a, E_1(t), N_1(t) > \sqrt{t}) + \sum_{t=0}^{T-1} \mathbb{P}((E_1(t))^c, N_1(t) > \sqrt{t}) + C_{1/2}
\end{aligned}$$

Whereas the second sum in the last display is upper bounded using classical concentration techniques, for the first term, we need the second statement of Lemma 3.11 to show

$$\mathbb{P}(\mu(\theta_1(t)) \leq \mu_1 - \Delta'_a | \mathcal{F}_t) \mathbb{1}_{E_1(t)} \leq C_{1,1} e^{-(N_1(t)-1)(1-\delta_1 C_{2,1})\mathrm{K}(\theta_1, \mu^{-1}(\mu_1 - \Delta'_a)) + \log(N_1(t))}$$

and finally lower bound term (A) by some constant plus the series

$$\sum_{t=1}^{\infty} C_{1,1} e^{-(\sqrt{t}-1)(1-\delta_1 C_{2,1})\mathrm{K}(\theta_1, \mu^{-1}(\mu_1 - \Delta'_a)) + (1/2)\log(t)} < +\infty.$$

To conclude, one has shown that there exists a constant $\mathcal{C}(\epsilon, \boldsymbol{\theta}, \delta_a, \Delta_a)$ such that

$$\mathbb{E}[N_a(T)] \leq \frac{\ln(T)}{(1-\delta_a C_{2,a}) K(\theta_a, \mu^{-1}(\mu_a + \Delta_a))(1-\epsilon)} + \mathcal{C}(\epsilon, \boldsymbol{\theta}, \delta_a, \Delta_a)$$

The constant is of course increasing (dramatically) when δ_a goes to zero, Δ_a to $\mu_1 - \mu_a$, or ϵ to zero. But one can choose Δ_a close enough to $\mu_1 - \mu_a$ and δ_a small enough, such that

$$(1 - C_{2,a}(\Delta_a)\delta_a)\mathrm{K}(\theta_a, \mu^{-1}(\mu_a + \Delta_a)) \geq \frac{\mathrm{K}(\theta_a, \theta_1)}{(1+\epsilon)},$$

and this choice leads to

$$\mathbb{E}[N_a(T)] \leq \frac{1+\epsilon}{1-\epsilon} \frac{\ln(T)}{\mathrm{K}(\theta_a, \theta_1)} + \mathcal{C}(\epsilon, \boldsymbol{\theta}, \delta_a, \Delta_a).$$

3.4 Numerical experiments and discussion

We illustrate here the performance of Thompson Sampling on numerical experiments with Bernoulli rewards. We compare both the regret and Bayes risk of Thompson Sampling to those of state-of-the-art frequentist algorithms and Bayes-UCB.

3.4.1 Regret of Thompson Sampling

We start by comparing, on several two-armed bandit models and for a quite small horizon ($T = 1000$), the regret of Thompson Sampling, Bayes-UCB, KL-UCB and its two variants KL-UCB$^+$ and KL-UCB-H$^+$ and the FH-Gittins algorithms. Preliminary experiments in Chapter 2 showed that FH-Gittins, that we believe to be a good approximation of the Bayesian optimal solution, displays good performance for several fixed bandit models. Here we investigate this trend further, adding also a comparison with Thompson Sampling. Results are reported in Figure 3.2. On the four different two-armed bandit considered with small (left) and high mean rewards (right), we see that FH-Gittins compares well to asymptotically optimal algorithms, even if it does not always outperform them. In particular, the performance of FH-Gittins seems to deteriorate on problems with high mean rewards. On this small scale, we also see that Thompson Sampling and Bayes-UCB do not always outperform KL-UCB, and that Bayes- UCB performs better than Thompson Sampling. This trend seems to be reversed when we consider larger horizons.

In our second experiment, we study regret up to some larger horizon $T = 20000$ for a 10-armed bandit problem, already studied by [Cappé et al., 2013], with means given by

$$\mu = (0.1, 0.05, 0.05, 0.05, 0.02, 0.02, 0.02, 0.01, 0.01, 0.01). \tag{3.11}$$

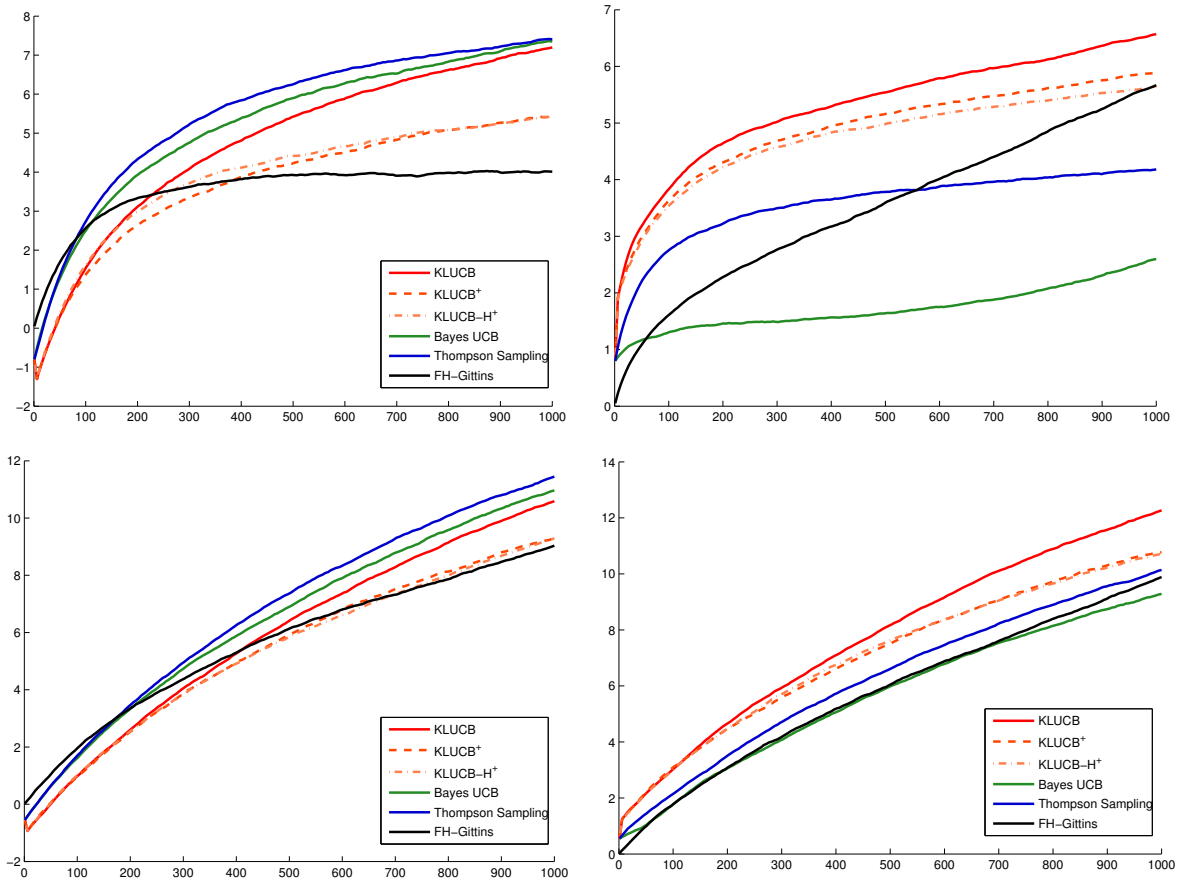

Figure 3.2: Regret of several algorithms including FH-Gittins on four different two-armed bandit problems: 0.05-0.15 (top, left) 0.85-0.95 (top, right) and two more difficult problems, 0.2-0.25 (bottom, left) and 0.75-0.8 (bottom, right). Regret is estimated based on $N = 10000$ simulations.

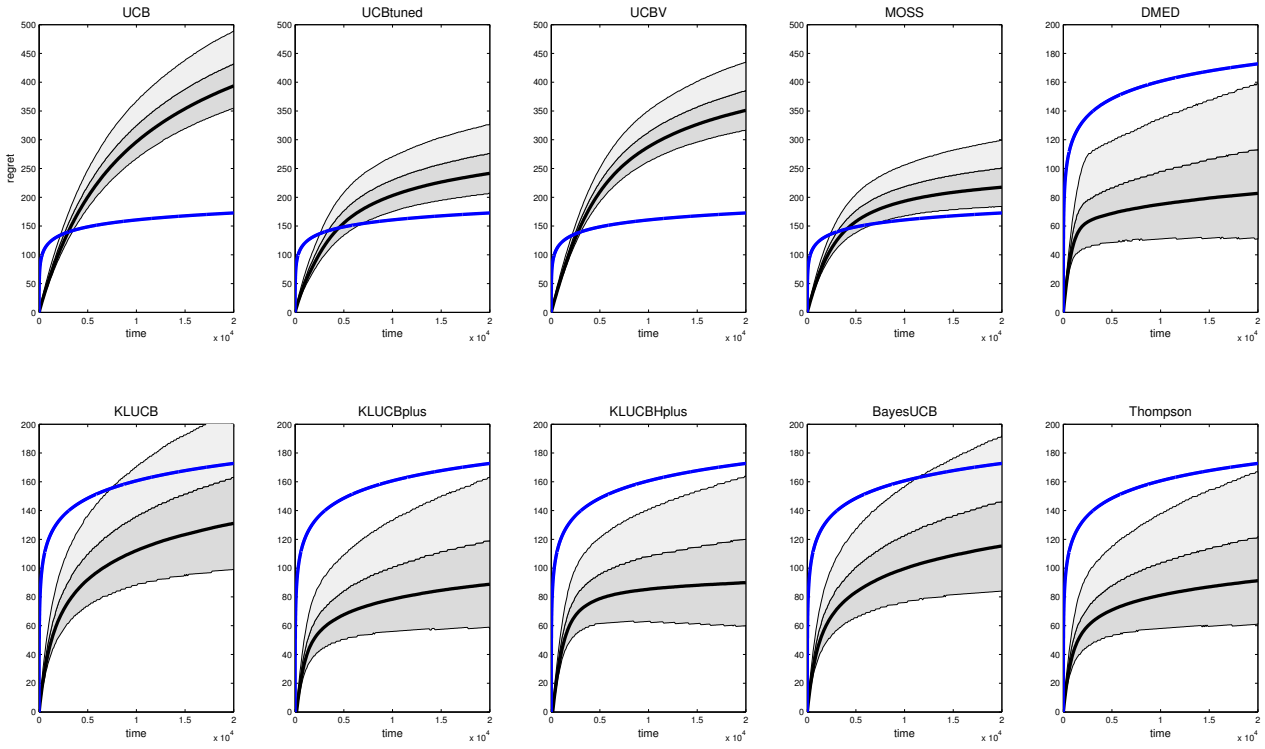

Figure 3.3: Regret of the various algorithms as a function of time. On each graph, the blue line shows the lower bound, the solid bold curve corresponds to the mean regret while the dark and light shaded regions show respectively the central 99% and the upper 0.05%

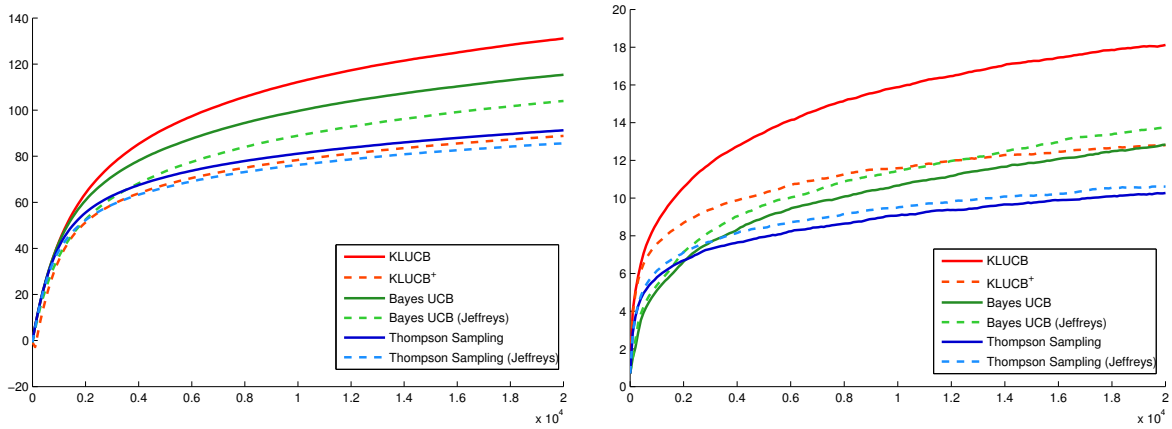

Figure 3.4: Influence of the prior. Regret for the 10-armed problem (3.11) (left) and the two-armed problem with means 0.8-0.9 (right), averaged over $N = 50000$ simulations.

Figure 3.3 displays for several algorithms an estimation of the distribution of the cumulative regret based on $N = 50000$ trials. The first four algorithms are variants of UCB, displayed using the same scale, that are not known to be optimal. Of these, the UCB-V algorithm of [Audibert et al., 2009] is close to the index policy to which Thompson Sampling is compared in [Chapelle and Li, 2011] in the Bernoulli setting. This algorithm incorporates an estimation of the variance of the rewards in the index which is defined to be, for an arm that have produced k rewards in n draws,

$$\frac{k}{n} + \sqrt{\frac{2\log(t)}{n}\frac{k}{n}\left(1-\frac{k}{n}\right)} + \frac{3\log(t)}{n}$$

UCB-Tuned is an heuristic proposed by [Auer et al., 2002a] that also uses estimates of the variance, whereas MOSS ([Audibert and Bubeck, 2010]) is a variant of UCB using an alternative exploration rate that is reminiscent of KL-UCB-H$^+$: the $\log(t)$ in UCB is replaced by $\log(T/(KN_a(t)))$.

The other six algorithms displayed in Figure 3.3, on a different scale, have a mean regret closer to (sometimes smaller than) the lower bound of Lai and Robbins, displayed in blue, which we recall is only asymptotic. All these algorithms are provably asymptotically optimal. Among them, Thompson Sampling outperforms Bayes-UCB and on this specific bandit model its performance is comparable to the best frequentist algorithms, KL-UCB$^+$ and KL-UCB-H$^+$. It is also the easiest optimal policy to implement, since at each round the indices computed by KL-UCB (and its variants) and even the quantiles computed by Bayes-UCB are more costly than the production of posterior samples.

In Section 3.2 and 3.3 we proved the asymptotic optimality of two versions of Thompson Sampling: the first using the uniform prior (Beta$(1,1)$) and the second the Jeffreys' prior, which corresponds to a Beta $(1/2, 1/2)$ prior distribution over each mean. In Figure 3.4, we compare the two resulting algorithms. While for the 10-armed bandit studied before using Jeffreys' prior appears to reduce the regret for both Bayes-UCB and Thompson Sampling, one can find other problems (like the two-armed bandit with means 0.8 and 0.9) for which it does not. Besides, the performance of the two variants is quite close on both problems presented in Figure 3.4. Thus we may conjecture that for distributions depending on a single parameter, the prior distribution chosen has little influence on the asymptotic optimality of Thompson Sampling. However, [Honda and Takemura, 2014] show that this conjecture in not true for the particular (two-parameter) case of Gaussian distributions with unknown mean and variance. Indeed, the authors prove that for a choice of prior $\pi_a^0(\mu_a, \sigma_a) \sim (\sigma_a)^{-1-2\alpha}$ with $\alpha < 0$, Thompson Sampling is asymptotically optimal (i.e. its regret matches the lower bound of Theorem 1.2), whereas for $\alpha \geq 0$, the regret is not asymptotically logarithmic.

3.4.2 Bayes risk of Thompson Sampling

As discussed in Chapter 1, KL-UCB-H$^+$ is asymptotically optimal with respect to the Bayes risk, in the sense of the lower bound of [Lai, 1987]. Theorem 1.16 also shows that KL-UCB is close to optimal since its Bayes risk is within a multiplicative factor 2 of the lower bound. Compared to the experiments in Chapter 1, we propose here experiments for a larger horizon, including also Bayes-UCB and Thompson Sampling using the uniform prior. This large horizon does not allow for a comparison with FH-Gittins, that would be extremely heavy to implement.

For $K = 5$ arms (left plot in Figure 3.5) and $K = 10$ arms (right plot in Figure 3.5), for each algorithm we approximate the Bayes risk up to horizon $T = 20000$ by sampling $N = 50000$ bandit models with K arms from the prior distribution and playing the algorithm on each bandit model up to horizon T. As already observed on smaller horizons, KL-UCB$^+$ appears as a good anytime approximation of KL-UCB-H$^+$. Meanwhile, the gap between KL-UCB and KL-UCB-H$^+$ increases, indicating KL-UCB might

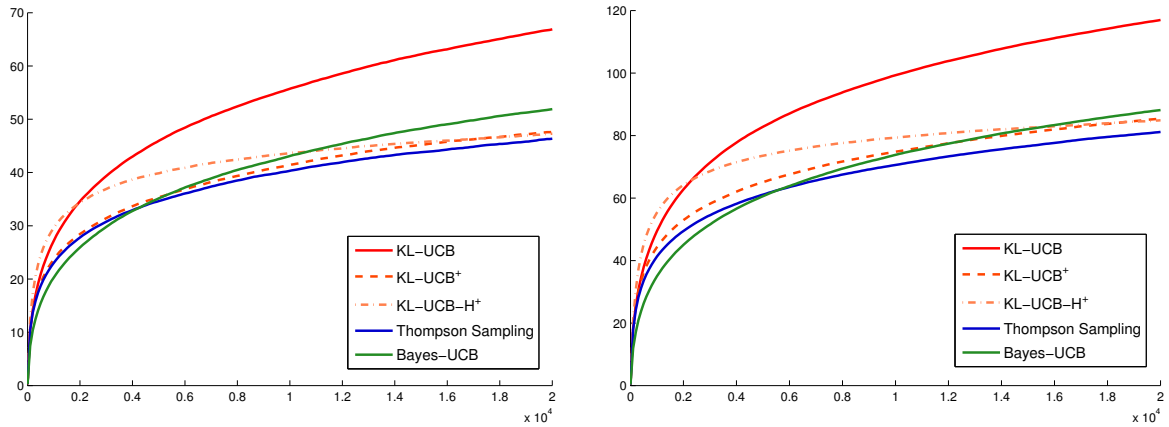

Figure 3.5: Bayes risk of several algorithm with a uniform prior distribution on the means of the 5 arms (left) or 10 arms (right).

not be asymptotically optimal. As for the Bayesian algorithms, Thompson Sampling outperforms KL-UCB-H$^+$ and Bayes-UCB, despite its slightly larger Bayes risk, still appears as a fair competitor to KL-UCB-H$^+$ and KL-UCB$^+$. We therefore conjecture that Bayes-UCB and Thompson Sampling are good approximations of the Bayesian optimal policy, at least with a uniform prior. Recently, [Guha and Munagala, 2014] have investigated the Bayesian optimality of Thompson Sampling, showing that for two-armed bandits, with arbitrary prior Π_0, for all T,

$$\mathbb{E}_{\Pi_0}\left[N_b^{\mathrm{TS}}(T)\right] \leq 2 \times \min_{\mathcal{A}} \mathbb{E}_{\Pi_0}\left[N_b^{\mathcal{A}}(T)\right],$$

where $N_a^{\mathcal{A}}(t)$ denotes the number of draws of arm a by the algorithm $\mathcal{A}$ up to time t and b denotes the (random) suboptimal arm. As the Bayes risk of an algorithm in that case is $\mathbb{E}_{\Pi_0}[(\mu^* - \mu_b)N_b^{\mathcal{A}}(T)]$, this result does not exactly state that the Bayes risk of Thompson Sampling is within a multiplicative factor 2 of the Bayes-risk of the Bayesian optimal solution, but it still provides good performance guarantees for Thompson Sampling in the Bayesian framework

On the two bandit problems displayed in Figure 3.5, one can check that the prior-independent upper bound on the Bayes risk given by [Bubeck and Liu, 2013] is quite pessimistic: $14\sqrt{TK}$ is more than one hundred times the actual regret obtained when there are ten arms. However, this upper bound holds for any bandit model with a finite number of arms and any prior distribution π_0 on reward distributions bounded in $[0, 1]$, which includes more general situations with possibly correlated arms.

3.4.3 Thompson Sampling in more general frameworks

In the Bernoulli case, we have seen that Thompson Sampling is the easiest to implement asymptotically optimal policy. This computational advantage will be even stronger in more complex models, in which it might not be possible to design a UCB-like algorithm, or for which a complicated prior distribution is used, such that the associated posterior distributions can only be sampled from using MCMC simulation. In the latter case, an (approximate) implementation of Bayes-UCB needs several samples from the posterior distribution to estimate the quantiles, whereas Thompson Sampling only needs to produce one sample per round.

The performance of Thompson Sampling beyond one-parameter exponential family bandits has been

recently investigated. [Honda and Takemura, 2014] propose the first analysis of Thompson Sampling in the particular case of Gaussian bandit models with unknown means and variances. They prove that this algorithm is asymptotically optimal for some choices of independent prior distributions (but not for all the possible choices of independent priors, as explained above). [Bubeck and Liu, 2013] study Thompson Sampling for Gaussian bandits with known variance but for a special form of non-independent prior. For two-armed bandit models such that the means μ_1, μ_2 are known up to a permutation, they show that Thompson Sampling has a finite regret. This regret has an optimal dependency in $|\mu_1 - \mu_2|$, in a sense specified by [Bubeck et al., 2013a], who study the particular setting in which the mean of the best arm and (a lower bound on) the gap between the best and second best means are known. Recently, [Gopalan et al., 2014] have proposed the first analysis for Thompson Sampling that holds for quite general bandit problems. The authors derive a logarithmic upper bound on the regret that involves a (non-explicit) constant that captures correlations between arms. However, this bound holds for finitely supported arms and prior distribution, and it would be interesting to investigate whether these assumptions could be relaxed to recover the upper bound obtained in the Bernoulli case with independent arms. Thompson Sampling has also been successfully used for bandit problems with switching environments (see [Mellor and Shapiro, 2013]), yet without theoretical guarantees.

As already mentioned in the Introduction, the good empirical performance of Thompson Sampling in contextual bandit models was known before any theoretical guarantee even in the Bernoulli case was available. Chapter 4 is dedicated to the presentation of contextual bandit models. We will notably review (and prove new) regret and Bayes risk upper bounds for Thompson Sampling in this setting. Thompson Sampling has also been successfully used in the more general framework of reinforcement learning. In model-based reinforcement learning, under some assumptions on the transition and reward functions in a Markov Decision Process, the goal is to design algorithms using estimates of these functions that act (almost) optimally in the MDP. Optimistic approaches (building set of statistically plausible MDPs and acting as in the best possible MDP) have been considered, leading to the UCRL2 ([Jaksch et al., 2010]) or KL-UCRL ([Filippi et al., 2010a]) algorithms. [Strens, 2000] introduces the following algorithm, inspired by Thompson Sampling, that consists in several episodes. A prior distribution over the rewards and transitions is maintained. At the beginning of each episode a MDP is drawn from the current posterior distribution, the optimal policy for this sampled MDP is computed and played until the end of the episode. Recently [Osband et al., 2013] have provided the first (Bayesian) regret guarantees for this algorithm, while illustrating its practical performance. Other algorithms, sampling several MDPs at each round (or at the beginning of each episode) have been considered. [Asmuth et al., 2009] merge the samples obtained into a mixed MDP, whereas [Fonteneau et al., 2013] suggest to combine Thompson Sampling with the optimism principle: at each round, the best action in the best possible sampled MDP is chosen.

3.5 Elements of proof

3.5.1 Proof of Lemma 3.8

On the event $\{\mathcal{J} \subseteq [\tau_j, \tau_{j+1}[\}$, for all $s \in \mathcal{J}$ the posterior distribution $\pi_1^s = \pi_1^{\tau_j}$ is fixed and the $(\theta_1(s))$ are then, when conditioned on $S_1(\tau_j)$, an i.i.d. sequence with common distribution $\text{Beta}(S_1(\tau_j) + 1, j -$

$S_1(\tau_j)+1)$. Thus, one can write

$$\begin{aligned}\mathbb{P}\left((\mathcal{J}\subseteq[\tau_j,\tau_{j+1}[)\cap(\forall s\in\mathcal{J},\theta_1(s)\le\mu_2+\delta)\right) &= \mathbb{E}\Big[\mathbb{E}\Big[\mathbb{1}_{(\mathcal{J}\subseteq[\tau_j,\tau_{j+1}[)}\mathbb{1}_{(\forall s\in\mathcal{J},\theta_1(s)\le\mu_2+\delta)}\Big|\tau_j,S_1(\tau_j)\Big]\Big]\\ &\le \mathbb{E}\Big[\mathbb{E}\Big[\mathbb{1}_{(\forall s\in\mathcal{J},\tilde{\theta}_1(s)\le\mu_2+\delta)}\Big|\tau_j,S_1(\tau_j)\Big]\Big]\end{aligned}$$

where $\tilde{\theta}_1(s)$ is an i.i.d. sequence conditionally to $S_1(\tau_j)$ with distribution Beta$(S_1(\tau_j)+1, j-S_1(\tau_j)+1)$. Using that $\mathcal{J}$ is τ_j-measurable,

$$\begin{aligned}\mathbb{P}\left((\mathcal{J}\subseteq[\tau_j,\tau_{j+1}[)\cap(\forall s\in\mathcal{J},\theta_1(s)\le\mu_2+\delta)\right) &\le \mathbb{E}\left[\left(F^{\text{Beta}}_{(S_1(\tau_j)+1,j-S_1(\tau_j)+1)}(\mu_2+\delta)\right)^{|\mathcal{J}|}\right]\\ &\le \mathbb{E}\left[\left(1-F^B_{(j+1,\mu_2+\delta)}(S_1(\tau_j))\right)^{f(t)}\right],\end{aligned}$$

where we use the link between the tail of Beta and Bernoulli distribution mentioned above and the fact that $|\mathcal{J}|\ge f(t)$. It remains to upper bound this last expectation. An exact computation yields

$$\mathbb{E}\left[(1-F^B_{(j+1,\mu_2+\delta)}(S_1(\tau_j)))^{f(t)}\right]=\sum_{s=0}^{j}(1-F^B_{(j+1,\mu_2+\delta)}(s))^{f(t)}f^B_{j,\mu_1}(s)$$

To simplify notation, from now on let $y=\mu_2+\delta$. Using, as [Agrawal and Goyal, 2012], that

$$F^B_{j+1,y}(s)=(1-y)F^B_{j,y}(s)+yF^B_{j,y}(s-1)\ge(1-y)F^B_{j,y}(s),$$

we get:

$$(1-F^B_{(j+1,y)}(s))^{f(t)}\le\exp\left(-f(t)F^B_{(j+1,y)}(s)\right)\le\exp\left(-f(t)(1-y)F^B_{(j,y)}(s)\right)$$

Therefore,

$$\mathbb{E}\left[(1-F^B_{(j+1,\mu_2+\delta)}(S_1(\tau_j)))^{f(t)}\right]\le\sum_{s=0}^{j}\exp\left(-f(t)(1-y)F^B_{(j,y)}(s)\right)f^B_{j,\mu_1}(s)$$

Using the fact that for $s\ge\lceil yj\rceil$, $F^B_{j,y}(s)\ge\frac{1}{2}$ (since the median of a binomial distribution with parameters j and y is $\lceil yj\rceil$ or $\lfloor yj\rfloor$), we get

$$\begin{aligned}&\mathbb{E}\left[(1-F^B_{(j+1,\mu_2+\delta)}(S_1(\tau_j)))^{f(t)}\right]\\ &\le\sum_{s=0}^{\lfloor jy\rfloor}\exp\left(-f(t)(1-y)F^B_{(j,y)}(s)\right)f^B_{j,\mu_1}(s)+\sum_{s=\lceil jy\rceil}^{j}\left(\frac{1}{2}\right)^{(1-y)f(t)}f^B_{j,\mu_1}(s)\\ &\le\underbrace{\sum_{s=0}^{\lfloor jy\rfloor}\exp\left(-f(t)(1-y)F^B_{(j,y)}(s)\right)f^B_{j,\mu_1}(s)}_{E}+\left(\frac{1}{2}\right)^{(1-y)f(t)}.\end{aligned}$$

It is easy to show that for every $\lambda>1,\forall x>0, x^\lambda\exp(-x)\le\left(\frac{\lambda}{e}\right)^\lambda$ This allows us to upper-bound the exponential for all $\lambda>1$, using $C_\lambda=\left(\frac{\lambda}{e}\right)^\lambda$,by:

$$(E)\le\frac{C_\lambda}{(f(t)(1-y))^\lambda}\sum_{s=0}^{\lfloor jy\rfloor}\frac{f^B_{j,\mu_1}(s)}{\left(F^B_{(j,y)}(s)\right)^\lambda}\le\frac{C_\lambda}{(f(t)(1-y))^\lambda}\sum_{s=0}^{\lfloor jy\rfloor}\frac{f^B_{j,\mu_1}(s)}{\left(f^B_{(j,y)}(s)\right)^\lambda}$$

Now, inspired by Agrawal and Goyal's work (proof of Lemma 3) we compute:

$$\frac{f^B_{j,\mu_1}(s)}{\left(f^B_{(j,y)}(s)\right)^\lambda} = \frac{\binom{j}{s}\mu_1^s(1-\mu_1)^{j-s}}{\binom{j}{s}^\lambda (y^\lambda)^s((1-y)^\lambda)^{j-s}} \leq \frac{\mu_1^s(1-\mu_1)^{j-s}}{(y^\lambda)^s((1-y)^\lambda)^{j-s}}$$
$$= \left(\frac{1-\mu_1}{(1-y)^\lambda}\right)^j \left(\frac{\mu_1(1-y)^\lambda}{y^\lambda(1-\mu_1)}\right)^s$$

Let $R_\lambda(\mu_1, y) = \frac{\mu_1(1-y)^\lambda}{y^\lambda(1-\mu_1)}$. There exists some $\lambda_1 > 1$ such that, if $\lambda < \lambda_1$, $R_\lambda > 1$. More precisely,

$$R_\lambda > 1 \Leftrightarrow \frac{\mu_1}{1-\mu_1} > \left(\frac{y}{1-y}\right)^\lambda \Leftrightarrow \ln\left(\frac{\mu_1}{1-\mu_1}\right) > \lambda \ln\left(\frac{y}{1-y}\right)$$

and so

$$\lambda_1(\mu_1, y) = \begin{cases} \frac{\ln\left(\frac{\mu_1}{1-\mu_1}\right)}{\ln\left(\frac{y}{1-y}\right)} & \text{if } y > \frac{1}{2} \\ +\infty & \text{if } y < \frac{1}{2} \end{cases}$$

For $1 < \lambda < \lambda_1$:

$$\sum_{s=0}^{\lfloor jy \rfloor} \frac{f^B_{j,\mu_1}(s)}{(f^B_{(j,\mu_2+\delta)}(s))^\lambda} \leq \left(\frac{1-\mu_1}{(1-y)^\lambda}\right)^j \sum_{s=0}^{\lfloor jy \rfloor} R_\lambda^s = \left(\frac{1-\mu_1}{(1-y)^\lambda}\right)^j \frac{R_\lambda^{\lfloor jy \rfloor + 1} - 1}{R_\lambda - 1}$$
$$\leq \left(\frac{1-\mu_1}{(1-y)^\lambda}\right)^j \frac{R_\lambda}{R_\lambda - 1} R_\lambda^{jy} = \frac{R_\lambda}{R_\lambda - 1}\left(\frac{1-\mu_1}{(1-y)^\lambda}\right)^{j-jy}\left(\frac{\mu_1}{y^\lambda}\right)^{jy}$$
$$= \frac{R_\lambda}{R_\lambda - 1} e^{-j d_\lambda(y,\mu_1)}$$

where $d_\lambda(y, \mu_1) = y \ln\left(\frac{y^\lambda}{\mu_1}\right) + (1-y)\ln\left(\frac{(1-y)^\lambda}{1-\mu_1}\right)$. Rearranging we can write

$$d_\lambda(y, \mu_1) = \lambda\left[y\ln(y) + (1-y)\ln(1-y)\right] - \left[y\ln(\mu_1) + (1-y)\ln(1-\mu_1)\right]$$

which is an affine function of λ with negative slope ($y\ln(y) + (1-y)\ln(1-y) < 0$ for all $y \in (0,1)$) and $d_1(y, \mu_1) = K(y, \mu_1) > 0$. Hence, for fixed $0 < y < \mu_1 \leq 1$ this function is positive whenever

$$\lambda < \frac{y\ln(\mu_1) + (1-y)\ln(1-\mu_1)}{y\ln(y) + (1-y)\ln(1-y)} =: \lambda_2(\mu_1, y).$$

Clearly, $\lambda_2(\mu_1, y) > 1$ and we choose $\lambda_0 = \min(\lambda_1, \lambda_2)$. After some calculation one can show that $\lambda_2 \leq \lambda_1$, and therefore that

$$\lambda_0(\mu_1, \mu_2) = \lambda_2(\mu_1, \mu_2 + \delta) = 1 + \frac{K(\mu_2 + \delta, \mu_1)}{(\mu_2 + \delta)\ln\frac{1}{\mu_2+\delta} + (1 - \mu_2 - \delta)\ln\frac{1}{1-\mu_2-\delta}}.$$

To obtain the constants used in the statement of the lemma we define $d_{\lambda,\mu_1,\mu_2} := d_\lambda(y, \mu_1)$

$$C_{\lambda,\mu_1,\mu_2} := C_{\lambda_0}(1 - \mu_2 - \delta)^{-\lambda} \frac{R_\lambda}{1 - R_\lambda}.$$

This concludes the proof.

3.5.2 Proof of Lemma 3.9

Let $a \in \{2, \dots K\}$. Recall that $C_a = \frac{32}{\Delta_a^2}$ and $\delta_a = \frac{\Delta_a}{2}$.

$$\mathbb{P}\left(\exists s \leq t : \theta_a(s) > \mu_a + \delta_a, N_a(s) > C_a \log t\right)$$
$$\leq \underbrace{\mathbb{P}\left(\exists s \leq t : \frac{S_a(t)}{N_a(t)} > \mu_a + \frac{\delta_a}{2}, N_a(s) > C_a \log t\right)}_{A} + \underbrace{\mathbb{P}\left(\exists s \leq t : \theta_a(s) > \frac{S_a(t)}{N_a(t)} + \frac{\delta_a}{2}, N_a(s) > C_a \log t\right)}_{B}$$

Term A is easily upper bounded using a union bound and Hoeffding inequality:

$$(A) \leq \sum_{s=\lceil C_a \log t \rceil}^{t} \mathbb{P}\left(\frac{\sum_{k=1}^{s} Y_{a,s}}{s} > \mu_a + \frac{\delta_a}{2}\right) \leq \sum_{s=\lceil C_a \log t \rceil}^{t} \exp\left(-\frac{2s\delta_a^2}{4}\right) \leq t \exp\left(-\frac{C_a \log t \Delta_a^2}{8}\right) = \frac{1}{t^3} \leq \frac{1}{t^2}.$$

Term B is upper bounded as follows:

$$\begin{aligned}
(B) &\leq \sum_{s=\lceil C_a \log t \rceil}^{t} \sum_{r=1}^{s} \mathbb{P}\left(\theta_a(s) > \frac{r}{s} + \frac{\delta_a}{2} | S_a(s) = r, N_a(s) = s\right) \mathbb{P}(S_a(t) = r, N_a(t) = s) \\
&= \sum_{s=\lceil C_a \log t \rceil}^{t} \sum_{r=1}^{s} \left(1 - F^{\text{Beta}}_{r+1, s-r+1}\left(\frac{r}{s} + \frac{\delta_a}{2}\right)\right) \mathbb{P}(S_a(t) = r, N_a(t) = s) \\
&= \sum_{s=\lceil C_a \log t \rceil}^{t} \sum_{r=1}^{s} F^{\text{Bin}}_{s+1, \frac{r}{s}+\frac{\delta_a}{2}}(r) \mathbb{P}(S_a(t) = r, N_a(t) = s)
\end{aligned}$$

Introducing a sequence (Z_i) of i.i.d Bernoulli random variables with mean $\frac{r}{s} + \delta_a$, one has

$$F^{\text{Bin}}_{s+1, \frac{r}{s}+\frac{\delta_a}{2}}(r) = \mathbb{P}\left(\sum_{i=1}^{s+1} Z_i < r\right) \leq \mathbb{P}\left(\sum_{i=1}^{s+1} \left(Z_i - \frac{r}{s} - \frac{\delta_a}{2}\right) \leq -\frac{\delta_a}{2}(s+1)\right) \leq \exp\left(-2C_a \log t \frac{\delta_a^2}{4}\right) \leq \frac{1}{t^3}.$$

Hence,

$$(B) \leq \sum_{s=\lceil C_a \log t \rceil}^{t} \sum_{r=1}^{s} \frac{1}{t^3} \mathbb{P}(S_a(s) = r, N_a(s) = s) = \sum_{s=\lceil C_a \log t \rceil}^{t} \frac{1}{t^3} \leq \frac{1}{t^2}.$$

Finally a union bound over the arm yields Lemma 3.9.

Chapter 4

Bayesian algorithms for linear contextual bandits

Recall that the initial motivation for studying the stochastic multi-armed bandit problem with Bernoulli rewards was sequential allocation of medical treatments: the response of patients to each treatment are assumed to be i.i.d. binary random variables (indicating whether the patient is cured or not). This assumption is however oversimplified, since the doctor has information about both the patient and the treatments, and the optimal treatment for two patients might actually differ. One way to incorporate this side information is to consider *contextual bandit problems*. In a more recent motivation for bandit problems, online advertisement, contexts are also of utmost importance, since a lot of information on the add, user or webpage is available. For all these reasons, we chose to dedicate a chapter of this thesis to the presentation of these more general bandit models of practical interest. We especially focus on linear contextual bandit problems and we argue that Bayes-UCB and Thompson Sampling can also be used in this more general framework.

The contributions of this chapter are the following. If d is the dimension of the context space, we give upper bounds on the Bayes risk of Bayes-UCB and Thompson Sampling using a Gaussian prior on the regression parameter that scale in $O(d\sqrt{T})$ and $O(\sqrt{dT\log(K)})$ in the particular case of a finite number of contexts K. For Bayes-UCB, we give a high-probability result that is inspired by standard analysis of optimistic algorithms for linear bandits, whereas for Thompson Sampling, we use ideas introduced by [Russo and Van Roy, 2014] and provide upper bounds in expectation. Both analyses rely on the use of Bayesian confidence regions instead of frequentist confidence regions used in previous work.

Contents

4.1 Introduction

Let $\mathcal{D}_t \subset \mathbb{R}^d$ be a set of contexts (or 'contextualized actions') available at time t. In a contextual bandit model, at time t an agent chooses a context $x_t \in \mathcal{D}_t$ and receives a reward

$$y_t = f(x_t) + \epsilon_t, \tag{4.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a (unknown) real-valued function and ϵ_t some centered noise. The agent aims at maximizing the sum of his rewards.

The context set $\mathcal{D}_t$ can be seen as a set of structured actions that evolves over time. A more general contextual bandit model involves both actions (in an action set $\mathcal{A}$) and contexts (in a context set $\mathcal{C}$). At each time t, a context $c_t \in \mathcal{C}$ is revealed, the agent chooses an action $a_t \in \mathcal{A}$ an receives a reward

$$y_t = g(c_t, a_t) + \epsilon_t.$$

In the online advertisement application, the context c_t could be a feature vector relative to the t^{th} user of the website, while action a corresponds to some advertisement that could be shown to him (a could be itself a feature vector for this add). It is reasonable to assume that some feature vector for the pair user/add $x_t^a = \phi(c_t, a)$, is built for each add a available and that there exists f such that $g(c, a) = f(\phi(c, a))$. In that case, we are in the model (4.1) with $\mathcal{D}_t = (x_t^a)_{a \in \mathcal{A}}$. For example, [Chapelle et al., 2014] suggest that such coupled features are used in some add prediction systems. Besides, if the users can be clustered into different categories, one can assume that in each category, the model (4.1) holds, with $\mathcal{D}_t$ the set of feature vectors of adds available for display at time t. In the rest of this chapter, we will study only the model (4.1).

Several assumptions on the function f have been considered in the literature. If f is a linear function, that is $f(x) = x^T\theta$ with $\theta \in \mathbb{R}^d$ some unknown vector we are in a *linear contextual bandit model*, first considered by [Auer, 2002] (under the name 'associative reinforcement learning with linear value function'). [Filippi et al., 2010b] consider the richer *generalized linear bandit model* for which $f(x) = \mu(x^T\theta)$, where μ is some link function and $\theta \in \mathbb{R}^d$ some unknown parameter. [Valko et al., 2013] consider *kernelized contextual bandit models* for which $f(x) = \phi(x)^T\theta$, where x belongs to some set $\mathcal{X}$, $\phi : \mathcal{X} \to \mathcal{H}$ is a mapping to some Hilbert space $\mathcal{H}$, and $\theta \in \mathcal{H}$. In the rest of this chapter, we will consider only the linear case, and present —as for classical bandit models in previous chapters— optimistic approaches and Bayesian alternatives for the linear contextual bandit problem.

Contextual linear bandits. The contextual model mostly considered in this chapter is the following. Let θ be a parameter in $\mathbb{R}^d$. At time t, the agent chooses a context $x_t \in \mathcal{D}_t$ based on past observations, according to his strategy (or bandit algorithm) $\mathcal{A}$, and receives a reward

$$y_t = x_t^T\theta + \epsilon_t.$$

The σ-field $\mathcal{H}_t = \sigma(\mathcal{D}_1, x_1, y_1, \ldots, \mathcal{D}_t, x_t, y_t, \mathcal{D}_{t+1})$ represents the information available at the end of round t: contexts chosen and rewards observed up to the end of round t, as well as the new set of contexts

$\mathcal{D}_{t+1}$ from which the agent has to choose from at round $t+1$. The noise ϵ_t satisfies $\mathbb{E}[\epsilon_t|\mathcal{H}_{t-1}] = 0$. In a deterministic strategy, x_t is assumed to be $\mathcal{H}_{t-1}$-measurable, whereas in a randomized strategy, x_t is drawn from some distribution p_t on $\mathcal{D}_t$, such that p_t is $\mathcal{F}_{t-1}$-measurable.

The best context (or arm) at time t, i.e. the one with highest mean, is

$$x_t^* = \underset{x\in\mathcal{D}_t}{\operatorname{argmax}}\, x^T\theta$$

and the agent aims at minimizing the following random quantity, called *pseudo-regret*: [1]

$$\mathcal{R}_\theta(T,\mathcal{A}) = \sum_{t=1}^{T} r_t \quad \text{where} \quad r_t = (x_t^*)^T\theta - x_t^T\theta.$$

In the literature, this model has been often introduced in the static case in which $\forall t, \mathcal{D}_t = \mathcal{D}$ (as in the papers by [Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010] for example). In Chapter 2, Bayes-UCB was also applied to a linear contextual bandit problem with a static, finite set of contexts $\mathcal{D} = \{U_1, \dots, U_K\}$. We call this problem a 'linear bandit problem', reserving the adjective 'contextual' to changing sets of contexts.

Depending on whether the regression parameter θ is regarded as an unknown parameter or is assumed to be drawn from a prior distribution π_0 on $\mathbb{R}^d$, one may consider two different probabilistic frameworks. We denote by $\mathbb{P}$,$\mathbb{E}$ the probability and expectation under the Bayesian model (with an implicit dependency on the prior distribution π_0) and by $\mathbb{P}_\theta$, $\mathbb{E}_\theta$ the probability and expectation in the frequentist framework (that is, conditionally to θ). As for classical bandits, one defines Bayesian algorithms to be algorithms using at round t the posterior distribution on θ to make a decision. Independently, for any algorithm, Bayesian or 'frequentist' (i.e. that does not use a prior in its routine) one can carry out two types of analyses:

- a *frequentist analysis* bounds the pseudo-regret of an algorithm either in probability or in expectation, conditionally to θ. For example one can bound the *regret* (or expected regret) defined by

$$\mathrm{R}_\theta(T,\mathcal{A}) = \mathbb{E}_\theta[\mathcal{R}_\theta(T,\mathcal{A})]$$

- a *Bayesian analysis* bounds the pseudo-regret of an algorithm in the Bayesian modeling (including an average over the prior π_0). For example one can bound the *Bayes risk*, defined by

$$\mathrm{BR}_{\pi_0}(T,\mathcal{A}) = \mathbb{E}[\mathcal{R}_\theta(T,\mathcal{A})] = \mathbb{E}[\mathrm{R}_\theta(T,\mathcal{A})].$$

The goal of this chapter is to present the implementation of Bayes-UCB and Thompson Sampling in linear contextual bandit models, as well as Bayesian and frequentist analyses of these two algorithms. We make the following classical assumptions:

Assumption 1. The contexts are bounded: there exists $L > 0$ such that $\forall t \in \mathbb{N}, \forall x \in \mathcal{D}_t, ||x||_2 \leq L$.

Assumption 2. The noise is centered and σ^2-subgaussian : $\mathbb{E}_\theta[\eta_t|\mathcal{H}_{t-1}] = 0$ and

$$\forall \lambda > 0, \mathbb{E}_\theta[e^{\lambda\eta_t}|\mathcal{H}_t] = \mathbb{E}[e^{\lambda\eta_t}|\theta,\mathcal{H}_t] \leq e^{\frac{\lambda^2\sigma^2}{2}}.$$

The Bayesian analyses presented here hold under an additional assumption presented below: Gaussian prior and Gaussian noise.

1. In some papers, pseudo-regret may be called regret, and what we call regret should be called 'expected regret'. We chose these denominations to be consistent with what we call regret in previous chapters: for classical bandits, $\mathrm{R}_\theta(T,\mathcal{A})$ is the expectation of the pseudo-regret $\sum_{t=1}^{t}(\mu^* - \mu_{A_t})$.

Contextual linear bandit with Gaussian prior. A natural prior distribution on the parameter θ is a Gaussian prior with covariance $\kappa^2 I_d$; that is $\theta \sim \mathcal{N}\left(0, \kappa^2 I_d\right)$. Indeed, under this prior distribution, if one assumes that the noise is Gaussian with known variance, the posterior distribution is explicitly computable. Introducing for every $t \geq 1$ the matrix and vectors

$$X_t = \begin{pmatrix} x_1^T \\ x_2^T \\ . \\ x_t^T \end{pmatrix} \in \mathcal{M}_{t,d}(\mathbb{R}), \quad Y_t = \begin{pmatrix} y_1 \\ y_2 \\ . \\ y_t \end{pmatrix} \in \mathbb{R}^t, \quad \text{and} \quad E_t = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ \epsilon_t \end{pmatrix} \in \mathbb{R}^t,$$

one has $Y_t = X_t\theta + E_t$. If the noise is such that $\epsilon_t \sim \mathcal{N}\left(0, \sigma^2\right)$, the posterior distribution on θ at the beginning of round $t+1$ is

$$p(\theta|\mathcal{H}_t) = \mathcal{N}\left(\hat{\theta}(t), \Sigma_t\right)$$

where

$$\begin{cases} \hat{\theta}(t) &= (B(t))^{-1} X_t^T Y_t \quad \text{with} \quad B(t) = \frac{\sigma^2}{\kappa^2} I_d + X_t^T X_t \\ \Sigma_t &= \sigma^2 (B(t))^{-1}. \end{cases}$$

The posterior mean $\hat{\theta}(t)$ is the regularized least-square estimator of θ with regularization parameter $\lambda = \frac{\sigma^2}{\kappa^2}$.

The Bayes-UCB, Bayes-LinUCB and Thompson Sampling algorithms presented in Sections 4.3 and 4.4 use the Gaussian prior distribution defined above. It is sometimes possible to implement (approximations of) these algorithms using a more general prior distribution π_0 by resorting to MCMC simulation when there is no close form for the posterior distribution, as illustrated in Chapter 2 for Bayes-UCB with a sparsity-inducing prior. For our Bayesian analyses, Assumption 2 above will be replaced by the assumption of a Gaussian noise with variance σ^2:

$$y_t = \theta^T x_t + \epsilon_t, \quad \text{with} \quad \theta \sim \mathcal{N}\left(0, \kappa^2 I_d\right) \quad \text{and} \quad \epsilon_t \sim \mathcal{N}\left(0, \sigma^2\right). \tag{4.2}$$

Optimistic approaches and related works. Optimistic algorithms for linear contextual bandits build a confidence region C_t in $\mathbb{R}^d$ for the unknown parameter θ and choose

$$x_{t+1} = \underset{x \in \mathcal{D}_{t+1}}{\text{argmax}} \ \max_{\theta' \in C_t} x^T \theta'. \tag{4.3}$$

This optimism-in-face-of-uncertainty principle is implemented in the algorithms proposed by [Auer, 2002, Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010, Chu et al., 2011, Abbasi-Yadkori et al., 2011], with successive refinement on how to build the confidence region C_t. Some of the confidence regions used are built using a (potentially regularized) least-square estimate of θ. Keeping the notation $B(t)$, Σ_t and $\hat{\theta}(t)$ of the previous section (with $\kappa = +\infty$ allowed when there is no regularization), some confidence region used are of the form

$$C_t = \left(\theta' \in \mathbb{R}^d : ||\hat{\theta}(t) - \theta'||_{\Sigma_t^{-1}} \leq \beta(t+1, \delta)\right),$$

where we recall that $||v||_A = \sqrt{v^T A v}$ is the L^2-norm associated to the matrix A. For C_t of the following form, the maximum in (4.3) can be computed explicitly and the algorithm rewrites

$$x_{t+1} = \underset{x \in \mathcal{D}_{t+1}}{\text{argmax}} \ \left[\hat{\theta}(t)^T x + ||x||_{\Sigma_t} \beta(t+1, \delta)\right].$$

When the number of contexts is finite, the above algorithm appears as an index policy. It can also be implemented when $\mathcal{D}_t$ is infinite and convex as the maximization of a convex differentiable function. When $\mathcal{D}_t$ is a polytope, [Dani et al., 2008] suggest to use L^1 confidence regions.

Among theses optimistic algorithms, the one using the tightest confidence region is the OFUL algorithm of [Abbasi-Yadkori et al., 2011], which picks at time t

$$x_{t+1} = \underset{x \in \mathcal{D}_{t+1}}{\operatorname{argmax}} \left[\hat{\theta}(t)^T x + ||x||_{\Sigma_t} \left(\sqrt{2 \log \frac{1}{\delta} + d \log \left(1 + (t+1) \frac{L^2 \kappa^2}{d\sigma^2} \right)} + \frac{1}{\kappa} ||\theta|| \right) \right].$$

OFUL was introduced in a frequentist setting, in which κ is no longer seen as the parameter of a prior distribution but is a parameter of the algorithm such that $\frac{\sigma^2}{\kappa^2}$ is the regularization coefficient in the regularized least square estimate $\hat{\theta}(t)$.

[Abbasi-Yadkori et al., 2011] propose a frequentist analysis of OFUL under Assumption 1 and 2 and the additional assumption that for all $t \in \mathbb{N}^*$, for all $x \in \mathcal{D}_t$, $|x^T \theta| \leq 1$. They show that, for a fixed parameter θ, the pseudo-regret of OFUL is of order $\tilde{O}(d\sqrt{T})$ with high probability. The $\tilde{O}$ notation means that we ignore logarithmic factors in T. In the static case, that is when $\mathcal{D}_t = \mathcal{D}$, there exists some (worst-case) lower bounds on the regret and Bayes risk, under specific assumptions on the set $\mathcal{D}$. For $\mathcal{D}$ the unit sphere in $\mathbb{R}^d$, [Rusmevichientong and Tsitsiklis, 2010] show that there exists a prior distribution π_0 (namely a Gaussian prior with covariance matrix $\frac{1}{d} I_d$) such that every bandit algorithm $\mathcal{A}$ satisfies $\mathrm{BR}_{\pi_0}(T, \mathcal{A}) \geq 0.006 d\sqrt{T}$. Consequently, for every algorithm $\mathcal{A}$, there exists $\theta \in \mathbb{R}^d$ such that in the associated linear bandit model $\mathrm{R}_\theta(T, \mathcal{A}) \geq 0.006 d\sqrt{T}$. When $\mathcal{D}$ is the hypercube $\{0,1\}^d$, of cardinal $K = 2^d$, [Dani et al., 2007] show that for any bandit algorithm $\mathcal{A}$, there exists a vector θ and a centered noise ϵ_t such that $\mathrm{R}_\theta(T, \mathcal{A})$ is at least of order $d\sqrt{T}$. In this example, $d\sqrt{T} = O(\sqrt{dT \log(K)})$, and more generally when the number of contexts is finite, one may expect to have an upper bound on the regret or Bayes risk that scales in $\sqrt{dT \log(K)}$ in place of $d\sqrt{T}$.

This has been shown to be possible in a more general, adversarial, setting. In an adversarial linear bandit problem (rather called bandit linear optimization in the literature), at each time t the agent chooses a context $x_t \in \mathcal{D}$ while an 'adversary' simultaneously chooses a vector $l_t \in \mathbb{R}^d$. The agent receives the rewards $x_t^T l_t$ and wants to minimize its pseudo-regret, defined in this setting by

$$\mathcal{R}(T, \mathcal{A}) = \max_{x \in \mathcal{D}} \mathbb{E} \left[\sum_{t=1}^{T} x^T l_t \right] - \mathbb{E} \left[\sum_{t=1}^{T} x_t^T l_t \right].$$

No stochastic assumptions are made on l_t unlike in the (stochastic) linear bandit problem considered so far. When the number of contexts in $\mathcal{D}$ is finite, algorithms for online linear optimization mostly consist in adapting the EXP3 algorithm of [Auer et al., 2002b] suited for classical adversarial bandits (see e.g. [Bubeck and Cesa-Bianchi, 2012]) by incorporating estimates of l_t at each step. Among them, EXP2 with John exploration, proposed by [Bubeck et al., 2012] can be applied to the stochastic linear contextual bandit problem with a static, finite set of context $\mathcal{D}$ and is such that $\mathrm{R}_\theta(T, \mathcal{A}) \leq 2\sqrt{3}\sqrt{dT \log(K)}$. However, this algorithm is difficult to implement. We will show below that the pseudo-regret of Bayes-UCB and Thompson Sampling also scales in $\tilde{O}(d\sqrt{T})$ or $\tilde{O}(\sqrt{dT \log(K)})$, whichever is smaller.

4.2 Bayesian and frequentist confidence regions

Confidence regions on θ in $\mathbb{R}^d$ but also confidence intervals (in $\mathbb{R}$) on the mean of each context $x^T \theta$ are necessary to define and analyse optimistic algorithms, like the algorithms mentioned above or the

two variants of Bayes-UCB presented in the next section. In Lemma 4.1, we recall a confidence region given by [Abbasi-Yadkori et al., 2011]. This confidence region was obtained using the so-called 'method of mixtures', that we discuss in Appendix A. One can find in Section A.3 therein the deviation inequality for vector-valued martingales proved by [Abbasi-Yadkori et al., 2011] that allows to show the following lemma.

Lemma 4.1. *Under Assumptions 1 and 2, for any algorithm $\mathcal{A}$,*

$$\mathbb{P}_\theta\left(\forall t\in\mathbb{N}, \|\theta-\hat{\theta}(t)\|_{\Sigma_t^{-1}} \leq \left(\sqrt{2\log\frac{1}{\delta}+d\log\left(1+(t+1)\frac{L^2\kappa^2}{d\sigma^2}\right)}+\frac{1}{\kappa}\|\theta\|\right)\right)\geq 1-\delta.$$

Proof. Introducing the martingale $S_t = X_t^T E_t = \sum_{s=1}^t \epsilon_s x_s$, and letting $V_t = B(t)$, one can write

$$\begin{aligned}\hat{\theta}(t)-\theta &= V_t^{-1}X_t^TY_t-\theta = V_t^{-1}X_t^TX_t\theta+V_t^{-1}X_t^TE_t-\theta = -\frac{\sigma^2}{\kappa^2}V_t^{-1}\theta+V_t^{-1}S_t\\ \|\hat{\theta}(t)-\theta\|_{\Sigma_t^{-1}} &= \sigma^{-1}\|\hat{\theta}(t)-\theta\|_{V_t}\leq\sigma^{-1}\|S_t\|_{V_t^{-1}}+\frac{\sigma}{\kappa^2}\|\theta\|_{V_t^{-1}}.\end{aligned}$$

$S_t = \sum_{s=1}^t \epsilon_s x_s$ is such that the x_s are $\mathcal{F}_{s-1}$ measurable and ϵ_s is centered and σ-subgaussian. Thus the deviation inequality of Lemma A.8 in Appendix A can be applied with $V = \frac{\sigma^2}{\kappa^2}\mathrm{I}_d$. With probability larger than $1-\delta$, for all $t\in\mathbb{N}$,

$$\|\hat{\theta}(t)-\theta\|_{\Sigma_t^{-1}}\leq\sqrt{2\log\frac{1}{\delta}+\log\frac{\det(V_t)}{\det\left(\frac{\sigma^2}{\kappa^2}\mathrm{I}_d\right)}}+\frac{1}{\kappa}\|\theta\|.$$

Following Lemma 10 of [Abbasi-Yadkori et al., 2011], the determinant of V_t is bounded as

$$\det(V_t)\leq\left(\frac{\sigma^2}{\kappa^2}+t\frac{L^2}{d}\right)^d, \tag{4.4}$$

which concludes the proof.

□

Lemma 4.1 also leads to a confidence intervals on $x^T\theta$ since from the Cauchy-Schwarz inequality

$$|\theta^Tx-\hat{\theta}(t)^Tx|\leq\|\theta-\hat{\theta}(t)\|_{\Sigma_t^{-1}}\|x\|_{\Sigma_t}.$$

This confidence region on θ given in Lemma 4.1 holds conditionally to θ and is therefore a 'frequentist' confidence region. It is used to define and analyse the OFUL algorithm ([Abbasi-Yadkori et al., 2011]) as well as in the frequentist analysis of a version of Thompson Sampling proposed by [Agrawal and Goyal, 2013b]. Integrating over the prior yields a Bayesian confidence region, that is used by [Russo and Van Roy, 2014] in a general Bayesian analysis of Thompson Sampling.

The Bayesian analysis we present here for Thompson Sampling and Bayes-UCB are based on the Bayesian confidence regions we introduce here in a model with Gaussian noise and Gaussian prior.

Lemma 4.2. *Under the Bayesian model (4.2), for any algorithm $\mathcal{A}$,*

$$\mathbb{P}\left(\|\theta - \hat{\theta}(t)\|_{\Sigma_t^{-1}} \leq \sqrt{Q\left(1-\delta;\chi^2_d\right)}\right) \geq 1-\delta. \tag{4.5}$$

Moreover, if at each round $|\mathcal{D}_t| = K$, for any algorithm $\mathcal{A}$, for all $t \geq 1$,

$$\mathbb{P}\left(\forall x \in \mathcal{D}_{t+1}, |x^T\theta - x^T\hat{\theta}(t)| \leq \|x\|_{\Sigma_t} Q\left(1-\frac{\delta}{2K};\mathcal{N}(0,1)\right)\right) \geq 1-\delta. \tag{4.6}$$

where $Q(\alpha,\pi)$ is the quantile of order α of the distribution π.

Proof. We first write

$$\begin{aligned}\|\theta - \hat{\theta}(t)\|^2_{\Sigma_t^{-1}} &= (\theta-\hat{\theta}(t))^T\Sigma_t^{-1}(\theta-\hat{\theta}(t)) = [(\theta-\hat{\theta}(t))^T\Sigma_t^{-\frac{1}{2}}]\Sigma_t^{-\frac{1}{2}}(\theta-\hat{\theta}(t)) \\ &= \|\Sigma_t^{-\frac{1}{2}}(\theta-\hat{\theta}(t))\|^2.\end{aligned}$$

Given $\mathcal{H}_t$, θ has distribution $\mathcal{N}\left(\hat{\theta}(t),\Sigma_t\right)$, hence $\Sigma_t^{-\frac{1}{2}}(\theta-\hat{\theta}(t)) \sim \mathcal{N}(0,I_d)$ and $\|\Sigma_t^{-\frac{1}{2}}(\theta-\hat{\theta}(t))\|^2$ follows a chi-square distribution with d degrees of freedom. We have shown that

$$\|\theta - \hat{\theta}(t)\|^2_{\Sigma_t^{-1}}|\mathcal{H}_t \sim \chi^2_d.$$

Hence, we have, by definition of the quantile, that

$$\mathbb{P}\left(\|\theta - \hat{\theta}(t)\|_{\Sigma_t^{-1}} \leq \sqrt{Q\left(1-\delta;\chi^2_d\right)}\middle|\mathcal{H}_t\right) = \mathbb{P}\left(\|\theta - \hat{\theta}(t)\|^2_{\Sigma_t^{-1}} \leq Q\left(1-\delta;\chi^2_d\right)\middle|\mathcal{H}_t\right) \geq 1-\delta$$

and inequality (4.5) follows in conditioning.

In the particular case where $\mathcal{D}_t = K$ for all $t \in \mathbb{N}$, we can introduce an (arbitrary) ordering of the contexts and write $\mathcal{D}_t = (b_1(t),\dots,b_K(t))$. Then, for each context, we use that conditionally to $\mathcal{H}_t$, $b_i(t+1)^T\theta \sim \mathcal{N}\left(b_i(t+1)^T\hat{\theta}(t), \|b_i(t+1)\|_{\Sigma_t}\right)$. Thus

$$\begin{aligned}&\mathbb{P}\left(\exists x \in \mathcal{D}_{t+1}, |x^T\theta - x^T\hat{\theta}(t)| > \|x\|_{\Sigma_t} Q\left(1-\frac{\delta}{2K};\mathcal{N}(0,1)\right)\right) \\ &\quad\leq \sum_{i=1}^K \mathbb{P}\left(|b_i(t+1)^T\theta - b_i(t+1)^T\hat{\theta}(t)| > \|b_i(t+1)\|_{\Sigma_t} Q\left(1-\frac{\delta}{2K};\mathcal{N}(0,1)\right)\right) \\ &\quad= \sum_{i=1}^K \mathbb{E}\left[\mathbb{P}\left(|b_i(t+1)^T\theta - b_i(t+1)^T\hat{\theta}(t)| > \|b_i(t+1)\|_{\Sigma_t}\, Q\left(1-\frac{\delta}{2K};\mathcal{N}(0,1)\right)\middle|\mathcal{H}_t\right)\right] \\ &\quad= \sum_{i=1}^K \frac{\delta}{K} = \delta\end{aligned}$$

which yields inequality (4.6).

□

4.3 The Bayes-UCB algorithm and a generalization

4.3.1 The algorithms

The Bayes-UCB algorithm presented in Chapter 2 picks at time t the arm whose quantile of order $1 - \frac{1}{t}$ of the marginal posterior distribution on the mean is maximal. Ignoring dependencies among the arms, we explained in Chapter 2 how this algorithm could be applied in a (static) linear bandit model. Bayes-UCB can also naturally be extended to a linear contextual bandit (with changing contexts). To obtain theoretical guarantees, however, we need to consider a different level of confidence, and $1/t$ is replaced in the definition below by $e^{-f(t,\delta)}$, for some exploration rate $f(t,\delta)$. The Bayes-UCB algorithm using the Bayesian model (4.2) chooses at time $t+1$

$$\begin{aligned} x_{t+1} &= \underset{x\in\mathcal{D}_{t+1}}{\text{argmax}}\, Q\left(1-e^{-f(t+1,\delta)}; \mathcal{N}\left(x^T\hat{\theta}(t), \|x\|_{\Sigma_t}\right)\right), \\ x_{t+1} &= \underset{x\in\mathcal{D}_{t+1}}{\text{argmax}}\, \left[x^T\hat{\theta}(t) + \|x\|_{\Sigma_t} Q\left(1-e^{-f(t+1,\delta)}; \mathcal{N}(0,1)\right)\right], \end{aligned}$$

where $f(t,\delta)$ is some exploration rate. Indeed, the posterior distribution on the mean $x^T\theta$ of arm x at the beginning of round $t+1$ is $\mathcal{N}\left(x^T\hat{\theta}(t), \|x\|_{\Sigma_t}\right)$.

Bayes-UCB can be seen as a variant of a UCB-type algorithm that uses Bayesian confidence regions on the mean of each arm. When we go from classical bandit to linear bandits, the optimism-in-face-of-uncertainty principle is used differently : one rather picks the context for which there exists a 'possible' regression parameter θ (in the sense that it lies in some confidence region) for which the associated mean $x^T\theta$ is the highest among all contexts and all 'possible' regression parameters (see (4.3)). A natural Bayesian optimistic algorithm for the linear contextual bandit problem consists therefore in applying this principle with a Bayesian confidence region. It follows from inequality (4.5) that

$$\mathbb{P}\left(\|\theta - \hat{\theta}(t)\|_{\Sigma_t^{-1}} \leq \sqrt{Q\left(1-e^{-f(t+1,\delta)}; \chi^2_d\right)}\right) \geq 1 - e^{-f(t+1,\delta)}.$$

We define the Bayes-LinUCB algorithm to be the algorithm of the form (4.3) using the Bayesian confidence region

$$C_t = \left\{\theta' : \|\theta' - \hat{\theta}(t)\|_{\Sigma_t^{-1}} \leq \sqrt{Q\left(1-e^{-f(t+1,\delta)}; \chi^2_d\right)}\right\}.$$

As explained above, using a bit of algebra one can show Bayes-LinUCB picks at time $t+1$ the context

$$x_{t+1} = \underset{x\in\mathcal{D}_{t+1}}{\text{argmax}}\, \left[\hat{\theta}(t)^T x + \|x\|_{\Sigma_t}\sqrt{Q\left(1-e^{-f(t+1,\delta)}; \chi^2_d\right)}\right].$$

4.3.2 Bayesian analysis of Bayes-UCB and Bayes-LinUCB

Bayes-UCB and Bayes-LinUCB are algorithms of the form

$$x_{t+1} = \underset{x\in\mathcal{D}_{t+1}}{\text{argmax}}\, \left[\hat{\theta}(t)^T x + \|x\|_{\Sigma_t}\beta(t+1,\delta)\right]. \tag{4.7}$$

One of the first analyses proposed for an algorithm of this form was that of ConfidenceBall given by [Dani et al., 2008]. A closer examination of their analysis shows that its cornerstone can be summarized in Lemma 4.3 below, whose proof is given in Section 4.6.1. This Lemma is also implicitly used in the analysis of the Uncertainty Ellipsoid algorithm of [Rusmevichientong and Tsitsiklis, 2010] and of the OFUL algorithm of [Abbasi-Yadkori et al., 2011].

Lemma 4.3. *For an algorithm $\mathcal{A}$ that picks at time $t+1$ the context x_{t+1} according to (4.7), on the event*

$$E = \bigcap_{t\in\mathbb{N}} \left(\forall x \in \mathcal{D}_{t+1}, |x^T\theta - x^T\hat{\theta}(t)| \leq \|x\|_{\Sigma_t}\beta(t+1,\delta)\right)$$

the pseudo-regret of $\mathcal{A}$ satisfies, under Assumption 1,

$$\forall T \in \mathbb{N}^*,\ \mathcal{R}_\theta(\mathcal{A},T) \leq \beta(T,\delta)\sqrt{Td}\sqrt{C_1 \log\left(1+T\frac{L^2\kappa^2}{\sigma^2 d}\right)},$$

with $C_1 := \frac{4L^2\kappa^2}{\log(1+L^2\kappa^2\sigma^{-2})}$.

This result states that on some event E, the pseudo-regret $\mathcal{R}_\theta(T,\mathcal{A})$ is upper bounded by a deterministic quantity. For the OFUL algorithm, Lemma 4.1 shows that the event E in Lemma 4.3 holds with probability at least $1-\delta$, leading to a high-probability upper bound on the pseudo-regret in the frequentist framework. Upper bounding the pseudo-regret of Bayes-UCB or Bayes-LinUCB with high probability, then boils down to choosing the exploration rate $f(t,\delta)$ in such a way that the probability of the event E associated is larger than $1-\delta$. Using the Bayesian confidence regions given in Lemma 4.2, and upper bounds on the quantiles of the normal and chi-square distribution leads to PAC-Bayesian bounds on the pseudo-regret stated in Theorem 4.4.

Up to logarithmic factor in T, the pseudo-regret of Bayes-UCB when there is a finite number of contexts K is of order $\tilde{O}(\sqrt{dT\log(K)})$, in high probability under the Bayesian model (4.2). For Bayes-LinUCB, suited for the general case (with a potentially infinite set of context $\mathcal{D}_t$), the pseudo-regret is of order $\tilde{O}(d\sqrt{T})$, also with high probability.

Theorem 4.4. *Choosing $f(t,\delta) = \log\frac{K\pi^2t^2}{3\delta}$, when the number of context is finite and $|\mathcal{D}_t| \leq K$, the Bayes-UCB algorithm satisfies, under Assumption 1 and the Bayesian model (4.2)*

$$\mathbb{P}\left(\forall T\in\mathbb{N}, \mathcal{R}_\theta(T,\mathcal{A}) \leq \sqrt{Td}\sqrt{2C_1\log\left(\frac{K\pi^2T^2}{6\delta}\right)\log\left(1+T\frac{L^2\sigma^2}{d\kappa^2}\right)}\right) \geq 1-\delta.$$

With the exploration rate $f(t,\delta) = \log\frac{\pi^2t^2}{6\delta}$, the Bayes-LinUCB algorithm satisfies

$$\mathbb{P}\left(\forall T\in\mathbb{N}, \mathcal{R}_\theta(T,\mathcal{A}) \leq d\sqrt{T}\sqrt{C_1\log\left(1+T\frac{L^2\sigma^2}{d\kappa^2}\right)\left(1+\frac{2}{d}\log\frac{\pi^2T^2}{6\delta}+2\sqrt{\frac{1}{d}\log\frac{\pi^2T^2}{6\delta}}\right)}\right) \geq 1-\delta.$$

where C_1 is the constant introduced in Lemma 4.3.

Proof of Theorem 4.4 Assume that $\forall t\in\mathbb{N}, |\mathcal{D}_t| = K$. With $f(t,\delta) = \log\frac{K\pi^2t^2}{3\delta}$, Bayes-UCB picks at time $t+1$ the context x_{t+1} according to the Equation (4.7) with $\beta(t,\delta) = Q\left(1-\frac{3\delta}{K\pi^2t^2};\mathcal{N}(0,1)\right)$. Let E be the event

$$E = \bigcap_{t\in\mathbb{N}}\left(\forall x\in\mathcal{D}_{t+1}, |x^T\theta - x^T\hat{\theta}(t)| \leq \|x\|_{\Sigma_t}Q\left(1-\frac{3\delta}{K\pi^2(t+1)^2};\mathcal{N}(0,1)\right)\right).$$

Inequality (4.6) yields, for all $t\in\mathbb{N}^*$

$$\mathbb{P}\left(\forall x\in\mathcal{D}_t, |x^T\theta - x^T\hat{\theta}(t-1)| \leq \|x\|_{\Sigma_{t-1}}Q\left(1-\frac{6\delta}{2K\pi^2t^2};\mathcal{N}(0,1)\right)\right) \geq 1-\frac{6\delta}{\pi^2t^2}$$

and an union bound on $t \geq 1$ yields $\mathbb{P}(E) \geq 1-\delta$. On E, from Lemma 4.3 the pseudo-regret is upper bounded for all T as follows:

$$\begin{aligned}\mathcal{R}_\theta(T,\mathcal{A}) &\leq Q\left(1-\frac{3\delta}{K\pi^2T^2};\mathcal{N}(0,1)\right)\sqrt{Td}\sqrt{C_1\log\left(1+T\frac{L^2\kappa^2}{\sigma^2 d}\right)}\\ &\leq \sqrt{Td}\sqrt{2\log\left(\frac{K\pi^2T^2}{6\delta}\right)}\sqrt{C_1\log\left(1+T\frac{L^2\kappa^2}{\sigma^2 d}\right)},\end{aligned}$$

using a classic upper bound on the tail of a Gaussian distribution: if $X \sim \mathcal{N}(0,1)$ it can be shown that $\mathbb{P}(X>c) \leq (1/2)e^{-c^2/2}$.

Bayes-LinUCB using the exploration rate stated above picks at time $t+1$ the context x_{t+1} according to (4.7) with $\beta(t,\delta) = \sqrt{Q\left(1-\frac{6\delta}{\pi^2t^2};\chi^2_d\right)}$. Introducing

$$E = \bigcap_{t\in\mathbb{N}}\left(\forall x \in \mathcal{D}_{t+1}, |x^T\theta - x^T\hat{\theta}(t)| \leq \|x\|_{\Sigma_t}\sqrt{Q\left(1-\frac{6\delta}{\pi^2t^2};\chi^2_d\right)}\right),$$

inequality (4.5) and a union bounds also yields $\mathbb{P}(E) \geq 1-\delta$. By Lemma 4.3, on E one has

$$\forall T \in \mathbb{N}^*,\ \mathcal{R}_\theta(T,\mathcal{A}) \leq \sqrt{Q\left(1-\frac{6\delta}{\pi^2T^2};\chi^2_d\right)}\sqrt{Td}\sqrt{C_1\log\left(1+T\frac{L^2\kappa^2}{\sigma^2 d}\right)}.$$

Inequality (4.3) in [Laurent and Massart, 2000] provides an upper bound of the quantile of a chi-square distribution, namely

$$Q\left(1-\alpha;\chi^2_d\right) \leq d + 2\log\frac{1}{\alpha} + 2\sqrt{d\log\frac{1}{\alpha}}. \tag{4.8}$$

Using this inequality, on event E,

$$\mathcal{R}_\theta(T,\mathcal{A}) \leq \sqrt{d+\log\left(\frac{\pi^2T^2}{6\delta}\right)+2\sqrt{d\log\left(\frac{\pi^2t^2}{6\delta}\right)}}\sqrt{Td}\sqrt{C_1\log\left(1+T\frac{L^2\kappa^2}{\sigma^2 d}\right)}$$

which concludes the proof.

□

4.3.3 Comparison with other optimistic algorithms

A linear bandit model with Gaussian prior can also be seen as the simplest example of Gaussian process bandit models as presented by [Srinivas et al., 2010, Srinivas et al., 2012]. Indeed, the model $y_t = x_t^T\theta + \epsilon_t$ with the assumption that $\theta \sim \mathcal{N}\left(0,\kappa^2 I_d\right)$ is equivalent to the model $y_t = f(x_t) + \epsilon_t$ with the assumption that f is sampled from a Gaussian process $\sim GP(0,k(x,x'))$ where k is a linear kernel on $\mathcal{D}$ such that $k(x,x') = \kappa^2 x^T x'$. [Srinivas et al., 2010] also make the assumption that the noise is Gaussian with variance σ^2 and introduce the GP-UCB algorithm, that chooses at time $t+1$, when $|\mathcal{D}_{t+1}| = K$,

$$x_{t+1} = \underset{x\in\mathcal{D}_{t+1}}{\operatorname{argmax}}\left[\hat{\theta}(t) + \|x\|_{\Sigma_t}\sqrt{2\log\left(\frac{K(t+1)^2\pi^2}{6\delta}\right)}\right].$$

Algorithm	Value of $\beta(t,\delta)$
Bayes-UCB*	$Q\left(1-\frac{3\delta}{K\pi^2t^2};\mathcal{N}(0,1)\right)$
Bayes-LinUCB	$\sqrt{Q\left(1-\frac{6\delta}{\pi^2t^2};\chi^2_d\right)}$
GP-UCB*	$\sqrt{2\log\left(\frac{Kt^2\pi^2}{6\delta}\right)}$
OFUL	$\left(\sqrt{2\log\frac{1}{\delta}+d\log\left(1+t\frac{L^2\kappa^2}{d\sigma^2}\right)}+\frac{1}{\kappa}\|\theta\|\right)$

Table 4.1: Index for various policies, some of them* suited for $|\mathcal{D}_t| = K$.

While the original version of GP-UCB was suited for a static context set $\mathcal{D}$, the authors later proposed a contextualized version ([Krause and Ong, 2011]).

With this particular linear kernel, GP-UCB is very close to Bayes-UCB: more precisely, since $\sqrt{2\log\left(K\pi^2t^2/(6\delta)\right)}$ is a tight upper bound on the quantile $Q\left(1-3\delta/(K\pi^2t^2);\mathcal{N}(0,1)\right)$ used by the provably efficient version of Bayes-UCB given in Theorem 4.4. [Srinivas et al., 2010] also propose an algorithm in the case when $\mathcal{D}$ is a compact set, based on a discretization argument and on regularity properties of the Gaussian process. Bayes-LinUCB on the contrary can be implemented for any set $\mathcal{D}$ and is based on the simple Bayesian confidence region given by (4.5).

Optimistic algorithms for linear contextual bandits presented so far appear as generalized index policies, of the form

$$x_{t+1} = \underset{x\in\mathcal{D}_{t+1}}{\operatorname{argmax}} \left[\hat{\theta}(t)^T x + \|x\|_{\Sigma_t}\beta(t+1,\delta)\right],$$

with different exploration rates $\beta(t,\delta)$, that we recall in Table 4.1. From the analysis we gave, based on Lemma 4.3, the smaller $\beta(t,\delta)$, the smaller the upper bound on the regret obtained. However, the way the different values of $\beta(t,\delta)$ in Table 4.1 compare is not obvious. As just mentioned, the exploration rate of Bayes-UCB is always smaller than that of GP-UCB. Among the other three algorithms, one would expect the exploration of Bayes-UCB to be smaller than that of Bayes-LinUCB on bandit models with finite number of contexts such that $\log(K) \leq d$. Figure 4.1 illustrates this tendency: for a bandit model with $\log(K) \simeq 7$, we show that the exploration rate of Bayes-UCB is smaller than that of Bayes-LinUCB when $d = 5$, and larger when $d = 10$. In these cases, both exploration rates are smaller than that of OFUL.

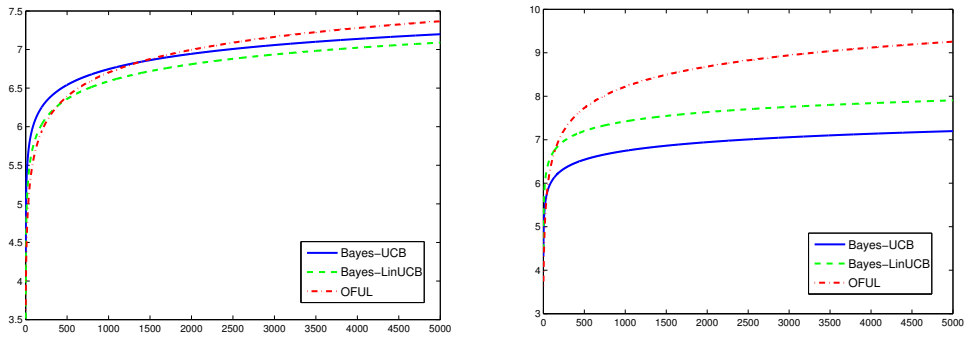

Figure 4.1: Exploration rates $\beta(t,\delta)$ used by the different algorithms as a function of t for $\delta = 0.05$, on a bandit model for which $K = 2000$ and $d = 5$ (left) and $K = 2000$ and $d = 10$ (right)

The original Bayes-UCB algorithm (see Chapter 2) is proved optimal for classical bandits with

Bernoulli rewards in a frequentist sense: the algorithm matches the problem-dependent lower bound on the regret given by [Lai and Robbins, 1985] for every Bernoulli bandit model. In the same spirit, we could expect to be able to give a frequentist analysis of Bayes-UCB (or Bayes-LinUCB) in the more general linear contextual bandit problem studied here. As already explained, most of analyses of algorithms of the form (4.7) rely on Lemma 4.3. For example, with a fixed number of contexts K, frequentist guarantees for Bayes-UCB would rely on an upper bound for the probability

$$\mathbb{P}_\theta\left(\exists t \in \mathbb{N}^*, \exists x \in \mathcal{D}_{t+1}, |x^T\theta - x^T\hat{\theta}(t)| > \|x\|_{\Sigma_t} Q\left(1 - \frac{\pi^2\delta}{6Kt^2}; \mathcal{N}(0,1)\right)\right).$$

However, the best deviation inequality currently available in the frequentist setting is the one deduced from Lemma 4.1:

$$\mathbb{P}_\theta\left(\exists t \in \mathbb{N}, \exists x \in \mathcal{D}_{t+1} : |\theta^T x - \hat{\theta}(t)^T x| > \|x\|_{\Sigma_t}\left(\sqrt{2\log\frac{1}{\delta} + d\log\left(1 + (t+1)\frac{L^2\kappa^2}{d\sigma^2}\right)} + \frac{1}{\kappa}\|\theta\|\right)\right) \leq \delta.$$

Even for a fixed context x, it is not known whether one can prove a deviation inequality of the form

$$\mathbb{P}_\theta\left(\exists t \in \mathbb{N}^* \ |\theta^T x - \hat{\theta}(t)^T x| > \|x\|_{\Sigma_t}\beta(t+1,\delta)\right) \leq \delta \tag{4.9}$$

with an exploration rate $\beta(t,\delta)$ independent on the dimension d.

4.4 Thompson Sampling

4.4.1 The algorithm

Thompson Sampling in linear contextual bandit models is easy to implement. Given a Bayesian model such that samples from the posterior distributions on θ at each round can be computed, the algorithm draws at round $t+1$ a sample $\tilde{\theta}(t)$ from the posterior distribution at the end of round t and picks the context x_{t+1} according to

$$x_{t+1} = \text{argmax}_{x\in\mathcal{D}_{t+1}} \ x^T\tilde{\theta}(t).$$

This ensures that the context chosen has the highest posterior probability of being the best context. Whereas optimistic algorithms reduce to solving the optimization problem (4.7), Thompson Sampling only requires to solve a linear optimization problem, which is in general easier.

Recently, Bayesian and frequentist guarantees for Thompson Sampling in linear contextual bandits have emerged in the literature. On the one hand [Russo and Van Roy, 2014] propose a general Bayesian analysis of Thompson Sampling that can be applied to linear contextual bandits when the prior distribution is such that θ remains bounded. On the other hand, [Agrawal and Goyal, 2013b] propose the first frequentist analysis of Thompson Sampling with a Gaussian prior. They give a high-probability upper bound on the pseudo-regret of Thompson Sampling for a specific choice of prior (and likelihood), that depends on δ and on assumptions on the noise.

As in the rest of the chapter, we focus here on Thompson Sampling with Gaussian prior. We review existing results from [Russo and Van Roy, 2014] and [Agrawal and Goyal, 2013b] and additionally propose a new Bayesian analysis suited for a potentially infinite set of contexts $\mathcal{D}_t$, a case not completely covered by the work of [Russo and Van Roy, 2014].

4.4.2 A Bayesian analysis of Thompson Sampling

[Russo and Van Roy, 2014] propose a Bayesian analysis of Thompson Sampling for a wide class of models, including linear contextual bandits. Their analysis holds for quite general prior distributions π_0 and uses the fact that conditionally to the history $\mathcal{H}_t$, θ and the sample $\tilde{\theta}(t)$ used in the algorithm have the same distribution, and so do x_{t+1} and x^*_{t+1}. This remark leads to a Bayes risk decomposition extensively used by [Russo and Van Roy, 2014], that rewrites for linear bandits

$$\mathrm{BR}_{\pi_0}(T, \mathrm{TS}_{\pi_0}) = \mathbb{E}[\mathcal{R}_\theta(T, \mathrm{TS}_{\pi_0})] \leq \mathbb{E}\sum_{t=1}^{T}[U_{t-1}(x_t) - \theta^T x_t] + \mathbb{E}\sum_{t=1}^{T}[\theta^T x^*_t - U_{t-1}(x^*_t)], \tag{4.10}$$

where (U_t) is *any* sequence of confidence bounds. The Bayes-risk bound sketched in [Russo and Van Roy, 2014] for linear contextual bandit uses classic upper bounds of the form $U_t(x) = \hat{\theta}(t)^T x + \|x\|_{\Sigma_t}\beta_{t+1}$ and holds for prior distribution satisfying $\mathbb{P}(\|\theta\|_2 \leq C) = 1$ for some constant C.

This assumption does not hold for a Gaussian prior distribution. However, as already pointed out in Section 4.3.3, a linear bandit model with Gaussian prior and Gaussian noise is equivalent to a particular case of Gaussian process bandit, for which [Russo and Van Roy, 2014] also provide a Bayes risk bound. This bound is based on the decomposition (4.10) and holds when the total number of contexts $\mathcal{D}_t$ is finite and static. Their argument can be easily extended to a set of changing contexts, leading to the following theorem.

Theorem 4.5 (adapted from [Russo and Van Roy, 2014], Proposition 5)**.** *In the Bayesian model (4.2), if for all t, $|\mathcal{D}_t| = K$, under Assumption 1, the Bayes-risk of Thompson sampling using a Gaussian prior $\pi_0 = \mathcal{N}\left(0, \kappa^2 I_d\right)$ is upper bounded as*

$$\mathrm{BR}_{\pi_0}(T, \mathrm{TS}_{\pi_0}) \leq \frac{\kappa^2 L^2}{\sqrt{2\pi}} + \sqrt{2dT\log\left(\frac{KT^2\pi^2}{6}\right)}\sqrt{C_1\log\left(1 + T\frac{L^2\kappa^2}{\sigma^2 d}\right)},$$

with C_1 defined as in Lemma 4.3.

The Bayes risk bound we now provide in Theorem 4.6 holds without the assumption that the context set $\mathcal{D}_t$ is finite. It is based on techniques similar to those of [Russo and Van Roy, 2014] (even if the decomposition (4.10) is not used explicitly) and on the fact that conditionally to $\mathcal{H}_t$, $\|\theta - \hat{\theta}(t)\|_{\Sigma_t^{-1}}$ follows a chi-square distribution with d degrees of freedom, as already used to obtain the Bayesian confidence region (4.5) in Lemma 4.2. The proof of this theorem is given in Section 4.6.2.

Theorem 4.6. *Under Assumption 1 and the Bayesian model (4.2), the Bayes risk of Thompson Sampling using a Gaussian prior $\pi_0 = \mathcal{N}\left(0, \kappa^2 I_d\right)$ is upper bounded in the following way:*

$$\mathrm{BR}_{\pi_0}(T, \mathrm{TS}_{\pi_0}) \leq d\sqrt{T}\sqrt{C_1\log\left(1 + T\frac{L^2\kappa^2}{\sigma^2 d}\right)\left(3 + 8\log T + 2\sqrt{\log(2T^4)}\right)} + \frac{\pi^2 L\kappa\sqrt{d}}{3\sigma},$$

with $C_1 = \frac{4L^2\kappa^2}{\log(1+L^2\kappa^2\sigma^{-2})}$.

In a linear contextual bandit problem with a finite number of contexts presented at each round, we saw in the previous section that either Bayes-UCB or Bayes-LinUCB could be used, with a high probability upper bound on the pseudo-regret of order $\tilde{O}(\sqrt{dT\log(K)})$ or $\tilde{O}(d\sqrt{T})$ respectively. Thus, depending on the relation between the dimension d an the number of contexts K, one of these two variants could

be preferred based on these theoretical bounds or on the observation that one version performs better than the other in practice. Conversely, for Thompson Sampling, no adaptation of the algorithm is needed to obtain the same guarantees. That is, the Bayes risk of Thompson Sampling is upper bounded by whichever is smaller between the two bounds given in Theorem 4.5 and Theorem 4.6.

4.4.3 A frequentist analysis of Thompson Sampling

The only frequentist guarantees available to date for Thompson Sampling in contextual bandits with linear payoff follow from the work of [Agrawal and Goyal, 2013b]. For a fixed value of θ, under Assumptions 1 and 2, the authors give a high-probability upper bound on the pseudo-regret of Thompson Sampling, but with a specific choice of prior *and* likelihood.

If the noise is assumed to be σ^2-subgaussian (cf. Assumption 2), letting

$$\kappa = v = \sigma\sqrt{9d\log\frac{T}{\delta}}, \tag{4.11}$$

[Agrawal and Goyal, 2013b] analyse the version of Thompson Sampling assuming $\theta \sim \mathcal{N}\left(0, \kappa^2\right)$ and $\epsilon_t \sim \mathcal{N}\left(0, v^2\right)$. This algorithm, denoted by TS$'$ chooses at round $t+1$ the context maximizing the dot product with the sample

$$\tilde{\theta}(t) \sim \mathcal{N}\left(\hat{\theta}(t), v^2 B(t)^{-1}\right) \text{ with } B(t) = \mathrm{I}_d + X_t^T X_t.$$

The parameters κ and v have to be chosen as a function of the horizon T. To circumvent this issue, the authors propose and analyse the algorithm drawing at each time a sample from a normal distribution with mean $\hat{\theta}(t)$ and covariance $v_t^2 B(t)^{-1}$, with $v_t = \sigma\sqrt{9d\log\frac{t}{\delta}}$. This distribution can be regarded as a posterior distribution if the prior distribution and the distribution of the noise are $\mathcal{N}\left(0, v_t^2\right)$. Due to this varying prior distribution, this horizon-free variant cannot really be interpreted as Thompson Sampling. For TS$'$ and its horizon-free variant, [Agrawal and Goyal, 2013b] prove the following theorem.

Theorem 4.7 ([Agrawal and Goyal, 2013b], Theorem 1)**.** *For a fixed value of θ, under Assumptions 1 and 2, the algorithm TS$'$ described above satisfies, with probability larger than $1-\delta$,*

$$\mathcal{R}_\theta(T, \mathrm{TS}') = O\left(\min\left\{d^{3/2}\sqrt{T}; d\sqrt{T\log(K)}\right\}\left(\log(T) + \sqrt{\log(T)\log\frac{1}{\delta}}\right)\right).$$

The proof of Theorem 4.7 given by [Agrawal and Goyal, 2013b] is interesting since it departs from frequentist analyses presented so far. Indeed, it does not reduce to a deterministic upper bound of the pseudo-regret on some event whose probability can be shown to be larger than $1-\delta$ (as is the case when using Lemma 4.3). The cornerstone of this new analysis lies in an upper bound on the conditional expectation $\mathbb{E}_\theta\left[r_t|\mathcal{F}_{t-1}\right]$ which takes the form

$$\mathbb{E}_\theta\left[r_t|\mathcal{F}_{t-1}\right] \leq C g_t \mathbb{E}_\theta\left[||x_t||_{B(t-1)^{-1}}|\mathcal{F}_{t-1}\right] + \frac{D g_t}{t^2},$$

C, D being two real constants and g_t some function growing logarithmically.

Theorem 4.7 leaves interesting open questions regarding the analysis of Thompson Sampling from a frequentist perspective. First, this bound displays an extra factor $\sqrt{d}$ when compared to the bounds obtained in the Bayesian framework, whereas we will see in the experimental section to follow that in practice, the regret of Thompson Sampling is smaller. Second, if the noise is assumed to be σ^2-subgaussian, it is natural to try to obtain performance guarantees for the version of Thompson Sampling that assumes a normal distribution for the noise with variance σ^2, and not v^2 given by (4.11)).

4.5 Numerical experiments

Experiments assessing the performance of Thompson Sampling beyond the Bernoulli case are mostly carried out for a particular instance of *generalized* linear bandit model, based on logistic regression (see [Scott, 2010, Chapelle and Li, 2011, May et al., 2012, Chapelle et al., 2014]). The rewards are assumed to be binary, such that $y_t \in \{-1, +1\}$, modeling for example click or no click from the users, and

$$\mathbb{P}(y_t = 1 | x_t, \theta) = \frac{1}{1 + \exp(-\theta^T x_t)}.$$

If θ is assumed to be drawn from a Gaussian prior distribution $\mathcal{N}\left(0, \frac{1}{\lambda} \mathrm{I}_d\right)$, under this model the posterior distribution is no longer exactly computable. However, using a Laplace approximation (see e.g. [Bishop, 2006]), it can be approximated by a Gaussian distribution:

$$p(\theta | x_1, y_1, \ldots, x_t, y_t) \sim \mathcal{N}\left(m, \mathrm{Diag}(q_i^{-1})\right)$$

with

$$\begin{aligned} m &= \operatorname*{argmin}_{w \in \mathbb{R}^d} \frac{\lambda}{2} \|w\|^2 + \sum_{i=1}^{n} \log(1 + \exp(-y_i w^T x_i)) \\ q_i &= \sum_{j=1}^{n} x_{j,i}^2 p_j (1 - p_j) \quad \text{with} \quad p_j = (1 + \exp(-m^T x_j))^{-1}. \end{aligned}$$

Thompson Sampling in this model then boils down to drawing at each round a sample $\tilde{\theta}(t)$ from the (approximate) posterior distribution at the end of round t and then choosing the context

$$x_{t+1} = \operatorname*{argmax}_{x \in \mathcal{D}_{t+1}} \frac{1}{1 + \exp(-\tilde{\theta}(t)^T x)} = \operatorname*{argmax}_{x \in \mathcal{D}_{t+1}} \tilde{\theta}(t)^T x.$$

Hence this version of Thompson Sampling is quite similar to Thompson Sampling as presented for linear contextual bandits, but with a different posterior distribution motivated by the logistic model. In the experiments below, we focus on the linear model.

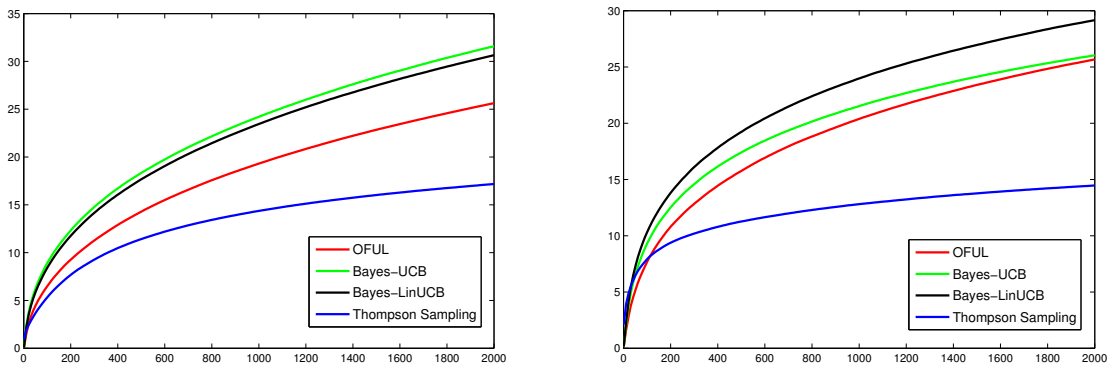

Figure 4.2: Regret curves for several algorithms in two different linear contextual models, in which $d = 5$ and $K = 2000$ (left) or $d = 10$ and $K = 2000$ (right).

Even if most of the results of this chapter are obtained under the Bayesian model (4.2), we propose here frequentist experiments. Indeed, for a fixed value of the regression parameter $\theta \in \mathbb{R}^d$, we repeat several bandit games up to some horizon T and estimate the regret (i.e. the expectation of the pseudo-regret, given this fixed value of θ). In each bandit game, the reward after choosing context x is drawn from $\mathcal{N}\left(x^T\theta, \sigma^2\right)$, with $\sigma = 0.25$. In order to simulate the changing contexts, we first choose (at random) a large set of $K_m = 100000$ contexts, and at each round t, a subset of $K = 2000$ contexts is chosen at random from this large set and forms the set $\mathcal{D}_t$. For this value of K, one has $\log(K) \simeq 7.6$. We propose experiments in two different bandit models in which the dimension d is either smaller or larger than this value. Results are reported in Figure 4.2.

In the first model, $d = 5$ (therefore $d < \log(K)$). Regret curves (averaged over $N = 1000$ bandit games) are displayed on the left of Figure 4.2 for OFUL and the three Bayesian strategies studied in this chapter: Bayes-UCB, Bayes-LinUCB and Thompson Sampling. In this bandit model, Bayes-LinUCB outperforms Bayes-UCB, as was expected from Figure 4.1, but OFUL outperforms both algorithms (maybe because for small horizons, the exploration rate used by OFUL is smaller than that used by both algorithms, as displayed in Figure 4.1). The regret of Thompson Sampling is significantly smaller than that of all other algorithms. In the second model, for which regret curves, averaged over $N = 500$ bandit games, are displayed on the right of Figure 4.2, one has $d = 10$ (therefore $d > \log(K)$). In this model, Bayes-UCB outperforms Bayes-LinUCB and performs almost as well as OFUL, whereas Thompson Sampling still outperforms all his competitors by a large margin.

4.6 Elements of proof

4.6.1 Proof of Lemma 4.3

Let $U_t(x) := \hat{\theta}(t)^T x + ||x||_{\Sigma_t^{-1}}\beta(t+1,\delta)$ and $\mathcal{A}$ be the algorithm picking at round t

$$x_t = \underset{x\in\mathcal{D}_t}{\text{argmax}}\ U_{t-1}(x).$$

We start by upper bounding the quantity $r_t = \theta^T x_t^* - \theta^T x_t$ on the event E. By definition, the algorithm is such that $U_{t-1}(x_t) \geq U_{t-1}(x)$ for all $x \in \mathcal{D}_t$, thus one has in particular $U_{t-1}(x_t^*) \leq U_{t-1}(x_t)$. On the event E, one also has $\theta^T x_t^* \leq U_{t-1}(x_t^*)$. Therefore, if E holds,

$$\begin{aligned} r_t &\leq U_{t-1}(x_t^*) - \theta^T x_t \leq U_{t-1}(x_t) - \theta^T x_t \\ &= \hat{\theta}(t-1)^T x_t + ||x_t||_{\Sigma_{t-1}}\beta(t,\delta) - \theta^T x_t \\ &= (\hat{\theta}(t-1) - \theta)^T x_t + ||x_t||_{\Sigma_{t-1}}\beta(t,\delta) \\ &\leq 2\beta(t,\delta)||x_t||_{\Sigma_{t-1}}. \end{aligned}$$

To obtain the last inequality, we use that on E, $(\hat{\theta}(t-1) - \theta)^T x_t \leq ||x_t||_{\Sigma_{t-1}}\beta(t,\delta)$. Technical lemma 4.8 stated below can now be used to upper bound deterministically the sum of the norms $||x_t||_{\Sigma_{t-1}}$. Using additionally the Cauchy-Schwarz (C.S.) inequality, one can write, if event E holds,

$$\begin{aligned}
\mathcal{R}_\theta(T,\mathcal{A}) &= \sum_{t=1}^{T} r_t \leq 2\sum_{t=1}^{T}\beta(t,\delta)\|x_t\|_{\Sigma_{t-1}} \\
&\leq 2\beta(T,\delta)\sum_{t=1}^{T}\|x_t\|_{\Sigma_{t-1}} \overset{\text{C.S.}}{\leq} 2\beta(T,\delta)\sqrt{T}\sqrt{\sum_{t=1}^{T}\|x_t\|^2_{\Sigma_{t-1}}} \\
&\overset{\text{Lemma 4.8}}{\leq} 2\beta(T,\delta)\sqrt{T}\sqrt{\frac{L^2\kappa^2}{\log(1+L^2\kappa^2\sigma^{-2})}d\log\left(1+T\frac{L^2\kappa^2}{\sigma^2 d}\right)} \\
&= \beta(T,\delta)\sqrt{Td}\sqrt{C_1\log\left(1+T\frac{L^2\kappa^2}{\sigma^2 d}\right)}
\end{aligned}$$

which concludes the proof.

□

Lemma 4.8.

$$\sum_{t=1}^{T}\|x_t\|^2_{\Sigma_{t-1}} \leq \frac{L^2\kappa^2}{\log(1+L^2\kappa^2\sigma^{-2})}d\log\left(1+T\frac{L^2\kappa^2}{\sigma^2 d}\right)$$

Proof of Lemma 4.8. First $\|x_t\|^2_{\Sigma_{t-1}} = \sigma^2\|x_t\|^2_{B(t-1)^{-1}}$. One has

$$\|x_t\|^2_{B(t-1)^{-1}} = x_t^T B(t-1)^{-1}x_t \underset{\text{CS}}{\leq} \|x_t\|\|B(t-1)^{-1}\|\|x_t\| \leq \frac{\kappa^2}{\sigma^2}\|x_t\|^2 \leq \frac{\kappa^2}{\sigma^2}L^2, \tag{4.12}$$

where we use that $\|B(t)^{-1}\| \leq \|B(t)\|^{-1} \leq (\sigma^2/\kappa^2)^{-1}$, since the eigenvalues of matrix $B(t)$ are lower bounded by σ^2/κ^2. For every $a > 0$ it can be easily shown that

$$\forall x \in [0,a],\ x \leq \frac{a}{\log(1+a)}\log(1+x).$$

Applying this inequality to the value $a = \kappa^2L^2\sigma^{-2}$ yields

$$\sum_{t=1}^{T}\|x_t\|^2_{B(t-1)^{-1}} \leq \frac{\kappa^2L^2\sigma^{-2}}{\log(1+\kappa^2L^2\sigma^{-2})}\sum_{t=1}^{T}\log\left(1+\|x_t\|^2_{B(t-1)^{-1}}\right). \tag{4.13}$$

Now, following for example [Dani et al., 2008], one can show that

$$\sum_{t=1}^{T}\log\left(1+\|x_t\|^2_{B(t-1)^{-1}}\right) = \log\left(\frac{\det(B(T))}{\det\left(\frac{\sigma^2}{\kappa^2}I_d\right)}\right). \tag{4.14}$$

Indeed,

$$\begin{aligned}
\det(B(t+1)) &= \det\left(\frac{\sigma^2}{\kappa^2}I_d + \sum_{s=1}^{t+1}x_s x_s^T\right) = \det\left(B(t)+x_{t+1}x_{t+1}^T\right) \\
&= \det(B(t))\det\left(I_d + B(t)^{-1/2}x_{t+1}x_{t+1}^T B(t)^{-1/2}\right) \\
&= \det(B(t))\det\left(I_d + (B(t)^{-1/2}x_{t+1})(B(t)^{-1/2}x_{t+1})^T\right) \\
&= \det(B(t))\left(1+\|B(t)^{-1/2}x_{t+1}\|^2\right) = \det(B(t))\left(1+\|x_{t+1}\|^2_{B(t)^{-1}}\right).
\end{aligned}$$

To obtain the last equality, we used that for every $v \in \mathbb{R}^d$, $\det(I_d + vv^T) = (1 + v^T v)$. By induction, it easily follows that $\det(B(T)) = \prod_{t=1}^{T}(1 + \|x_t\|^2_{B(t-1)^{-1}}) \det\left(\frac{\sigma^2}{\kappa^2} I_d\right)$ which proves (4.14). Using the bound (4.4) already used for the determinant of $B(T)$, one has

$$\det(B(T)) \leq \left(\frac{\sigma^2}{\kappa^2} + T\frac{L^2}{d}\right)^d .$$

This leads to an upper bound on (4.14), that yields, combined with Equation (4.13)

$$\sum_{t=1}^{T} \|x_t\|^2_{\Sigma_{t-1}} \leq \frac{\kappa^2 L^2}{\log(1 + \kappa^2 L^2 \sigma^{-2})} d \log\left(1 + T\frac{L^2\kappa^2}{d\sigma^2}\right)$$

□

Remark 4.9. *Introducing the notation $\sigma_t^2(x) = \|x\|^2_{\Sigma_{t-1}}$, the quantity introduced in (4.13) can be written*

$$\frac{1}{2}\sum_{t=1}^{T} \log\left(1 + \|x_t\|^2_{B(t-1)^{-1}}\right) = \frac{1}{2}\sum_{t=1}^{T} \log\left(1 + \sigma^{-2}\sigma_t^2(x_t)\right).$$

This last quantity is explicitly bounded by [Srinivas et al., 2010] in the more general context of Gaussian Process bandits. It is upper bounded by the maximum information gain called γ_T. Here we rather use techniques from [Dani et al., 2008] to upper bound this quantity in the simpler case of a linear kernel.

4.6.2 Proof of Theorem 4.6

As already used in [Russo and Van Roy, 2014], x_{t+1} and x^*_{t+1} have the same distribution conditionally to $\mathcal{H}_t$. Thus, for any function $f(x, y)$, if X is a $\mathcal{H}_t$ measurable random variable, the random variables $f(X, x_{t+1})$ and $f(X, x^*_{t+1})$ have the same conditional expectation with respect to $\mathcal{H}_t$. [Russo and Van Roy, 2014] use that $\mathbb{E}[U_t(x_{t+1})|\mathcal{H}_t] = \mathbb{E}[U_t(x^*_{t+1})|\mathcal{H}_t]$ for any confidence bound introduced in the analysis. Here we use similar equalities:

$$\mathbb{E}[\|x_{t+1}\|_{\Sigma_t}|\mathcal{H}_t] = \mathbb{E}[\|x^*_{t+1}\|_{\Sigma_t}|\mathcal{H}_t] \tag{4.15}$$
$$\mathbb{E}[\hat{\theta}(t)^T x_{t+1}|\mathcal{H}_t] = \mathbb{E}[\hat{\theta}(t)^T x^*_{t+1}|\mathcal{H}_t] \tag{4.16}$$

We start with the following decomposition:

$$r_{t+1} = \theta^T x^*_{t+1} - \theta^T x_{t+1} = \theta^T x^*_{t+1} - \hat{\theta}(t)^T x^*_{t+1} + \hat{\theta}(t)^T x^*_{t+1} - \theta^T x_{t+1}$$

Conditioning by $\mathcal{H}_t$ and using equality (4.16) yields, for any sequence β_t introduced below,

$$\begin{aligned}
\mathbb{E}[r_{t+1}] &\leq \mathbb{E}\left[(\theta - \hat{\theta}(t))^T x^*_{t+1}\right] + \mathbb{E}\left[(\hat{\theta}(t) - \theta)^T x_{t+1}\right] \\
&\leq \mathbb{E}\left[\|\theta - \hat{\theta}(t)\|_{\Sigma_t^{-1}} \|x^*_{t+1}\|_{\Sigma_t}\right] + \mathbb{E}\left[\|\theta - \hat{\theta}(t)\|_{\Sigma_t^{-1}} \|x_{t+1}\|_{\Sigma_t}\right] \\
&\leq \mathbb{E}\left[\|\theta - \hat{\theta}(t)\|_{\Sigma_t^{-1}} \left(\|x^*_{t+1}\|_{\Sigma_t} + \|x_{t+1}\|_{\Sigma_t}\right) \mathbb{1}_{\left(\|\theta - \hat{\theta}(t)\|_{\Sigma_t^{-1}} \leq \beta_{t+1}\right)}\right] \\
&\quad + \mathbb{E}\left[\|\theta - \hat{\theta}(t)\|_{\Sigma_t^{-1}} \left(\|x^*_{t+1}\|_{\Sigma_t} + \|x_{t+1}\|_{\Sigma_t}\right) \mathbb{1}_{\left(\|\theta - \hat{\theta}(t)\|_{\Sigma_t^{-1}} > \beta_{t+1}\right)}\right] \\
&\leq 2\beta_{t+1}\mathbb{E}\left[\|x_{t+1}\|_{\Sigma_t}\right] + 2L\kappa\,\mathbb{E}\left[\|\theta - \hat{\theta}(t)\|_{\Sigma_t^{-1}} \mathbb{1}_{\left(\|\theta - \hat{\theta}(t)\|_{\Sigma_t^{-1}} > \beta_{t+1}\right)}\right],
\end{aligned}$$

where the last inequality is obtained using (4.15) together with the upper bound $\|x\|_{\Sigma_t} \leq L\kappa$ for any $x \in \mathcal{D}_{t+1}$ (deduced from inequality (4.12) in Section 4.6). One obtains, by summing these terms and using the Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathrm{BR}_{\pi_0}(T, \mathrm{TS}_{\pi_0}) &\leq \sum_{t=1}^{T} 2\beta_t \mathbb{E}\left[\|x_t\|_{\Sigma_{t-1}}\right] + 2L\kappa \sum_{t=1}^{T} \mathbb{E}\left[\|\theta - \hat{\theta}(t-1)\|_{\Sigma_{t-1}^{-1}} \mathbb{1}_{\left(\|\theta-\hat{\theta}(t-1)\|_{\Sigma_{t-1}^{-1}} > \beta_t\right)}\right] \\
&\leq 2\beta_T \sqrt{T}\mathbb{E}\left[\sqrt{\sum_{t=1}^{T} \|x_t\|^2_{\Sigma_{t-1}}}\right] + 2L\kappa \sum_{t=0}^{T-1} \mathbb{E}\left[\|\theta - \hat{\theta}(t)\|_{\Sigma_t^{-1}} \mathbb{1}_{\left(\|\theta-\hat{\theta}(t)\|_{\Sigma_t^{-1}} > \beta_{t+1}\right)}\right]
\end{aligned}$$

The sum $\sum_{t=1}^{T} \|x_t\|^2_{\Sigma_{t-1}}$ that appears in the first term is upper bounded deterministically by Lemma 4.8, given in Section 4.6. To bound the second term, we use again that $\|\theta - \hat{\theta}(t)\|^2_{\Sigma_t^{-1}}$ follows a chi-square distribution with d degrees of freedom. Jensen inequality gives

$$\begin{aligned}
\left(\mathbb{E}\left[\|\theta - \hat{\theta}(t)\|_{\Sigma_t^{-1}} \mathbb{1}_{\left(\|\theta-\hat{\theta}(t)\|_{\Sigma_t^{-1}} > \beta_{t+1}\right)}\right]\right)^2 &\leq \mathbb{E}\left[\|\theta - \hat{\theta}(t)\|^2_{\Sigma_t^{-1}} \mathbb{1}_{\left(\|\theta-\hat{\theta}(t)\|^2_{\Sigma_t^{-1}} > \beta^2_{t+1}\right)}\right] \\
&= \int_{\beta^2_{t+1}}^{\infty} \frac{x}{2^{\frac{d}{2}}\Gamma\left(\frac{d}{2}\right)} x^{\frac{d}{2}-1} e^{-\frac{x}{2}} dx = d \int_{\beta^2_{t+1}}^{\infty} \frac{1}{2^{\frac{d+2}{2}}\Gamma\left(\frac{d+2}{2}\right)} x^{\frac{d+2}{2}-1} e^{-\frac{x}{2}} dx \\
&= d(1 - F_{\chi^2_{d+2}}(\beta^2_{t+1})).
\end{aligned}$$

where $F_{\chi^2_{d+2}}$ denotes the cdf of a chi-square distribution with $d+2$ degrees of freedom. Thus,

$$\sum_{t=0}^{T-1} \mathbb{E}\left[\|\theta - \hat{\theta}(t)\|_{\Sigma_t^{-1}} \mathbb{1}_{\left(\|\theta-\hat{\theta}(t)\|_{\Sigma_t^{-1}} > \beta_{t+1}\right)}\right] \leq \sqrt{d} \sum_{t=1}^{T} \sqrt{1 - F_{\chi^2_{d+2}}(\beta_t^2)} \tag{4.17}$$

Hence, choosing

$$\beta_t = \sqrt{Q\left(1 - \frac{1}{t^4}; \chi^2_{d+2}\right)} \leq \sqrt{d + 2 + 8\log t + 2\sqrt{(d+2)\log 2t^4}},$$

inequality (4.17) and Lemma 4.8 yield

$$\begin{aligned}
\mathrm{BR}_{\pi_0}(T, \mathrm{TS}_{\pi_0}) &\leq 2\beta_T \sqrt{T} \sqrt{\frac{L^2\kappa^2}{\log(1 + L^2\kappa^2\sigma^{-2})} d \log\left(1 + T\frac{L^2\kappa^2}{\sigma^2 d}\right)} + 2L\kappa\sqrt{d}\frac{\pi^2}{6} \\
&\leq d\sqrt{T} \sqrt{C_1 \log\left(1 + T\frac{L^2\kappa^2}{\sigma^2 d}\right)\left(3 + 8\log T + 2\sqrt{\log(2T^4)}\right)} + \frac{\pi^2 L\kappa\sqrt{d}}{3}
\end{aligned}$$

with $C_1 = \frac{4L^2\kappa^2}{\log(1+L^2\kappa^2\sigma^{-2})}$.

Chapter 5

Refined frequentist tools for best arm identification

In the first chapters of this thesis, we studied optimal algorithms for the objective of maximizing rewards in bandit models. A different objective can be considered, in which samples collected from the arms are not perceived as rewards, and the goal is to optimally explore the environment so as to identify the best arm(s) (without an incentive to exploration). This chapter gathers our contributions to this objective, called best arms identification (or pure-exploration) in bandit models.

Unlike that of regret minimization, the complexity of best arms identification is not so well understood, and our goal is to identify optimal algorithms in this framework. In this chapter, we introduce a dedicated notion of complexity in two different settings that have been considered in the literature: the fixed-budget and the fixed-confidence settings. We propose new lower bounds on these complexities, that involve information-theoretic quantities (like the Lai and Robbins' lower bound on the regret), as well as improved algorithms. Some of these algorithms, inspired by KL-UCB, are based on refined confidence intervals using Kullback-Leibler divergence. Whereas we were not able to close the gap between informational upper and lower bounds in the general case, we do identify the complexities for particular instances of two-armed bandits.

This chapter is based on two publications: a joint work with Shivaram Kalyanakrishnan ([Kaufmann and Kalyanakrishnan, 2013]), presented at the COLT conference in 2013, in which we study algorithms for m best arms identification, and a joint work with Olivier Cappé and Aurélien Garivier in which we introduce the complexities and notably present lower bounds ([Kaufmann et al., 2014b], submitted to JMLR). A shorter version of this last paper, [Kaufmann et al., 2014a], focused on two-armed bandits, was also presented at COLT in 2014.

Contents

5.1 Introduction

Recall a bandit model is a collection of K arms, where each arm ν_a $(1 \le a \le K)$ is a probability distribution on $\mathbb{R}$ with expectation μ_a. At each time $t = 1, 2, \dots$, an agent chooses an option $A_t \in \{1, \dots, K\}$ and receives an independent draw X_t of the corresponding arm ν_{A_t}. We denote by $\mathbb{P}_\nu$ (resp. $\mathbb{E}_\nu$) the probability law (resp. expectation) of the corresponding process (X_t). The agent's goal is now to identify the m best arms, that is, the set $\mathcal{S}_m^*$ of indices of the m arms with highest expectation. Letting $(\mu_{[1]}, \dots, \mu_{[K]})$ be the K-uple of the expectations $(\mu_1, \dots, \mu_K)$ sorted in decreasing order, we assume that the bandit model ν belong to a class $\mathcal{M}_m$ such that for every $\nu \in \mathcal{M}_m$, $\mu_{[m]} > \mu_{[m+1]}$, in which $\mathcal{S}_m^*$ is unambiguously defined. This last assumption is not necessary any more in an ϵ-relaxation of the problem that is sometimes considered in the literature: for some tolerance parameter $\epsilon \ge 0$ the agent has to ensure that $\hat{S}_m$ is included in the set of (ϵ, m)-optimal arms $\mathcal{S}_{m,\epsilon}^* = \{a : \mu_a \ge \mu_{[m]} - \epsilon\}$.

In order to identify $\mathcal{S}_m^*$, the agent must use a strategy defining which arms to sample from, but also when to stop sampling, and which set $\hat{S}_m$ to choose. The *sampling rule* determines how, at time t, the arm A_t is chosen based on the past observations; in other words, A_t is $\mathcal{F}_{t-1}$-measurable, with $\mathcal{F}_t = \sigma(A_1, Z_1, \dots, A_t, X_t)$. The *stopping rule* τ is a stopping time with respect to $(\mathcal{F}_t)_{t\in\mathbb{N}}$. The *recommendation rule* is a $\mathcal{F}_\tau$-measurable random subset $\hat{S}_m$ of $\{1, \dots, K\}$ of size m. This triple $((A_t), \tau, \hat{S}_m)$ entirely determines the strategy, which we denote in the sequel by $\mathcal{A}$.

In the bandit literature, two different settings have been considered. In the *fixed-confidence setting*, a risk parameter δ is fixed. A strategy $\mathcal{A}$ is called δ-*PAC* if, for every choice of $\nu \in \mathcal{M}_m$, $\mathbb{P}_\nu(\hat{S}_m = \mathcal{S}_m^*) \ge 1 - \delta$, or in the ϵ-relaxation described before, if $\mathbb{P}_\nu(\hat{S}_m \subset \mathcal{S}_{m,\epsilon}^*) \ge 1 - \delta$. The goal is, among the δ-*PAC* strategies, to minimize the expected number of draws $\mathbb{E}_\nu[\tau]$ (sometimes called *sample complexity*). In the *fixed-budget setting*, the number of draws τ is fixed in advance ($\tau = t$ almost surely) and the goal is to choose the sampling and recommendation rules so as to minimize $p_t(\nu) := \mathbb{P}_\nu(\hat{S}_m \ne \mathcal{S}_m^*)$. In the fixed-budget setting, a strategy $\mathcal{A}$ is called *consistent* if, for every choice of $\nu \in \mathcal{M}_m$, $p_t(\nu)$ goes to zero when t goes to infinity.

Recall that the complexity of regret minimization, the alternative objective considered so far, is well understood for parametric bandits. Indeed, [Lai and Robbins, 1985] define a dedicated notion of consistency and prove that, in generic one-parameter models,

$$\inf_{\mathcal{A} \text{ consistent}} \liminf_{t\to\infty} \frac{\mathbb{E}_\nu[R_\nu(T, \mathcal{A})]}{\log t} \ge \sum_{a:\mu_a<\mu_{[1]}} \frac{(\mu_{[1]} - \mu_a)}{\mathrm{KL}(\nu_a, \nu_{[1]})}.$$

We saw that there exists strategies whose regret attain this bound, including the two Bayesian strategies discussed in Chapter 2 and 3 and the KL-UCB algorithm of [Cappé et al., 2013] that uses informational upper-bounds.

Similarly, one would expect in the fixed-confidence setting a lower bound on the sample complexity $\mathbb{E}_\nu[\tau]$ of any δ-PAC algorithm (resp. in the fixed-budget setting a lower bound on the probability of error $p_t(\nu)$ of any consistent algorithm) that also features information-theoretic quantities, as well as algorithms whose sample complexity (resp. probability of error) matches the lower bound, called *matching algorithms*. In order to unify the two settings, we define the *complexity* $\kappa_C(\nu)$ (resp. $\kappa_B(\nu)$) of best arms identification in the fixed-confidence (resp. fixed-budget) setting, as follows:

$$\kappa_{\mathrm{C}}(\nu) = \inf_{\mathcal{A}\ \delta\text{–PAC}} \limsup_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\log\frac{1}{\delta}}, \qquad \kappa_{\mathrm{B}}(\nu) = \inf_{\mathcal{A}\ \text{consistent}} \left(\limsup_{t\to\infty} -\frac{1}{t}\log p_t(\nu)\right)^{-1}. \tag{5.1}$$

Heuristically, for a given bandit model ν and a small enough value of δ, a fixed-confidence optimal strategy needs an average number of samples of order $\kappa_C(\nu)\log\frac{1}{\delta}$ to identify the m best arms, whereas a fixed-budget optimal strategy requires approximately $t = \kappa_B(\nu)\log\frac{1}{\delta}$ draws in order to ensure a probability of error of order δ. Our goal in this chapter is to evaluate and compare these two complexities. We will do so by providing new lower bounds on $\kappa_C(\nu)$ and $\kappa_B(\nu)$ that feature information-theoretic quantities and by analyzing refined strategies, some of which matching the lower bounds in particular instances of two-armed bandit problems.

Most of the existing performance bounds for *pure-exploration* (a denomination that encompasses the fixed-budget and fixed-confidence settings) can be expressed using the two complexity measures defined above, as we see now.

The problem of best arms identification has been studied since the 1950s under the name 'ranking and identification problems'. The first advances on this topic are summarized in the monograph by [Bechhofer et al., 1968] who consider the fixed-confidence setting. More recently, in the same setting [Even-Dar et al., 2006] propose algorithms for (single) best arm identification in bounded bandit models, in which each arm ν_a is a probability distribution on $[0,1]$. m best arms identification with $m > 1$ was considered for example by [Kalyanakrishnan et al., 2012] (under the name Explore-m), who propose the LUCB (for Lower and Upper Confidence Bounds) algorithm, still for bounded bandit models. Bounded distributions are particular examples of subgaussian distributions, to which the proposed algorithms can be easily generalized. A relevant quantity introduced in the analysis of algorithms for bounded (or subgaussian) bandit models is the 'complexity term'

$$H(\nu) = \sum_{a\in\{1,2,\dots K\}} \frac{1}{\Delta_a^2} \quad \text{with} \quad \Delta_a = \begin{cases} \mu_a - \mu_{[m+1]} & \text{for } a \in \mathcal{S}_m^*, \\ \mu_{[m]} - \mu_a & \text{for } a \in (\mathcal{S}_m^*)^c. \end{cases} \tag{5.2}$$

The upper bound on the sample complexity of the LUCB algorithm of [Kalyanakrishnan et al., 2012] implies in particular that $\kappa_C(\nu) \le 292H(\nu)$. Some of the existing works on the fixed-confidence setting do not bound τ in expectation but rather show that $\mathbb{P}_\nu(\hat{\mathcal{S}}_m = \mathcal{S}_m^*, \tau = O(H(\nu))) \ge 1-\delta$. These results are not directly comparable with the complexity $\kappa_C(\nu)$, although no significant gap is to be observed yet.

For $m = 1$, the work of [Mannor and Tsitsiklis, 2004] provides a lower bound on $\kappa_C(\nu)$, who address Bernoulli bandit models with the ϵ-relaxation described before. The authors show that if an algorithm is δ-PAC, then in the bandit $\nu = (\mathcal{B}(\mu_1),\dots,\mathcal{B}(\mu_K))$ such that $\forall a,\ \mu_a \in [0,\alpha]$ for some $\alpha \in]0,1[$, there

exists two sets $\mathcal{M}_\alpha(\nu) \subset \mathcal{S}_1^*$ and $\mathcal{N}_\alpha(\nu) \subset \{1,\dots,K\}\backslash\mathcal{S}_1^*$ and a positive constant C_α such that

$$\mathbb{E}_\nu[\tau] \geq C_\alpha \left(\sum_{a\in\mathcal{M}_\alpha(\nu)} \frac{1}{\epsilon^2} + \sum_{a\in\mathcal{N}_\alpha(\nu)} \frac{1}{(\mu_{[1]}-\mu_a)^2} \right) \log\left(\frac{1}{8\delta}\right).$$

This bound is non asymptotic (as emphasized by the authors), although not completely explicit. In particular, the subset $\mathcal{M}_\alpha$ and $\mathcal{N}_\alpha$ do not always form a partition of the arms (it can happen that $\mathcal{M}_\alpha \cup \mathcal{N}_\alpha \neq \{1,\dots K\}$), hence the complexity term does not involve a sum over all the arms. For $m > 1$, the only lower bound available in the literature is the worst-case result of [Kalyanakrishnan et al., 2012]. It states that for every δ-PAC algorithm *there exists* a bandit model ν such that $\mathbb{E}_\nu[\tau] \geq K/(18375\epsilon^2)\log(m/8\delta)$. This yields, however, no lower bound on the complexity $\kappa_C(\nu)$.

The fixed-budget setting has been studied by [Audibert et al., 2010, Bubeck et al., 2011] for single best-arm identification in bounded bandit models. For multiple arm identification ($m > 1$), still in bounded bandit models, [Bubeck et al., 2013b] introduce the SAR (for Successive Accepts and Rejects) algorithm. An upper bound on the failure probability of the SAR algorithm yields $\kappa_B(\nu) \leq 8\log(K)H(\nu)$.

For $m = 1$, [Audibert et al., 2010] prove an asymptotic lower bound on the probability of error for Bernoulli bandit models. They state that for every algorithm and every bandit problem ν such that $\forall a$, $\mu_1 \in [\alpha, 1-\alpha]$, there exists a permutation of the arms ν' such that

$$p_t(\nu') \geq \exp(-t/C_\alpha H_2(\nu'))), \quad \text{with} \quad H_2(\nu) = \max_{i:\mu_{[i]}<\mu_{[1]}} \frac{i}{\left(\mu_{[1]}-\mu_{[i]}\right)^2}$$

and $C_\alpha = \alpha(1-\alpha)/(5+o(1))$. This result does not imply a lower bound on $\kappa_B(\nu)$ and for $m > 1$, no such lower bound exists either.

The gap between lower and upper bounds known so far does not permit to identify exactly the complexity terms $\kappa_B(\nu)$ and $\kappa_C(\nu)$ defined in (5.1). Not only do they involve imprecise multiplicative constants but by analogy with the Lai and Robbins' bound for the regret, the quantities $H(\nu)$ or $H_2(\nu)$ presented above are only expected to be relevant in the Gaussian case. Moreover, when $m > 1$, no lower bounds on the complexities are available.

Our contributions are the following. In the fixed-confidence setting, we first propose in Section 5.2 two algorithms, KL-LUCB and KL-Racing that use informational upper and lower bounds, transposing the improvements of KL-UCB from regret minimization to pure-exploration. The analysis of these algorithms leads to the first informational upper bound on $\kappa_C(\nu)$. We then propose in Section 5.3 a lower bound on $\kappa_C(\nu)$, that holds for general classes of bandit models, also when $m > 1$. This bound takes the form of a sum over all arms of an individual complexity term involving Kullback-Leibler divergence so that the quantity $H(\nu)$ appears as a subgaussian approximation. However, this lower bound is not attained by the KL-LUCB algorithm. For specific families of two-armed bandits (Gaussian bandits with known –but possibly different– variances, Bernoulli bandits) a refined lower bound as well as improved algorithms then lead to an exact expression of some complexity terms. Interestingly, we prove that for Gaussian bandit models $\kappa_C(\nu) = \kappa_B(\nu)$, whereas for Bernoulli bandit models $\kappa_C(\nu) > \kappa_B(\nu)$, showing that the two complexities are not equal in general. The particular case of two-armed bandits is studied in Section 5.4.

5.2 Algorithms: KL-LUCB and KL-Racing

In this Section, we present general algorithms for m best arms identification among K arms. [Bubeck et al., 2011] show that in the fixed-budget setting, for single best arm identification, any sampling strategy designed to minimize regret performs poorly with respect to the *simple regret* $r_t := \mu^* - \mu_{\hat{S}_1}$, a quantity closely related to the probability $p_t(\nu)$ of recommending the wrong arm. Therefore, good strategies for best arms identification are expected to be quite different from UCB-like strategies.

In Section 5.2, we present a review of existing algorithms for the fixed-confidence and fixed-budget settings, identifying two main classes of algorithms: those based on *uniform sampling and eliminations* and those based on *adaptive sampling*. We introduce, for the fixed-confidence setting, two generic algorithms based on confidence intervals (using both upper *and* lower confidence bounds) that belong to each of these classes and analyse in Section 5.2.2 two particular instances, KL-Racing and KL-LUCB. These algorithms are shown to perform well in practice (experiments are reported in Section 5.2.3) and the upper bounds obtained on their sample complexity are the first upper bounds featuring information-theoretic quantities.

To ease the presentation, we will restrict our attention to Bernoulli bandit models, but KL-LUCB and KL-Racing (and our results) can be extended to exponential bandits, using the appropriate d function (see Section 1.2.2 in Chapter 1). Recall a Bernoulli bandit model, of the form $\nu = (\mathcal{B}(\mu_1), \dots, \mathcal{B}(\mu_K))$, can be either regarded as a bounded bandit model or as an exponential bandit model. In the regret minimization framework, this allowed to use algorithms based on upper confidence bounds built either with Hoeffding's inequality (UCB1, designed for any bandit bounded bandit model) or on Chernoff inequality, that lead to Bernoulli-specific confidence regions based on KL-divergence (KL-UCB). In this Section, we introduce similar improved algorithms for the problem of m-best arms identification in Bernoulli bandit models in the fixed-confidence setting.

We consider in this section the ϵ-relaxation of the fixed-confidence setting described above, and assume to ease the notation that the arms are ordered such that $\mu_1 \geq \cdots \geq \mu_m \geq \mu_{m+1} \geq \cdots \geq \mu_K$.

5.2.1 Two classes of algorithms based on confidence intervals

Virtually all the algorithms proposed to date for pure-exploration problems can be classified according to their sampling strategy: algorithms using *uniform sampling and eliminations* maintain a set of remaining arms, and sample all these remaining arms at each round, whereas algorithms using *adaptive sampling* sample at each round one or two well-chosen arms.

Just as upper confidence bounds have been used successfully in the regret setting, most existing algorithms for the fixed-confidence setting have used both upper and lower confidence bounds on the means of the arms. We state here a generic version of an algorithm using uniform sampling and eliminations, Racing, and a generic version of an adaptive sampling algorithm, LUCB. To describe these contrasting heuristics, we use generic confidence intervals, denoted by $\mathcal{I}_a(t) = [L_a(t), U_a(t)]$, where t is the round of the algorithm, $L_a(t)$ and $U_a(t)$ are the lower and upper confidence bounds on the mean of arm a. Let $N_a(t)$ denote the number of draws, and $S_a(t)$ the sum of the rewards gathered from arm a up to time t. Let $\hat{\mu}_a(t) = \frac{S_a(t)}{N_a(t)}$ be the corresponding empirical mean reward, and let $\hat{\mu}_{a,u}$ be the empirical mean of the first u i.i.d. samples from arm a. Additionally, let $J(t)$ be the set of m arms with the highest empirical means at time t (for the Racing algorithm, $J(t)$ only includes $m' \leq m$ arms if $m - m'$ have already been selected). Also, l_t and u_t are two 'critical' arms from $J(t)$ and $J(t)^c$ that are likely to be misclassified:

$$u_t = \operatorname*{argmax}_{b \notin J(t)} U_b(t) \quad \text{and} \quad l_t = \operatorname*{argmin}_{a \in J(t)} L_a(t). \tag{5.3}$$

Algorithm 3 Racing

Require: $\epsilon \geq 0$ (tolerance level), U, L (confidence bounds)
$\mathcal{R} = \{1, ..., K\}$ set of remaining arms. $S = \varnothing$ set of selected arms.
$\mathcal{D} = \varnothing$ set of discarded arms. $t = 1$ (current round of the algorithm)
while $|\mathcal{S}| < m$ and $|\mathcal{D}| < K - m$ **do**
Sample all the arms in $\mathcal{R}$ update confidence intervals
Compute $J(t)$ the set of empirical $m - |\mathcal{S}|$ best arms and $J(t)^c = \mathcal{R} \backslash J(t)$
Compute u_t and l_t according to (5.3)
Compute a_B (resp. a_W) the empirical best (resp. worst) arm in $\mathcal{R}$
if $(U_{u_t}(t) - L_{a_B}(t) < \epsilon) \cup (U_{a_W}(t) - L_{l_t}(t) < \epsilon)$ **then**
$a = \underset{\{a_B, a_W\}}{\text{argmax}} \left((U_{u_t}(t) - L_{a_B}(t)) \mathbb{1}_{U_{u_t}(t) - L_{a_B}(t) < \epsilon}; (U_{a_W}(t) - L_{l_t}(t)) \mathbb{1}_{U_{a_W}(t) - L_{l_t}(t) < \epsilon} \right)$
Remove arm a: $\mathcal{R} = \mathcal{R} \backslash \{a\}$
If $a = a_B$ select a: $\mathcal{S} = \mathcal{S} \cup \{a\}$, else discard a: $\mathcal{D} = \mathcal{D} \cup \{a\}$
end if
t=t+1
end while
return $\mathcal{S}$ if $|\mathcal{S}| = m$, $\mathcal{S} \cup \mathcal{R}$ otherwise

The Racing algorithm. The first algorithms for best arms identification (see [Bechhofer et al., 1968]) used pure uniform sampling (sometimes called vector-at-a-time sampling): they proceed in rounds, where all the arms are drawn, and stop when a global stopping criterion is met. [Paulson, 1964] introduces the idea of coupling eliminations to uniform sampling to reduce the number of sample used. His goal was to find the (single) best arm in a Gaussian bandit model. [Jennison et al., 1982] in the same setup define general elimination procedures. An elimination procedure depends on some function g, samples all the remaining arms in $\mathcal{R}$ at each round t and eliminate arm b if there exists $a \in \mathcal{R}$ such that $S_a(t) - S_b(t) > g(t)$. This criterion can be rephrased as some lower confidence bound for μ_a being larger than some upper confidence bound for μ_b: this is the idea of Racing as introduced by [Maron and Moore, 1997] in the context of model selection.

For finding the m best arms with $m > 1$, [Levin and Leu, 2008] modify an existing procedure by introducing eliminations: both accepts and rejects (called eliminations and recruitments). Accepts and

Algorithm 4 LUCB

Require: $\epsilon \geq 0$ (tolerance level), U, L (confidence bounds)
$t = 1$ (number of stage of the algorithm), $B(1) = \infty$ (stopping index)
for a=1...K **do**
Sample arm a, compute confidence bounds $U_a(1)$, $L_a(1)$
end for
while $B(t) > \epsilon$ **do**
Draw arm u_t and l_t. $t = t + 1$.
Update confidence bounds, set $J(t)$ and arms u_t, l_t
$B(t) = U_{u_t}(t) - L_{l_t}(t)$
end while
return $J(t)$.

rejects based on confidence intervals are introduced by [Heidrich-Meisner and Igel, 2009] and applied within the context of reinforcement learning. The authors do not formally analyse the algorithm's sample complexity, as we do here. The Racing algorithm, stated precisely as Algorithm 3, samples at each round t all the remaining arms, and updates the confidence bounds. Then a decision is made to possibly select the empirical best arm if its lower confidence bound (LCB) is larger than the upper confidence bounds (UCBs) of all arms in $J(t)^c$, or to discard the empirical worst arm if its UCB is smaller than the LCBs of all arms in $J(t)$. The successive elimination algorithm ([Even-Dar et al., 2006]) for finding the single best arm can be regarded as a specification of Algorithm 3 using Hoeffding bounds.

The LUCB algorithm. A general version of the LUCB algorithm proposed by [Kalyanakrishnan et al., 2012] is stated in Algorithm 4, using generic confidence bounds U and L, while the original LUCB uses Hoeffding confidence regions. Unlike Racing, this algorithm does not sample the arms uniformly; rather, it draws at each round the two critical arms u_t and l_t. This sampling strategy is associated with the natural stopping criterion $(B(t) < \epsilon)$ where $B(t) := U_{u_t}(t) - L_{l_t}(t)$. That is the algorithm stops when the confidence intervals for the means of the arms in $J(t)$ and those for the means of the arms in $J(t)^c$ are well separeted. An illustration of the LUCB algorithm can be found in Figure 5.1

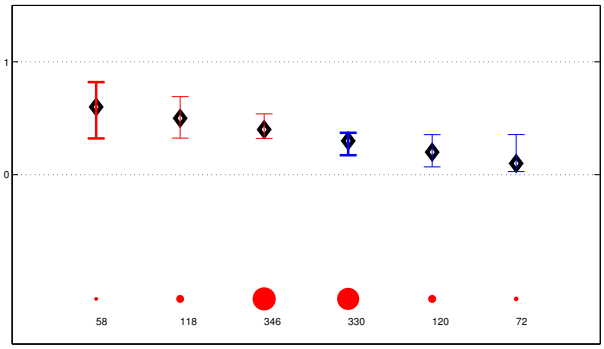

Figure 5.1: KL-LUCB for finding the $m = 3$ best arms among 6. In red (resp. blue) arms in $J(t)$ (resp. $J(t)^c$) and their confidence intervals. Arms in bold are l_t and u_t. Red circles represent the current number of draws of each arm.

The UGapEc algorithm of [Gabillon et al., 2012] also uses adaptive sampling and is very close to LUCB: it uses an alternative definition of $J(t)$ using confidence bounds on the simple regret, and a correspondingly different stopping criterion $B(t)$. But as LUCB, it also samples the corresponding critical arms u_t or l_t.

KL-Racing and KL-LUCB. The two algorithms mentioned above both use generic upper and lower confidence bounds on the mean of each arm, and one has the intuition that the smaller these confidence regions are, the smaller the sample complexity of these algorithms will be. Most of the previous algorithms use Hoeffding bounds, of the form

$$U_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{\beta(t,\delta)}{2N_a(t)}} \quad \text{and} \quad L_a(t) = \hat{\mu}_a(t) - \sqrt{\frac{\beta(t,\delta)}{2N_a(t)}} \tag{5.4}$$

for some *exploration rate* $\beta(t,\delta)$. Previous work [Mnih et al., 2008, Heidrich-Meisner and Igel, 2009, Gabillon et al., 2012] has also considered the use of empirical Bernstein bounds, that can be tighter. Here, we introduce the use of confidence regions based on KL-divergence for Explore-m, inspired by recent improvements in the regret setting ([Cappé et al., 2013]. We define, for some exploration rate $\beta(t,\delta)$ (that is a function of t and δ, and not only t as in the regret minimization framework),

$$u_a(t) := \max\{q \in [\hat{\mu}_a(t),1] : N_a(t)d(\hat{\mu}_a(t),q) \leq \beta(t,\delta)\}, \text{ and} \tag{5.5}$$
$$l_a(t) := \min\{q \in [0,\hat{\mu}_a(t)] : N_a(t)d(\hat{\mu}_a(t),q) \leq \beta(t,\delta)\}. \tag{5.6}$$

As already noted in Chapter 1 (Section 1.2.2) , for Bernoulli distributions KL-confidence regions are always smaller than those obtained with Hoeffding bounds, while sharing the same coverage probability:

$$\hat{\mu}_a(t) - \sqrt{\frac{\beta(t,\delta)}{2N_a(t)}} \leq l_a(t) \text{ and } u_a(t) \leq \hat{\mu}_a(t) + \sqrt{\frac{\beta(t,\delta)}{2N_a(t)}}. \tag{5.7}$$

We define, for a given function β, the **KL-Racing** and **KL-LUCB** algorithms with exploration rate β as the instances of Racing and LUCB, respectively, that use $u_a(t)$ and $l_a(t)$ as confidence bounds. Section 5.2.2 provides conditions on β for both algorithms to be δ-PAC and sample complexity bounds under these conditions. In our theoretical and experimental analysis to follow, we address the "KL versus Hoeffding" and "uniform versus adaptive sampling" questions.

Other algorithms for the fixed-confidence and fixed-budget setting. Among the family of algorithms using uniform sampling and eliminations, the algorithms presented so far, including (KL)-Racing, consider the possibility of eliminating one arm at each round. Algorithms inspired by the Median Elimination algorithm of [Even-Dar et al., 2006] for $m = 1$ - extended to Halving for $m > 1$ by [Kalyanakrishnan and Stone, 2010] - are quite different. The basic Median Elimination algorithm consists in phases and at the end of each phase, the empirical worst half of the arms is discarded, based on the samples gathered in this phase only. For Bernoulli bandits, the Exponential-Gap Elimination of [Karnin et al., 2013] uses the Median Elimination algorithm as a subroutine. For this δ-PAC algorithm, there exists a constant C such that, with high probability, the number of samples needed to identify the best arm is upper bounded by

$$C\sum_{a=2}^{K}\frac{1}{\Delta_a^2}\log\left(\frac{1}{\delta}\log\frac{1}{\Delta_a}\right) \tag{5.8}$$

This is an improvement when compared to most existing algorithm for which τ is upper bounded by some constants multiplied by $H(\nu)\log\left(H(\nu)/\delta\right)$: the relatively small factor $\log 1/\Delta_a$ compared to $H(\nu)$ can be a significant improvement, especially when the number of arms is large. Moreover, in the regime where δ is fixed and Δ_a goes to zero, [Jamieson et al., 2014] show that the term $\Delta_a^{-2}\log\log\Delta_a^{-2}$ is optimal, and propose an upper bound similar to (5.8) for the LIL-UCB algorithm. In both cases the constant C is quite large, which does not lead to an improved upper bound on the complexity term $\kappa_C(\nu)$. For the KL-Racing and KL-LUCB algorithms, we therefore propose in the next Section upper bound on τ with explicit constants, featuring moreover informational quantities, and no longer the squared gaps Δ_a.

In the fixed-budget setting, [Bubeck et al., 2013b] propose the Successive Accepts and Rejects (SAR) algorithm for the objective of finding the m best arms, generalizing the Successive Reject algorithm of [Audibert et al., 2010] for $m = 1$. This algorithm samples uniformly the arms in each of the $K-1$ phases

with predetermined length, and at the end of each phase exactly one arm is eliminated. The empirical best arm is selected or the empirical worst discarded, according to its empirical gap with $J(t)^c$ or $J(t)$ respectively (a criterion than cannot be formulated with confidence intervals). A variant of this algorithm in which at the end of each phase half of the remaining arms is eliminated has been recently proposed by [Karnin et al., 2013], showing improvements on the theoretical probability of error compared to that of SAR, but which does not improve the resulting upper bound on the (asymptotic) complexity $\kappa_B(\nu)$. Some adaptive sampling algorithms do exist for the fixed-budget setting too, namely UCB-E of [Audibert et al., 2010] for $m = 1$, or UGapEb of [Gabillon et al., 2012]. These algorithm are not efficient in practice since they share the need to know the complexity term $H(\nu)$. In the paper [Kaufmann and Kalyanakrishnan, 2013] we propose an other adaptive algorithm for the fixed-budget setting, KL-LUCB-E, derived from KL-LUCB by choosing the exploration rate β as a function of n, but suffering from the same weakness as its existing counterpart. In the practical experiments of Section 5.2.3, we will discuss further the interest of using adaptive sampling in the fixed-budget setting.

5.2.2 Analysis of KL-Racing and KL-LUCB

Theorem 5.1 gives a choice of β for which KL-Racing and KL-LUCB are correct with probability at least δ (δ-PAC). Note that these choices of β lead to the same guarantees for their Hoeffding counterpart, (Hoeffding)-Racing and LUCB.

Theorem 5.1. *The KL-Racing and KL-LUCB algorithms using* $\beta(t,\delta) = \log\left(\frac{k_1 K t^\alpha}{\delta}\right)$ *as an exploration rate, with* $\alpha > 1$ *and* $k_1 > 1 + \frac{1}{\alpha - 1}$*, are correct with probability at least* $1 - \delta$.

In order to introduce our results on the sample complexity of KL-Racing and KL-LUCB, one needs to introduce a new informational quantity: *Chernoff information*. The Chernoff information between two Bernoulli distributions $\mathcal{B}(x)$ and $\mathcal{B}(y)$, denoted by $d^*(x,y)$, is defined by

$$d^*(x,y) = d(z^*,x) = d(z^*,y) \text{ where } z^* \text{ is the unique } z \text{ such that } d(z,x) = d(z,y).$$

This definition can be generalized to distributions in an exponential family. We will discuss later the interpretation we can propose for the relevance of Chernoff information as a complexity measure for pure-exploration, here we first explain why Chernoff information in necessary in the proof of Theorem 5.6. In Theorem 5.5, an other quantity is involved, related to Chernoff information:

$$d^{**}(x,y) = d^*(z^{**},x) = d^*(z^{**},y) \text{ where } z^{**} \text{ is the unique } z \text{ such that } d^*(z,x) = d^*(z,y).$$

A tight concentration result involving Chernoff information. In our analysis of KL-LUCB, we need to bound the probability that some constant c belongs to the interval $\mathcal{I}_a(t)$ after this arm has already been sufficiently sampled. Deriving such a result for intervals based on KL-divergence brings up Chernoff information:

Lemma 5.2. *Let* $T \geq 1$ *be an integer. Let* $\gamma > 0$ *and* $c \in]0,1[$ *be such that* $\mu_a \neq c$.

$$\sum_{s=\lceil \frac{\gamma}{d^*(\mu_a,c)} \rceil + 1}^{T} \mathbb{P}\left(sd\left(\hat{\mu}_{a,s}, c\right) \leq \gamma\right) \leq \frac{\exp\left(-\gamma\right)}{d^*(\mu_a,c)}.$$

This result is a corollary of the important Lemma 5.4 below, proved in Section 5.6.2, that is in some sense an optimal deviation result involving KL-divergence. Some functions based on KL-divergence need to be defined in order to state this more general result.

Definition 5.3. *Let $C_1 > 1$, $(y,c) \in]0,1[^2$, $y \neq c$. Let $s_{C_1}(y,c)$ be the function implicitly defined by*

$$d(s_{C_1}(y,c),c) = \frac{d(y,c)}{C_1} \text{ and } s_{C_1}(y,c) \in (y,c),$$

where (y,c) denotes the interval $[y,c]$ if $y < c$, and $[c,y]$ otherwise. We define F_{C_1} as:

$$F_{C_1}(y,c) = \frac{C_1 d(s_{C_1}(y,c),y)}{d(y,c)}.$$

Lemma 5.4. *Let $C_1 > 1$, $\gamma > 0$ and $c \in]0,1[$ such that $\mu_a \neq c$. For any integer T,*

$$\sum_{u=\lceil \frac{C_1\gamma}{d(\mu_a,c)} \rceil + 1}^{T} \mathbb{P}\left(u d(\hat{\mu}_{a,u},c) \leq \gamma\right) \leq \frac{\exp\left(-F_{C_1}(\mu_a,c)\gamma\right)}{d\left(s_{C_1}(\mu_a,c),\mu_a\right)}. \tag{5.9}$$

The sum in Lemma 5.4 is bounded tightly in the recent analysis of KL-UCB by [Cappé et al., 2013] for the value $C_1 = 1$. However, the related bound shows no exponential decay in γ, unlike the one we prove for $C_1 > 1$ in Section 5.2.4. Whereas it was used to bound an expectation for KL-UCB, in the proof of Theorem 5.6 we will use it to bound a probability and thus need this exponential decay. This technical difference ushers in the bifurcation between Chernoff information and KL-divergence. Indeed, $F_{C_1}(\mu_a,c)$, that is the optimal rate in the exponential (see Section 5.2.4), depends on the problem and to be able to later choose an exploration rate that does not, we have to choose C_1 such that $F_{C_1}(\mu_a,c) = 1$. As we can see below, there is a unique constant $C_1(\mu_a,c)$ satisfying $F_{C_1(\mu_a,c)}(\mu_a,c) = 1$ and it is related to Chernoff information:

$$\begin{aligned} F_{C_1}(\mu_a,c) = 1 &\Leftrightarrow d(s_{C_1}(\mu_a,c),\mu_a) = \frac{d(\mu_a,c)}{C_1} \Leftrightarrow d(s_{C_1}(\mu_a,c),\mu_a) = d(s_{C_1}(\mu_a,c),c) \\ &\Leftrightarrow s_{C_1}(\mu_a,c) \text{ is the unique } z \text{ satisfying } d(z,\mu_a) = d(z,c). \end{aligned}$$

Hence, $C_1(\mu_a,c)$ can be rephrased using **Chernoff information** which is precisely defined for two Bernoulli by $d^*(\mu_a,c) = d(z^*,c) = d(z^*,\mu_a)$. One gets

$$C_1(\mu_a,c) = d(\mu_a,c)/d^*(\mu_a,c) \tag{5.10}$$

and invoking Lemma 5.4 with this particular value of C_1 leads to Lemma 5.2.

Sample complexity results and discussion. Theorem 5.5 gives an upper bound on the number of samples used by KL-Racing that holds with high probability, when $\epsilon = 0$ (i.e. when $\mu_m > \mu_{m+1}$ and no relaxation is considered). It involves the quantity d^{**} defined above and related to Chernoff information. We do not provide an upper bound on the expectation of τ for this algorithm, and the proof of Theorem 5.5 share a common structure with the analysis of many algorithms for which only a high-probability bound on τ is provided (e.g. [Even-Dar et al., 2006, Gabillon et al., 2012, Karnin et al., 2013]): on some event W on which the algorithm outputs the right subset, the number of draws of each arm is upper bounded deterministically (see details in Section 5.2.4).

Theorem 5.5. *Let $\beta(t,\delta) = \log\left(\frac{k_1 K t^\alpha}{\delta}\right)$, with $\alpha > 1$ and $k_1 > 1 + \frac{1}{\alpha - 1}$. The number of samples τ used in KL-Racing with $\epsilon = 0$ is such that*

$$\mathbb{P}_\nu\left(\tau \leq \alpha\left(\sum_{a=1}^{m} \frac{1}{d^{**}(\mu_a,\mu_{m+1})} + \sum_{a=m+1}^{K} \frac{1}{d^{**}(\mu_a,\mu_m)}\right)\log\frac{1}{\delta} + o\left(\log\frac{1}{\delta}\right), \hat{\mathcal{S}}_m = \mathcal{S}_m^*\right) \geq 1 - 2\delta.$$

For KL-LUCB, we provide in Theorem 5.6 an upper bound on $\mathbb{E}[\tau]$, in the ϵ-relaxation framework, for $\epsilon \geq 0$. It involves, for $c \in [\mu_{m+1}, \mu_m]$, the quantity

$$H^*_{\epsilon,c}(\nu) := \sum_{a \in \{1,\dots,K\}} \frac{1}{\max(d^*(\mu_a, c), \epsilon^2/2)}.$$

The proof of Theorem 5.6 follow from an upper bound on $\mathbb{P}(\tau \geq T)$ for any deterministic time T and is given in Section 5.2.4. In the sequel, the parameter $c \in [\mu_{m+1}, \mu_m]$ that we introduce in the analysis will be assumed to be in $]0,1[$, excluding the case $\mu_m = \mu_{m+1} = 0$ or 1.

Theorem 5.6. *Let $c \in [\mu_m, \mu_{m+1}]$, $\epsilon \geq 0$. Let $\beta(t,\delta) = \log\left(\frac{k_1 K t^\alpha}{\delta}\right)$ with $k_1 > 1 + \frac{1}{\alpha - 1}$. Then for $\alpha > 1$,*

$$\mathbb{P}_\nu\left(\tau \leq 2\alpha H^*_{\epsilon,c} \log\left(\frac{e k_1 K (H^*_{\epsilon,c})^\alpha}{\delta} \log\left(\frac{k_1 K (H^*_{\epsilon,c})^\alpha}{\delta}\right)\right), \hat{S}_m \subseteq \mathcal{S}^*_{m,\epsilon}\right) \geq 1 - 2\delta.$$

Moreover, for $\alpha > 2$,

$$\mathbb{E}_\nu[\tau] \leq 4\alpha H^*_{\epsilon,c} \log\left(\frac{e k_1 K (H^*_{\epsilon,c})^\alpha}{\delta} \log\left(\frac{k_1 K (H^*_{\epsilon,c})^\alpha}{\delta}\right)\right) + C_\alpha,$$

with $C_\alpha = \frac{2^{\alpha-1}\delta}{k_1} \sum_{t=1}^{\infty} \frac{\log(k_1 K t^\alpha/\delta)+1}{t^{\alpha-1}}$.

Theorem 5.5 and Theorem 5.6 provide upper bounds on τ for the KL-Racing and KL-LUCB algorithms that involve informational quantities and explicit constants. For KL-LUCB, we believe that the finite-time upper bound on the sample complexity is the first of its kind involving KL-divergence (through Chernoff information). This results yields (for $\epsilon = 0$) an upper bound on $\kappa_C(\nu)$, the complexity term introduced in (5.1):

$$\kappa_C(\nu) \leq 8 H^*_0(\nu), \quad \text{with} \quad H^*_0(\nu) := \min_{c \in [\mu_{m+1}, \mu_m]} \sum_{a=1}^{K} \frac{1}{d^*(\mu_a, c)}.$$

Pinsker's inequality shows that $d^*(x,y) \geq (x-y)^2/2$ and $d^{**}(x,y) \geq (x-y)^2/8$, which gives a relationship with the complexity term $H(\nu)$ that involves squared gaps defined in (5.2) : $H^*_0(\nu) \leq 8H(\nu)$. Although $H^*_0(\nu)$ cannot be shown to be strictly smaller than $H(\nu)$ on every problem (this will be the case for example when the parameters of the arms are small), the explicit bound in Theorem 5.6 still improves over that of [Kalyanakrishnan et al., 2012], which implies that $\kappa_C(\nu) \leq 192 H(\nu)$.

We believe that the constant c that appears in Theorem 5.6 is an artifact of our proof, but we are currently unable to eliminate it. We would therefore conjecture that it is possible to give an upper bound on $\kappa_C(\nu)$ that rather involves the quantity

$$\sum_{a=1}^{m} \frac{1}{d^*(\mu_a, \mu_{m+1})} + \sum_{a=m+1}^{K} \frac{1}{d^*(\mu_a, \mu_m)}.$$

We can propose an interpretation for this quantity, using that Chernoff information is a relevant quantity in testing problems ([Cover and Thomas, 2006]). Let $X_1, X_2, \dots, X_n$ be n i.i.d. samples and $H_1 : X_i \sim \mathcal{B}(x)$ against $H_2 : X_i \sim \mathcal{B}(y)$ be two alternative hypotheses. For a test ϕ, let $\alpha_n(\phi) = \mathbb{P}_1(\phi = 2)$ and $\beta_n = \mathbb{P}_2(\phi = 1)$ be respectively the type I and type II error. Chernoff's Theorem states that when the objective is to minimize both type I and type II error, the best achievable exponent is

$$d^*(x,y) = \lim_{n \to \infty} -\frac{1}{n} \log \min_{\phi} \max(\alpha_n(\phi), \beta_n(\phi)).$$

Hence, for small δ, $\frac{1}{d^*(\mu_a,\mu_m)}\log(\frac{1}{\delta})$ (resp. $\frac{1}{d^*(\mu_a,\mu_{m+1})}\log(\frac{1}{\delta})$) represents the minimal number of samples needed to discriminate between arm a and arm m (resp. arm a and arm $m+1$) with both error probabilities smaller than δ.

However, the general lower bound on $\kappa_C(\nu)$ we propose in the next Section does not involve Chernoff information but Kullback-Leibler divergence, so it is still to be determined whether KL-LUCB is optimal with respect to the complexity $\kappa_C(\nu)$. Nevertheless, Chernoff information will turn out to be a relevant measure of complexity for two-armed bandits, in the fixed-budget setting.

5.2.3 Numerical experiments

On the basis of our theoretical analysis from the previous section, could we expect the "KL-ized" versions of our algorithms to perform better in practice? Does being "fully sequential" make our adaptive sampling algorithms more efficient than uniform sampling algorithms *in practice*? In this section, we present numerical experiments that answer both these questions in the affirmative.

In our experiments, in addition to (KL-)LUCB and (KL-)Racing, we include (KL-)LSC, an adaptive sampling algorithm akin to (KL)-LUCB. This algorithm uses the same stopping criterion as (KL)-LUCB, but rather than sample arms u_t and l_t at stage t, (KL)-LSC samples the **l**east-**s**ampled arm from $J(t)$ (or $J(t)^c$) that **c**ollides (overlaps by at least ϵ) with some arm in $J(t)^c$ ($J(t)$). To ensure that all algorithms are provably PAC, we run them with the parameters $\alpha = 1.1$, $k_1 = 11.1$ justified by Theorem 5.1. Results are summarized in Figure 5.2.

As a first order of business, we consider bandit instances with $K = 10, 20, \ldots, 60$ arms; we generate 1000 random instances for each setting of K, with each arm's mean drawn uniformly at random from $[0, 1]$. We set $m = \frac{K}{5}$, $\epsilon = 0.1$, $\delta = 0.1$. The expected sample complexity of each algorithm on the bandit instances for each K are plotted in Figure 5.2(a). Indeed we observe for each K that (1) the KL-ized version of each algorithm enjoys a lower sample complexity, and (2) (KL)-LUCB outperforms (KL-)LSC, which outperforms (KL-)Racing.

These trends, aggregated from multiple bandit instances, indeed hold for nearly every individual bandit instance therein. In fact, we find that KL-izing has a more pronounced effect on bandit instances with means close to 0 or 1. For illustration, consider instance B_1 ($K = 15$; $\mu_1 = \frac{1}{2}$; $\mu_a = \frac{1}{2} - \frac{a}{40}$ for $a = 2, 3, \ldots, K$), an instance used by [Bubeck et al., 2013b] (see Experiment 5). Figure 5.2(b) compares the runs of LUCB and KL-LUCB both on B_1 (with $m = 3$, $\epsilon = 0.04$, $\delta = 0.1$), and a "scaled-down" version B_2 (with $m = 3$, $\epsilon = 0.02$, $\delta = 0.1$) in which each arm's mean is half that of the corresponding arm's in B_1 (and thus closer to 0). While LUCB and KL-LUCB both incur a higher sample complexity on the harder B_2, the latter's relative economy is clearly visible in the graph.

How conservative are the stopping criteria of our PAC algorithms? In our third experiment, we halt these algorithms at intervals of 1000 samples, and at each stage record the probability that the set $J(t)$ of m empirical best arms that would be returned at that stage is non-optimal. Results from this experiment, again on B_1 (with $m = 3$, $\epsilon = 0.04$, $\delta = 0.1$), are plotted in Figure 5.2(c). Notice that (KL)-LUCB indeed drives down the mistake probability much faster than its competitors. Yet, even if all the algorithms have an empirical mistake probability smaller than δ after 5,000 samples, they only stop after at least 20,000 episodes, leaving us to conclude that our formal bounds are rather conservative. On the low-reward instance B_2 (with $m = 3$, $\epsilon = 0.02$, $\delta = 0.1$), we observe that KL-LUCB indeed reduces the mistake probability more quickly than LUCB, indicating a superior sampling strategy. This difference is between LUCB and KL-LUCB is *not* apparent on B_1 in Figure 5.2(c).

We test KL-LUCB-$log(t)$, a version of KL-LUCB with an exploration rate of $log(t)$ (which yields no provable guarantees) as a candidate for the fixed-budget setting. On B_1 (with $n = 4000$), we compare this

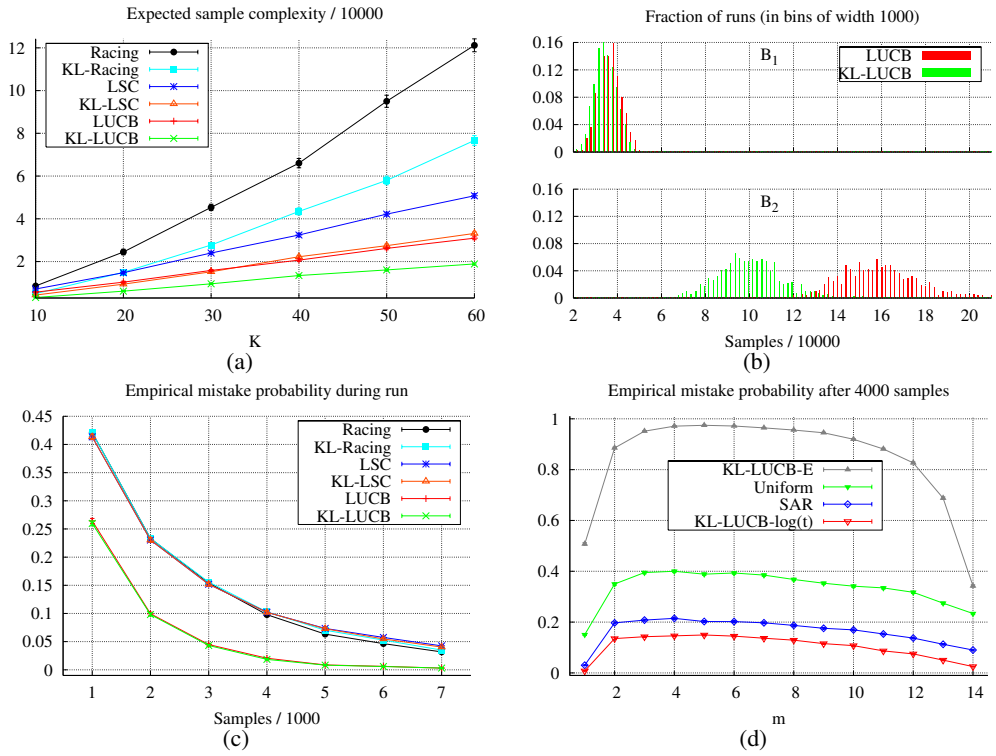

Figure 5.2: Experimental results (descriptions in text).

algorithm with KL-LUCB-E, discussed in the paper [Kaufmann and Kalyanakrishnan, 2013], which has a provably-optimal exploration rate involving the problem complexity ($H_\epsilon^\star \approx 13659$). Quite surprisingly, we find that KL-LUCB-$log(t)$ significantly outdoes KL-LUCB-E for every setting of m from 1 to 14. KL-LUCB-$log(t)$ also outperforms the SAR algorithm of [Bubeck et al., 2013b], yielding yet another result in favor of adaptive sampling. A *tuned version* of KL-LUCB-E (using an exploration rate of $\frac{n}{2\times 180}$) performs virtually identical to KL-LUCB-$log(t)$, and is not shown in the figure.

5.2.4 Proofs of the theorems of Section 5.2

Before giving the proofs of Theorems 5.1, 5.5 and 5.6, we introduce the following notation, already used in Chapter 1, that will be useful in the proof:

$$d^+(x,y) = d(x,y)\mathbb{1}_{(x<y)} \quad \text{and} \quad d^-(x,y) = d(x,y)\mathbb{1}_{(x>y)}. \tag{5.11}$$

Proof of Theorem 5.1. We first introduce the following lemma.

Lemma 5.7. KL-LUCB *and* KL-Racing *are such that* $\hat{S}_m \subseteq \mathcal{S}^*_{m,\epsilon}$ *on the event*

$$W = \bigcap_{t\in\mathbb{N}} \bigcap_{a\in\mathcal{S}^*_m} (U_a(t) > \mu_a) \bigcap_{b\in(\mathcal{S}^*_m)^c} (L_b(t) < \mu_b). \tag{5.12}$$

where U and L denote the generic confidence bounds used by these two algorithms.

Proof for Racing. If Racing is not correct, there exists some first round t on which either an arm in $(\mathcal{S}^*_{m,\epsilon})^c$ is selected (first situation), or an arm in $\mathcal{S}^*_m$ is dismissed (second situation). Before t, all

the arms in the set of selected arms $\mathcal{S}$ are in $\mathcal{S}^*_{m,\epsilon}$, and all the arms in set of discarded arms $\mathcal{D}$ are in $(\mathcal{S}^*)^c$. In the first situation, let b be the arm in $(\mathcal{S}^*_{m,\epsilon})^c$ selected : for all arms a in $J(t)^c$, one has $U_a(t) - L_b(t) < \epsilon$. Among these arms, at least one must be in $\mathcal{S}^*_m$. So there exists $a \in \mathcal{S}^*_m$ and $b \in (\mathcal{S}^*_\epsilon)^c$ such that $U_a(t) < L_b(t) + \epsilon$. The second situation leads to the same conclusion. Hence if the algorithm fails, the following event holds:

$$\begin{aligned}
&\bigcup_{t\in\mathbb{N}} (\exists a \in \mathcal{S}^*_m, \exists b \in (\mathcal{S}^*_{m,\epsilon})^c : \; U_a(t) - L_b(t) < \epsilon) \\
&\subset \bigcup_{t\in\mathbb{N}} \left(\exists a \in \mathcal{S}^*_m, \exists b \in (\mathcal{S}^*_{m,\epsilon})^c : \; (U_a(t) < \mu_a) \cup (L_b(t) > \mu_a - \epsilon > \mu_b)\right) \\
&\subset \bigcup_{t\in\mathbb{N}} \bigcup_{a\in\mathcal{S}^*_m} (U_a(t) < \mu_a) \bigcup_{b\in(\mathcal{S}^*_{m,\epsilon})^c} (L_b(t) > \mu_b) \subset W^c.
\end{aligned}$$

Proof for LUCB. If LUCB is not correct, there exists some stopping time τ, arm a in $\mathcal{S}^*_m$ and an arm b in $(\mathcal{S}^*_{m,\epsilon})^c$ such that $a \in J(\tau)$ and $b \in J(\tau)^c$. As the stopping condition holds, on has $U_a(t) - L_b(t) < \epsilon$. Using the same reasoning as above, if the algorithm fails, the following event holds:

$$\bigcup_{t\in\mathbb{N}} \bigcup_{a\in\mathcal{S}^*_m} (U_a(t) < \mu_a) \bigcup_{b\in(\mathcal{S}^*_{m,\epsilon})^c} (L_b(t) > \mu_b) \subset W^c.$$

□

The probability of error of both algorithms is upper bounded as

$$\mathbb{P}(\hat{S}_m \notin \mathcal{S}^*_{m,\epsilon}) \leq \sum_{a\in\mathcal{S}^*_m} \mathbb{P}\left(\exists t \in \mathbb{N}^* : u_a(t) < \mu_a\right) + \sum_{b\in(\mathcal{S}^*_m)^c} \mathbb{P}\left(\exists t \in \mathbb{N}^* : l_b(t) > \mu_b\right)$$

And for $a \in \mathcal{S}^*_m$

$$\begin{aligned}
\mathbb{P}\left(\exists t \in \mathbb{N}^* : u_a(t) < \mu_a\right) &= \mathbb{P}\left(\exists t \in \mathbb{N}^* : N_a(t) d(\hat{\mu}_a(t), \mu_a) \geq \beta(t,\delta), \hat{\mu}_a(t) < \mu_a\right) \\
&= \mathbb{P}\left(\exists t \in \mathbb{N}^* : N_a(t) d(\hat{\mu}_a(t), \mu_a) \geq \beta(N_a(t),\delta), \hat{\mu}_a(t) < \mu_a\right) \\
&\leq \mathbb{P}\left(\exists s \in \mathbb{N}^* : s d^+(\hat{\mu}_{a,s}, \mu_a) \geq \beta(s,\delta)\right) \\
&\leq \sum_{s=1}^{\infty} \mathbb{P}\left(s d^+(\hat{\mu}_{a,s}, \mu_a) \geq \beta(s,\delta)\right) \leq \sum_{s=1}^{\infty} \exp(-\beta(s,\delta))
\end{aligned}$$

from inequality (1.14). The probability $\mathbb{P}\left(\exists t \in \mathbb{N}^* : l_b(t) > \mu_b\right)$ is similarly upper bounded for $b \in (\mathcal{S}^*_m)^c$, thus we get

$$\mathbb{P}(\hat{S}_m \notin \mathcal{S}^*_{m,\epsilon}) \leq K \sum_{s=1}^{\infty} \exp(-\beta(s,\delta)) \leq \sum_{s=1}^{\infty} \frac{\delta}{C_1 t^\alpha} \leq \delta,$$

which proves Theorem 5.1.

Proof of Theorem 5.5. Let W be the event (5.12) defined in Lemma 5.7, on which the algorithm KL-Racing is correct. We upper bound *deterministically* the number of samples t_a from each arm a used by the algorithm assuming that this event hold.

On W, at every round of the algorithm the set $\mathcal{A}$ of accepted arms (resp. $\mathcal{D}$ of discarded arms) contains only arms from $\mathcal{S}^*_m$ (resp. $(\mathcal{S}^*_m)^c$). Let $a \in \mathcal{S}^*_m$. If $l_a(t)$ is larger than $u_b(t)$ for all b in $(\mathcal{S}^*_m)^c \cap \mathcal{R}$, as the set $(\mathcal{S}^*_m)^c \cap \mathcal{R}$ contains $K - m - |\mathcal{D}|$ arms, $l_a(t)$ is also larger that $u_b(t)$ for all b in $(J(t))^c$, and is thus accepted (and no longer drawn). Indeed, $(J(t))^c$ also contains $K - m - |\mathcal{D}|$ arms

and if any two arms still in the race are such that $\hat{\mu}_i(t) > \hat{\mu}_j(t)$, then $u_i(t) > u_j(t)$ (this holds because the arms have been drawn the same number of time). This shows that, for $a \in \mathcal{S}_m^*$,

$$t_a \leq \inf \left\{ t \in \mathbb{N} : \forall b \in (\mathcal{S}_m^*)^c \cap \mathcal{R},\ l_a(t) > u_b(t) \right\}.$$

Similarly, for $b \in (\mathcal{S}_m^*)^c$,

$$t_b \leq \inf \left\{ t \in \mathbb{N} : \forall a \in (\mathcal{S}_m^*) \cap \mathcal{R},\ l_a(t) > u_b(t) \right\}.$$

The following lemma then allows to further upper bound t_a and t_b.

Lemma 5.8. *Let $a \in \mathcal{S}_m^*$, $b \in (\mathcal{S}_m^*)^c$ and $T_{a,b}^*$ be the unique solution of*

$$td^{**}(\mu_a, \mu_b) = \beta(t, \delta).$$

On W, if $t \geq T_{a,b}^*$ and a and b are still in the race, then $l_a(t) > u_b(t)$.

Proof of Lemma 5.8. Let c be such that $d^*(c, \mu_a) = d^*(c, \mu_b) = d^{**}(\mu_a, \mu_b)$. One also introduce $\tilde{l}_a$ and $\tilde{u}_b$ such that

$$d^*(c, \mu_a) = d(\tilde{l}_a, \mu_a) = d(\tilde{l}_a, c) \quad \text{and} \quad d^*(c, \mu_b) = d(\tilde{u}_b, \mu_b) = d(\tilde{u}_b, c).$$

Assume that $t \geq T_{a,b}^*$ and W holds. We prove that $l_a(t) > c$ and $u_b(t) < c$, which leads to the result. We give here only the proof that $l_a(t) > c$, the proof that $u_b(t) < c$ is similar.

We start by showing that $\hat{\mu}_{a,t} > \tilde{l}_a$. As $\mu_a < u_a(t)$, one has $td^+(\hat{\mu}_{a,t}, \mu_a) \leq \beta(t, \delta)$. For $t \geq T_{a,b}^*$,

$$\beta(t, \delta) \leq td^{**}(\mu_a, \mu_b) = td(\tilde{l}_a, \mu_a) = td^+(\tilde{l}_a, \mu_a).$$

Hence, the facts that $d^+(\hat{\mu}_{a,t}, \mu_a) \leq d^+(\tilde{l}_a, \mu_a)$ and that $x \mapsto d^+(x, \mu_a)$ is non-increasing (cf. Figure 5.3) implies that $\hat{\mu}_{a,t} > \tilde{l}_a$.

By definition, $td^-(\hat{\mu}_{a,t}, l_a(t)) = \beta(t, \delta)$. On the one hand the fact that the mapping $x \mapsto d^-(x, l_a(t))$ is non-decreasing (Figure 5.3) and that, as proved above, $\hat{\mu}_{a,t} > \tilde{l}_a$, yields $td^-(\hat{\mu}_{a,t}, l_a(t)) > td^-(\tilde{l}_a, l_a(t))$. On the other hand, as $t \geq T_{a,b}^*$, one has

$$\beta(t, \delta) \leq td^{**}(\mu_a, \mu_b) = td(\tilde{l}_a, c) = td^-(\tilde{l}_a, c).$$

Putting everything together gives $d^-(\tilde{l}_a, l_a(t)) \leq d^-(\tilde{l}_a, c)$. This together with the fact that $x \mapsto d^-(\tilde{l}_a, x)$ is non-increasing (Figure 5.3) implies that $l_a(t) > c$.

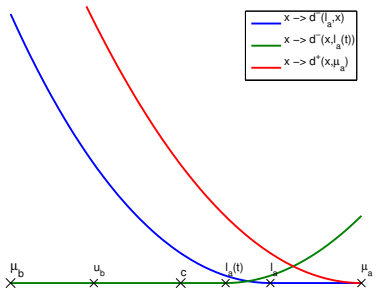

Figure 5.3: Functions based on KL-divergence used in the proof of Lemma 5.8

□

It follows from Lemma 5.8 that

$$\begin{array}{rcccl} a \in \mathcal{S}_m^* & \Rightarrow & t_a & \leq & \max_{b\in(\mathcal{S}_m^*)^c} T_{a,b}^* = T_{a,m+1}^* \\ b \in (\mathcal{S}_m^*)^c & \Rightarrow & t_b & \leq & \max_{a\in(\mathcal{S}_m^*)} T_{a,b}^* = T_{m,b}^* \end{array}$$

To conclude the proof, it remains to give an upper bound on $T_{a,m+1}^*$ and $T_{m,b}^*$, and Lemma 5.22, given in Appendix 5.6.1, permits to show that

$$T_{a,b}^* \leq \frac{\alpha}{d^{**}(\mu_a,\mu_b)}\left[\log\left(\frac{k_1 K}{\delta d^{**}((\mu_a,\mu_b))^\alpha}\right) + \log\log\left(\frac{k_1 K}{\delta d^{**}((\mu_a,\mu_b))^\alpha}\right) + 1\right]$$

which gives Theorem 5.5.

Proof of Theorem 5.6. At each round of the KL-LUCB algorithm, exactly two arms are drawn. The total number of samples used τ is therefore such that $\tau = 2\sigma$ where σ is the random number of rounds of the algorithm. Theorem 5.6 easily follows from these two inequalities that holds for any exploration rate:

$$\text{for } \alpha > 1, T \geq T_1^*,\ \mathbb{P}(\sigma \geq T) \leq H_{\epsilon,c}^* e^{-\beta(T,\delta)} + \sum_{t=1}^{\infty}(\beta(t,\delta)\log(t)+1)e^{-\beta(t,\delta)} \tag{5.13}$$

$$\text{for } \alpha > 2, T \geq T_2^*,\ \mathbb{P}(\sigma \geq T) \leq H_{\epsilon,c}^* e^{-\beta(T,\delta)} + \frac{KT}{2}(\beta(T,\delta)\log(T)+1)e^{-\beta(T/2,\delta)}, \tag{5.14}$$

with

$$T_1^* = \min\{T_0 : \forall T \geq T_0, H_{c,\epsilon}^*\beta(T,\delta) < T\} \quad \text{and} \quad T_2^* = \min\{T_0 : \forall T \geq T_0, 2H_{c,\epsilon}^*\beta(T,\delta) < T\},$$

and from upper bounds on T_1^* and T_2^* obtained from Lemma 5.22.

We now prove (5.14). For $c \in [\mu_{m+1}, \mu_m]$, if the algorithm hasn't stopped at time t, then one of the two intervals $\mathcal{I}_{u_t}(t)$ or $\mathcal{I}_{l_t}(t)$ is quite large and contains the parameter c. This simple idea is expressed in Proposition 5.7, which is proved below. To state it, we need to define the event

$$W_t = \bigcap_{a\in\mathcal{S}_m^*}(u_a(t) > \mu_a) \bigcap_{b\in(\mathcal{S}_m^*)^c}(l_b(t) < \mu_b).$$

Proposition 5.9. *If $U_{u_t} - L_{l_t} > \epsilon$ and W_t holds, then there exists $a \in \{l_t, u_t\}$ such that*

$$c \in \mathcal{I}_a(t) \ \text{ and } \ \tilde{\beta}_a(t) > \frac{\epsilon}{2},$$

where we define $\tilde{\beta}_a(t) := \sqrt{\frac{\beta(t,\delta)}{2N_a(t)}}$.

The remainder of this proof borrows from Lemma 5 of [Kalyanakrishnan et al., 2012]. Let T be some fixed time and σ the random number of rounds of the algorithm. Our goal is to find an event on

which $\min(\sigma, T) < T$; that is, the algorithm must have stopped after T rounds. Writing $\overline{T} = \lceil \frac{T}{2} \rceil$, we upper bound $\min(\sigma, T)$:

$$\begin{aligned}
\min(\sigma, T) &= \overline{T} + \sum_{t=\overline{T}}^{T} \mathbb{1}_{(\sigma \geq t)} = \overline{T} + \sum_{t=\overline{T}}^{T} \mathbb{1}_{(U_{u_t} - L_{l_t} > \epsilon)} \leq \overline{T} + \sum_{t=\overline{T}}^{T} \mathbb{1}_{(U_{u_t} - L_{l_t} > \epsilon)} \mathbb{1}_{W_t} + \sum_{t=\overline{T}}^{T} \mathbb{1}_{W_t^c} \\
&\leq \overline{T} + \sum_{t=\overline{T}}^{T} \mathbb{1}_{\left(\exists a \in \{u_t, l_t\} : c \in \mathcal{I}_a(t) \cap \tilde{\beta}_a(t) > \frac{\epsilon}{2}\right)} \mathbb{1}_{W_t} + \sum_{t=\overline{T}}^{T} \mathbb{1}_{W_t^c} \\
&\leq \overline{T} + \sum_{a=1}^{K} \underbrace{\sum_{t=\overline{T}}^{T} \mathbb{1}_{(a \in \{u_t, l_t\})} \mathbb{1}_{(c \in \mathcal{I}_a(t))} \mathbb{1}_{(\tilde{\beta}_a(t) > \frac{\epsilon}{2})}}_{(B_a)} + \sum_{t=\overline{T}}^{T} \mathbb{1}_{W_t^c}
\end{aligned}$$

where the first inequality comes from Proposition 5.7. Let $\mathcal{A}_\epsilon := \left\{a \in \{1, 2, \ldots, K\} : d^*(\mu_a, c) < \epsilon^2/2\right\}$. For $a \in \mathcal{A}_\epsilon$, the term (B_a) is upper bounded as

$$(B_a) \leq \sum_{t=\overline{T}}^{T} \mathbb{1}_{(a \in \{u_t, l_t\})} \mathbb{1}_{(\tilde{\beta}_a(t) > \frac{\epsilon}{2})} = \sum_{t=\overline{T}}^{T} \mathbb{1}_{(a \in \{u_t, l_t\})} \mathbb{1}_{\left(N_a(t) < \frac{\beta(t,\delta)}{\epsilon^2/2}\right)} \leq \frac{\beta(T, \delta)}{\epsilon^2/2}$$

whereas for $a \notin \mathcal{A}_\epsilon$, (B_a) is upper bounded as

$$\begin{aligned}
(B_a) &\leq \sum_{t=\overline{T}}^{T} \mathbb{1}_{(a \in \{u_t, l_t\})} \mathbb{1}_{(c \in \mathcal{I}_a(t))} \\
&\leq \sum_{t=\overline{T}}^{T} \mathbb{1}_{(a \in \{u_t, l_t\})} \mathbb{1}_{\left(N_a(t) \leq \frac{\beta(T,\delta)}{d^*(\mu_a, c)}\right)} + \sum_{t=\overline{T}}^{T} \mathbb{1}_{(a \in \{u_t, l_t\})} \mathbb{1}_{\left(N_a(t) > \frac{\beta(T,\delta)}{d^*(\mu_a, c)}\right)} \mathbb{1}_{(N_a(t) d(\hat{\mu}_a(t), c) \leq \beta(T, \delta))} \\
&\leq \frac{\beta(T, \delta)}{d^*(\mu_a, c)} + \sum_{s=\lceil \frac{\beta(T,\delta)}{d^*(\mu_a, c)} \rceil + 1}^{T} \mathbb{1}_{(s d(\hat{\mu}_{a,s}, c) \leq \beta(T, \delta))}
\end{aligned}$$

Let A_T and B_T be the two events

$$A_T = \bigcap_{a \in \mathcal{A}_\epsilon^c} \bigcap_{s \geq \lceil \frac{\beta(T,\delta)}{d^*(\mu_a, c)} \rceil + 1} (s d(\hat{\mu}_{a,s}, c) \geq \beta(T, \delta)) \quad \text{and} \quad B_T = \bigcap_{\overline{T}}^{T} W_t.$$

On $A_T \cap B_T$, one has $\min(\sigma, T) \leq \overline{T} + H_\epsilon^* \beta(T, \delta)$, thus $(\sigma \leq T)$ for all $T \geq T_2^*$, with

$$T_2^* = \min\{T_0 : \forall T \geq T_0, 2 H_{c,\epsilon}^* \beta(T, \delta) < T\}$$

Hence for $T \geq T_2^*$, $\mathbb{P}(\sigma \geq T) \leq \mathbb{P}(A_T^c) + \mathbb{P}(B_T^c)$. From Lemma 1.9,

$$\begin{aligned}
\mathbb{P}(W_t^c) &\leq \sum_{a \in \mathcal{S}_m^*} \mathbb{P}(\exists s \leq t : s d^+(\mu_{a,s}, \mu_a) \geq \beta(t, \delta)) + \sum_{a \notin \mathcal{S}_m^*} \mathbb{P}(\exists s \leq t : s d^-(\mu_{a,s}, \mu_a) \geq \beta(t, \delta)) \\
&\leq K e \beta(t, \delta) (\log t + 1) \exp(-\beta(t, \delta))
\end{aligned}$$

Using moreover Lemma 5.2, one obtains

$$\mathbb{P}(A_T^c) \leq \sum_{a \in \mathcal{A}_\epsilon^c} \sum_{s=\lceil \frac{\beta(T,\delta)}{d^*(\mu_a, c)} \rceil + 1}^{\infty} \mathbb{P}\left(s d(\hat{\mu}_{a,s}, c) \leq \beta(T, \delta)\right) \leq \sum_{a \in \mathcal{A}_\epsilon^c} \frac{e^{-\beta(T,\delta)}}{d^*(\mu_a, c)} \leq H_{\epsilon,c}^* \exp\left(-\beta(T, \delta)\right),$$

$$\mathbb{P}(B_T^c) \leq K \sum_{t=\overline{T}}^{T} e \beta(t, \delta) (\log t + 1) \exp(-\beta(t, \delta)) \leq K \frac{T}{2} \beta(T, \delta) (\log T + 1) \exp\left(-\beta\left(T/2, \delta\right)\right).$$

This proves (5.14). The proof of (5.13) follows along the same lines, except we do not introduce $\overline{T}$ and replace it by zero in the above equations. The introduction of $\overline{T}$ to show (5.13) is necessary to be able to upper-bound $\mathbb{P}(\sigma \geq T)$ by the general term of a convergent series. We now give the proof of Proposition 5.9 .

Proof of Proposition 5.9. We first show that at time t, if the stopping condition does not hold ($U_{u_t} - L_{l_t} > \epsilon$) and the event W_t holds, then either $c \in \mathcal{I}_{u_t}(t)$ or $c \in \mathcal{I}_{l_t}(t)$. This comes from a straightforward adaptation of the beginning of the proof of Lemma 2 from [Kalyanakrishnan et al., 2012]. Then we also observe that if $U_{u_t} - L_{l_t} > \epsilon$, the two intervals $\mathcal{I}_{u_t}(t)$ and $\mathcal{I}_{l_t}(t)$ cannot be too small simultaneously. Indeed, Pinsker's inequality (5.7) and the fact that $\hat{p}_{u_t}(t) < \hat{p}_{l_t}(t)$ leads to

$$\tilde{\beta}_{u_t}(t) + \tilde{\beta}_{l_t}(t) > \epsilon \quad \text{with} \quad \tilde{\beta}_a(t) := \sqrt{\frac{\beta(t,\delta)}{2N_a(t)}}. \tag{5.15}$$

Hence either $\tilde{\beta}_{u_t}(t) > \frac{\epsilon}{2}$ or $\tilde{\beta}_{u_t}(t) > \frac{\epsilon}{2}$. It remains to show that one of $k = l_t$ and $k = u_t$ such that $c \in \mathcal{I}_k(t)$ also satisfies this second condition. This part is the Proof uses properties of KL-divergence, and *cannot* directly be adapted from [Kalyanakrishnan et al., 2012].

It remains to show that if $U_{u_t}(t) - L_{l_t}(t) > \epsilon$, then the four statements below hold.

$$c \in \mathcal{I}_{u_t}(t) \text{ and } c > U_{l_t}(t) \quad \Rightarrow \quad \tilde{\beta}_{u_t}(t) > \frac{\epsilon}{2}. \tag{5.16}$$

$$c \in \mathcal{I}_{u_t}(t) \text{ and } c < L_{l_t}(t) \quad \Rightarrow \quad \tilde{\beta}_{u_t}(t) > \frac{\epsilon}{2}. \tag{5.17}$$

$$c \in \mathcal{I}_{l_t}(t) \text{ and } c > U_{u_t}(t) \quad \Rightarrow \quad \tilde{\beta}_{l_t}(t) > \frac{\epsilon}{2}. \tag{5.18}$$

$$c \in \mathcal{I}_{l_t}(t) \text{ and } c < L_{u_t}(t) \quad \Rightarrow \quad \tilde{\beta}_{l_t}(t) > \frac{\epsilon}{2}. \tag{5.19}$$

To prove (5.16), note that if $c \in \mathcal{I}_{u_t}(t)$ and $c > U_{l_t}(t)$, one has

$$d(\hat{p}_{u_t}(t), c) \leq 2\tilde{\beta}_{u_t}(t)^2 \quad \text{and} \quad d(\hat{p}_{l_t}(t), c) \geq 2\tilde{\beta}_{l_t}(t)^2.$$

Moreover, as $c > U_{l_t}$, $c > \hat{p}_{l_t}(t) > \hat{p}_{u_t}(t)$ holds, and therefore $d(\hat{p}_{l_t}(t), c) \leq d(\hat{p}_{u_t}(t), c)$. Hence,

$$2\tilde{\beta}_{l_t}(t)^2 \leq d(\hat{p}_{l_t}(t), c) \leq d(\hat{p}_{u_t}(t), c) \leq 2\tilde{\beta}_{u_t}(t)^2 \quad \text{and} \quad \tilde{\beta}_{l_t}(t) \leq \tilde{\beta}_{u_t}(t)$$

This together with $\tilde{\beta}_{l_t}(t) + \tilde{\beta}_{u_t}(t) > \epsilon$ leads to $\tilde{\beta}_{u_t}(t) > \frac{\epsilon}{2}$ and proves statement (5.16). The proof of statement (5.18) use identical arguments.

The proof of statement (5.17) goes as follows :

$$\begin{aligned}
&(U_{u_t}(t) - L_{l_t}(t) > \epsilon) \cap (L_{u_t} < c) \cap (c < L_{l_t}(t)) \\
&\Rightarrow (U_{u_t}(t) > c + \epsilon) \cap (L_{u_t} < c) \\
&\Rightarrow (\hat{p}_{u_t}(t) + \tilde{\beta}_{u_t}(t) > c + \epsilon) \cap (\hat{p}_{u_t}(t) - \tilde{\beta}_{u_t}(t) < c) \\
&\Rightarrow 2\tilde{\beta}_{u_t}(t) > \epsilon.
\end{aligned}$$

And the proof of statement (5.19) is similar.

5.3 Generic lower bound on the complexity in the fixed-confidence setting

The lower bound we provide here is not restricted to Bernoulli bandit models discussed in the previous section. Rather, we focus on *identifiable* classes of bandit models. A class $\mathcal{M}_m$ of bandit models is called identifiable is there exists a set et probability measures $\mathcal{P}$ satisfying

$$\forall p, q \in \mathcal{P},\ p \neq q \ \Rightarrow\ 0 < \mathrm{KL}(p, q) < +\infty,$$

such that for all $\nu = (\nu_1, \dots, \nu_K) \in \mathcal{M}_m$ and for $a \in \{1, \dots, K\}, \nu_a \in \mathcal{P}$.

All the lower bounds we propose in this chapter rely on the powerful technical Lemma 5.10, that was already introduced in Chapter 1 as Lemma 1.3 to prove the lower bound on the regret given in Theorem 1.2. This new, simple expression of a change of distribution was first presented in the paper [Kaufmann et al., 2014a].

Lemma 5.10. *Let ν and ν' be two bandit models. Let $\mathcal{F}_t = \sigma(A_1, Z_1, \dots, A_t, X_t)$ be the filtration associated to a sampling strategy (A_t). If σ is a stopping time with respect to $\mathcal{F}_t$, for any $A \in \mathcal{F}_\sigma$ such that $0 < \mathbb{P}_\nu(A) < 1$, one has*

$$\sum_{a=1}^{K} \mathbb{E}_\nu[N_a(\sigma)]\mathrm{KL}(\nu_a, \nu'_a) \geq d(\mathbb{P}_\nu(A), \mathbb{P}_{\nu'}(A)), \tag{5.20}$$

where $d(x, y) := \mathrm{KL}(\mathcal{B}(x), \mathcal{B}(y)) = x\log(x/y) + (1-x)\log((1-x)/(1-y))$.

We now propose a non asymptotic lower bound on the expected number of samples needed to identify the m best arms in the fixed-confidence setting, which straightforwardly yields a lower bound on $\kappa_C(\nu)$. Theorem 5.11 holds for an identifiable class of bandit models of the form:

$$\mathcal{M}_m = \{\nu = (\nu_1, \dots, \nu_K) : \nu_i \in \mathcal{P}, \mu_{[m]} > \mu_{[m+1]}\} \tag{5.21}$$

such that the set of probability measures $\mathcal{P}$ satisfies assumption 1 below.

Assumption 1. For all $\nu, \nu' \in \mathcal{P}^2$ such that $\nu \neq \nu'$, for all $\alpha > 0$,
there exists $\nu_1 \in \mathcal{P}$: $\mathrm{KL}(\nu, \nu') < \mathrm{KL}(\nu, \nu_1) < \mathrm{KL}(\nu, \nu') + \alpha$ and $\mathbb{E}_{X\sim\nu_1}[X] > \mathbb{E}_{X\sim\nu'}[X]$,
there exists $\nu_2 \in \mathcal{P}$: $\mathrm{KL}(\nu, \nu') < \mathrm{KL}(\nu, \nu_2) < \mathrm{KL}(\nu, \nu') + \alpha$ and $\mathbb{E}_{X\sim\nu_2}[X] < \mathbb{E}_{X\sim\nu'}[X]$.

These conditions are reminiscent of assumptions made by [Lai and Robbins, 1985]; they include simple classes of parametric bandits continuously parameterized by their means.

Theorem 5.11. *Let $\nu \in \mathcal{M}_m$, where $\mathcal{M}_m$ is defined by (5.21), and assume that $\mathcal{P}$ satisfies Assumption 1; any algorithm that is δ-PAC on $\mathcal{M}_m$ satisfies, for $\delta \leq 0.15$,*

$$\mathbb{E}_\nu[\tau] \geq \left[\sum_{a\in\mathcal{S}_m^*} \frac{1}{\mathrm{KL}(\nu_a, \nu_{[m+1]})} + \sum_{a\notin\mathcal{S}_m^*} \frac{1}{\mathrm{KL}(\nu_a, \nu_{[m]})}\right] \log\left(\frac{1}{2\delta}\right).$$

Proof Without loss of generality, one may assume that the arms are ordered such that $\mu_1 \geq \cdots \geq \mu_K$. Thus $\mathcal{S}_m^* = \{1, ..., m\}$. Let $\mathcal{A} = ((A_t), \tau, \hat{S}_m)$ be a δ-PAC algorithm and fix $\alpha > 0$. For all $a \in \{1, \dots, K\}$, from Assumption 1 there exists an alternative model

$$\nu' = (\nu_1, \dots, \nu_{a-1}, \nu'_a, \nu_{a+1}, \dots, \nu_K)$$

in which the only arm modified is arm a, and ν'_a is such that:

– $\mathrm{KL}(\nu_a, \nu_{m+1}) < \mathrm{KL}(\nu_a, \nu'_a) < \mathrm{KL}(\nu_a, \nu_{m+1}) + \alpha$ and $\mu'_a < \mu_{m+1}$ if $a \in \{1, \dots, m\}$,
– $\mathrm{KL}(\nu_a, \nu_m) < \mathrm{KL}(\nu_a, \nu'_a) < \mathrm{KL}(\nu_a, \nu_m) + \alpha$ and $\mu'_a > \mu_m$ if $a \in \{m+1, \dots, K\}$.

In particular, in the bandit model ν', the set of optimal arms is no longer $\{1, \dots, m\}$. Thus, introducing the event $A = (\hat{\mathcal{S}}_m = \{1, \dots, m\}) \in \mathcal{F}_\tau$, any δ-PAC algorithm satisfies $\mathbb{P}_\nu(A) \geq 1 - \delta$ and $\mathbb{P}_{\nu'}(A) \leq \delta$. Lemma 5.10 applied to the stopping time τ (such that $N_a(\tau) = N_a$ is the total number of draws of arm a) and the monotonicity properties of $d(x, y)$ ($x \mapsto d(x, y)$ is increasing when $x > y$ and decreasing when $x < y$) yield

$$\mathrm{KL}(\nu_a, \nu')\mathbb{E}_\nu[N_a] \geq d(1 - \delta, \delta).$$

From the definition of the alternative model, one obtains for $a \in \{1, \dots, m\}$ or $b \in \{m+1, \dots, K\}$ respectively, for every $\alpha > 0$,

$$\mathbb{E}_\nu[N_a] \geq \frac{d(1-\delta,\delta)}{\mathrm{KL}(\nu_a, \nu_{m+1}) + \alpha} \quad \text{and} \quad \mathbb{E}_\nu[N_b] \geq \frac{d(1-\delta,\delta)}{\mathrm{KL}(\nu_b, \nu_m) + \alpha}.$$

For $\delta \leq 0.15$, it can be shown that $d(1 - \delta, \delta) \geq \log(1/(2\delta))$. Thus, letting α tend to zero and summing over the arms leads to the lower bound on $\mathbb{E}_\nu[\tau] = \sum_{a=1}^K \mathbb{E}_\nu[N_a]$.

□

Remark 5.12. *Lemma 5.10 can also be used to improve the result of [Mannor and Tsitsiklis, 2004] that holds for $m = 1$ under the ϵ-relaxation described before. Combining the changes of distribution of this paper with Lemma 5.10 yields, for every $\epsilon > 0$ and $\delta \leq 0.15$,*

$$\mathbb{E}_\nu[\tau] \geq \left(\frac{|\{a : \mu_a \geq \mu_{[1]} - \epsilon\}| - 1}{\mathrm{KL}\left(\mathcal{B}(\mu_{[1]}), \mathcal{B}(\mu_{[1]} - \epsilon)\right)} + \sum_{a:\mu_a \leq \mu_{[1]} - \epsilon} \frac{1}{\mathrm{KL}\left(\mathcal{B}(\mu_a), \mathcal{B}(\mu_{[1]} + \epsilon)\right)} \right) \log \frac{1}{2\delta},$$

where $|\mathcal{X}|$ denotes the cardinal of the set $\mathcal{X}$ and $\mathcal{B}(\mu)$ the Bernoulli distribution of mean μ.

A class of exponential bandit models, such that

$$\mathcal{M}_m = \left\{ \nu = (\nu_{\theta_1}, \dots, \nu_{\theta_K}) : (\theta_1, \dots, \theta_K) \in \Theta^K, \theta_{[m]} > \theta_{[m+1]} \right\},$$

where ν_θ belongs to a canonical one-parameter exponential family and has a density with respect to some reference measure given by

$$f_\theta(x) = A(x) \exp(\theta x - b(\theta)), \quad \text{for } \theta \in \Theta \subset \mathbb{R} \tag{5.22}$$

is an example of class satisfying Assumption 1. Using the shorthand $\mathrm{K}(\theta, \theta') = \mathrm{KL}(\nu_\theta, \nu_{\theta'})$ for $(\theta, \theta') \in \Theta^2$, the lower bound of Theorem 5.11 together of the upper bound on $\mathbb{E}[\tau]$ for KL-LUCB (that can be generalized to exponential families) yield

$$\sum_{a=1}^m \frac{1}{\mathrm{K}(\theta_a, \theta_{m+1})} + \sum_{a=m}^K \frac{1}{\mathrm{K}(\theta_a, \theta_m)} \leq \kappa_C(\nu) \leq 8 \min_{\theta \in [\theta_{m+1}, \theta_m]} \sum_{a=1}^K \frac{1}{\mathrm{K}^*(\theta_a, \theta)}, \tag{5.23}$$

where $\mathrm{K}^*(\theta_1, \theta_2)$ is the Chernoff information between the distributions $\nu_{\theta_1}, \nu_{\theta_2}$, defined, as a function of the natural parameters, as

$$\mathrm{K}^*(\theta_1, \theta_2) = \mathrm{K}(\theta^*, \theta_1), \quad \text{where } \mathrm{K}(\theta^*, \theta_1) = \mathrm{K}(\theta^*, \theta_2).$$

A gap remains between the upper and lower bounds in (5.23), even when $K = 2$. For two armed-bandits we propose in the next Section a refined lower bounds on both $\kappa_C(\nu)$ and $\kappa_B(\nu)$ along with matching algorithms.

5.4 The complexity of A/B Testing

In this section, we present our contributions related to the complexity of best arm identification in two-armed bandit models, following closely the paper [Kaufmann et al., 2014a]. Two armed-bandits are of particular interest as they offer a theoretical framework for sequential A/B Testing. A/B Testing is a popular procedure used, for instance, for website optimization: two versions of a webpage, say A and B, are empirically compared by being presented to users. Each user is shown only one version $A_t \in \{1, 2\}$ and provides a real-valued index of the quality of the page, X_t, which is modeled as a sample of a probability distribution ν_1 or ν_2. For example, a standard objective is to determine which webpage has the highest conversion rate (probability that a user actually becomes a customer) by receiving binary feedback from the users.

In standard A/B Testing algorithms, the two versions are presented equally often. It is thus of particular interest to investigate whether an algorithm using a (pure) *uniform sampling strategy*, such that the arms are sampled in a round-robin fashion, can be efficient in two-armed bandit model. We saw in Section 5.2 that when there are more than $K > 2$ uniform sampling in the above sense is not desirable (it should at least be coupled with eliminations). However, when there are two arms, KL-LUCB and KL-Racing reduce to the same algorithm, sampling both arms at each round. An algorithm using uniform sampling can be regarded as a statistical test of the hypothesis $H_0 : (\mu_1 \leq \mu_2)$ against $H_1 : (\mu_1 > \mu_2)$ based on paired samples (X_s, Y_s) of ν_1, ν_2; namely a test based on a fixed number of samples in the fixed-budget setting, and, a *sequential test* in the fixed-confidence setting, in which a randomized stopping rule determines when the experiment is to be terminated.

In two-armed bandit models, classical sequential testing theory provides a first element of comparison between the fixed-budget and fixed-confidence settings, in the simpler case of fully specified alternatives. Consider for instance the case where ν_1 and ν_2 are Gaussian laws with the same known variance σ^2, the means μ_1 and μ_2 known up to a permutation. Denoting by P the joint distribution of the paired samples (X_s, Y_s), one must choose between the hypotheses $H_0 : P = \mathcal{N}\left(\mu_1, \sigma^2\right) \otimes \mathcal{N}\left(\mu_2, \sigma^2\right)$ and $H_1 : P = \mathcal{N}\left(\mu_2, \sigma^2\right) \otimes \mathcal{N}\left(\mu_1, \sigma^2\right)$. It is known since [Wald, 1945] that among the sequential tests such that type I and type II error probabilities are both smaller than δ, the Sequential Probability Ratio Test (SPRT) minimizes the expected number of required samples, and is such that $\mathbb{E}_\nu[\tau] = 2\sigma^2/(\mu_1 - \mu_2)^2 \log(1/\delta)$. However, the batch test that minimizes both probabilities of error is the Likelihood Ratio test; it can be shown to require a sample size of order $8\sigma^2/(\mu_1 - \mu_2)^2 \log(1/\delta)$ in order to ensure that both type I and type II error probabilities are smaller than δ. Thus, when the sampling strategy is uniform and the parameters are known, there is a clear gain in using randomized stopping strategies. We will show below that this conclusion is not valid anymore when the values of μ_1 and μ_2 are not assumed to be known. Indeed, for two-armed Gaussian bandit models we show that $\kappa_B(\nu) = \kappa_C(\nu)$ and for two-armed Bernoulli bandit models we show that $\kappa_C(\nu) > \kappa_B(\nu)$.

To prove this, we start by giving in Section 5.4.1 a refined lower bound on $\kappa_C(\nu)$, based on a different change of distribution, as well as a lower bound on $\kappa_B(\nu)$. We then provide in two particular cases, Gaussian bandits with known variances (Section 5.4.2) and Bernoulli bandits (Section 5.4.3) efficient algorithms (almost) matching these bounds. In particular, we show that for Bernoulli bandits only little can be gained by departing from uniform sampling, and propose an algorithm for the fixed-confidence setting based on a non-trivial stopping criterion that is reminiscent of KL-LUCB.

5.4.1 Lower bounds on the two complexities

For $\mathcal{M}_1 = \mathcal{M}$ an identifiable class of two-armed bandit models, Theorem 5.13 provides lower bounds on $\kappa_B(\nu)$ and $\kappa_C(\nu)$ for every $\nu \in \mathcal{M}$.

Theorem 5.13. *Let $\nu = (\nu_1, \nu_2)$ be a two-armed bandit model such that $\mu_1 > \mu_2$. In the fixed-budget setting, any consistent algorithm satisfies*

$$\limsup_{t\to\infty} -\frac{1}{t}\log p_t(\nu) \leq c^*(\nu), \quad \textit{where} \quad c^*(\nu) := \inf_{(\nu_1',\nu_2')\in\mathcal{M}:\mu_1'<\mu_2'} \max\left\{\mathrm{KL}(\nu_1',\nu_1), \mathrm{KL}(\nu_2',\nu_2)\right\}.$$

In the fixed-confidence setting any algorithm that is δ-PAC on $\mathcal{M}$ satisfies, when $\delta \leq 0.15$,

$$\mathbb{E}_\nu[\tau] \geq \frac{1}{c_*(\nu)}\log\left(\frac{1}{2\delta}\right), \quad \textit{where} \quad c_*(\nu) := \inf_{(\nu_1',\nu_2')\in\mathcal{M}:\mu_1'<\mu_2'} \max\left\{\mathrm{KL}(\nu_1,\nu_1'), \mathrm{KL}(\nu_2,\nu_2')\right\}.$$

In particular, Theorem 5.13 implies that $\kappa_B(\nu) \geq 1/c^*(\nu)$ and $\kappa_C(\nu) \geq 1/c_*(\nu)$. Proceeding similarly, one can obtain lower bounds for the algorithms that use uniform sampling of both arms. The proof of both results is provided below.

Theorem 5.14. *Let $\nu = (\nu_1, \nu_2)$ be a two-armed bandit model such that $\mu_1 > \mu_2$. In the fixed-budget setting, any consistent algorithm using a uniform sampling strategy satisfies*

$$\limsup_{t\to\infty} -\frac{1}{t}\log p_t(\nu) \leq I^*(\nu) \quad \textit{where} \quad I^*(\nu) := \inf_{(\nu_1',\nu_2')\in\mathcal{M}:\mu_1'<\mu_2'} \frac{\mathrm{KL}(\nu_1',\nu_1) + \mathrm{KL}(\nu_2',\nu_2)}{2}.$$

In the fixed-confidence setting, any algorithm that is δ-PAC on $\mathcal{M}$ and uses a uniform sampling strategy satisfies, for $\delta \leq 0.15$,

$$\mathbb{E}_\nu[\tau] \geq \frac{1}{I_*(\nu)}\log\frac{1}{2\delta} \quad \textit{where} \quad I_*(\nu) := \inf_{(\nu_1',\nu_2')\in\mathcal{M}:\mu_1'<\mu_2'} \frac{\mathrm{KL}(\nu_1,\nu_1') + \mathrm{KL}(\nu_2,\nu_2')}{2}.$$

Obviously, one always has $I^*(\nu) \leq c^*(\nu)$ and $I_*(\nu) \leq c_*(\nu)$ suggesting that uniform sampling can be sub-optimal. It is possible to give explicit expressions for the quantities $c^*(\nu), c_*(\nu)$ and $I^*(\nu), I_*(\nu)$ for specific classes of parametric bandit models that will be studied in the next Sections. In the case of two-armed Gaussian bandits with known variance (see Section 5.4.2):

$$\mathcal{M} = \{\nu = \left(\mathcal{N}\left(\mu_1, \sigma_1^2\right), \mathcal{N}\left(\mu_2, \sigma_2^2\right)\right) : (\mu_1, \mu_2) \in \mathbb{R}^2, \mu_1 \neq \mu_2\}, \tag{5.24}$$

using that

$$\mathrm{KL}(\mathcal{N}(\mu_1, \sigma_1), \mathcal{N}(\mu_2, \sigma_2)) = \frac{(\mu_1-\mu_2)^2}{2\sigma_2^2} + \frac{1}{2}\left[\frac{\sigma_1^2}{\sigma_2^2} - 1 - \log\frac{\sigma_1^2}{\sigma_2^2}\right], \tag{5.25}$$

one obtains

$$c^*(\nu) = c_*(\nu) = \frac{(\mu_1-\mu_2)^2}{2(\sigma_1+\sigma_2)^2} \quad \text{and} \quad I^*(\nu) = I_*(\nu) = \frac{(\mu_1-\mu_2)^2}{4(\sigma_1^2+\sigma_2^2)}.$$

Hence, the lower bounds of Theorem 5.13 are equal in this case, and we provide in Section 5.4.2 matching upper bounds confirming that indeed $\kappa_B(\nu) = \kappa_C(\nu)$. In addition, the observation that, when the

variances are different $c_*(\nu) > I_*(\nu)$, will be shown to imply that strategies based on uniform sampling are sub-optimal.

The values of $c^*(\nu)$ and $c_*(\nu)$ can also be computed in the class of two-armed exponential bandit models,

$$\mathcal{M} = \{\nu = (\nu_{\theta_1}, \nu_{\theta_2}) : (\theta_1, \theta_2) \in \Theta^2, \theta_1 \neq \theta_2\}$$

where ν_{θ_a} has density f_{θ_a} given by (5.22). One can show that

$$\begin{array}{rcccl} c^*(\nu) & = & \inf_{\theta\in\Theta} \max\left(\mathrm{K}(\theta,\theta_1), \mathrm{K}(\theta,\theta_2)\right) & = & \mathrm{K}(\theta^*,\theta_1), \quad \text{where } \mathrm{K}(\theta^*,\theta_1) = \mathrm{K}(\theta^*,\theta_2), \\ c_*(\nu) & = & \inf_{\theta\in\Theta} \max\left(\mathrm{K}(\theta_1,\theta), \mathrm{K}(\theta_2,\theta)\right) & = & \mathrm{K}(\theta_1,\theta_*), \quad \text{where } \mathrm{K}(\theta_1,\theta_*) = \mathrm{K}(\theta_2,\theta_*). \end{array}$$

The coefficient $c^*(\nu)$ is equal to the Chernoff information $\mathrm{K}^*(\theta_1, \theta_2)$ between the arms, already introduced in Section 5.2, whereas $c_*(\nu)$ corresponds to a quantity close to the Chernoff information but with 'reversed' roles for the arguments. By analogy, we denote this quantity by $\mathrm{K}_*(\theta_1, \theta_2) = \mathrm{K}(\theta_1, \theta_*)$.

For exponential bandits the quantities $c^*(\nu)$ and $c_*(\nu)$ are not equal in general, although it can be shown that it is the case when the log-partition function $b(\theta)$ is (Fenchel) self-conjugate (e.g., for Gaussian and exponential variables). In Section 5.4.3, we will focus on the case of Bernoulli models for which $c^*(\nu) > c_*(\nu)$. By exhibiting a matching strategy in the fixed-budget setting, we will show that this implies that $\kappa_C(\nu) > \kappa_B(\nu)$ in this case.

On the changes of distribution used. Theorem 5.11 applied to a two-armed exponential bandit model $\nu = (\nu_{\theta_1}, \nu_{\theta_2})$ yields

$$\kappa_C(\nu) \geq \left(\frac{1}{\mathrm{K}(\theta_1,\theta_2)} + \frac{1}{\mathrm{K}(\theta_2,\theta_1)}\right), \tag{5.26}$$

while the lower bound given in Theorem 5.13 is

$$\kappa_C(\nu) \geq \left(\frac{1}{\mathrm{K}_*(\theta_1,\theta_2)}\right). \tag{5.27}$$

which can be shown to be always tighter than (5.26).

Interestingly, the changes of distribution used to derive the two results are not the same. On the one hand, for inequality (5.26), the changes of distribution involved modify a single arm at a time: one of the arms is moved just below (or just above) the other (see Figure 5.4, left). This is the idea also used, for example, to obtain the lower bound of [Lai and Robbins, 1985] on the cumulative regret (see the proof of Theorem 1.2 in Chapter 1). On the other hand, for inequality (5.27), both arms are modified at the same time: they are moved close to the common intermediate value θ_* but with a reversed ordering (see Figure 5.4, right). In the fixed-budget setting, the changes of distribution used to obtain the lower bound on $\kappa_B(\nu)$ that follow from Theorem 5.13 moves both arms close to the value θ_*.

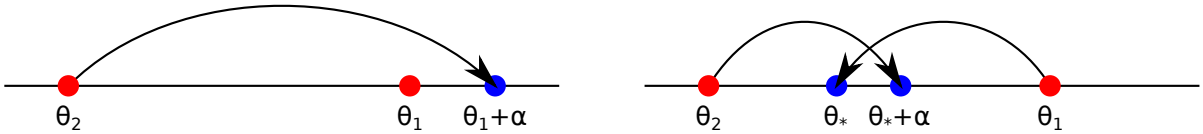

Figure 5.4: Alternative bandit models considered to obtain the lower bounds of Theorem 5.11 (left) and Theorem 5.13 (right), in the fixed-confidence setting.

Proof of Theorem 5.13 and 5.14. Without loss of generality, assume that the bandit model $\nu = (\nu_1, \nu_2)$ is such that $a^* = 1$. Consider any alternative bandit model $\nu' = (\nu_1', \nu_2')$ in which $a^* = 2$. For a given strategy $\mathcal{A}$, let A be the event $A = (\hat{S}_1 = 1)$, which belongs to $\mathcal{F}_\tau$, for τ the stopping rule of $\mathcal{A}$.

Fixed-budget setting Assume the strategy $\mathcal{A}$ is consistent. For every $t \in \mathbb{N}$, if the budget is $\tau = t$ a.s., Lemma 5.10 applied to the stopping time $\sigma = t$ and the event $A = (\hat{S}_1 = 1) \in \mathcal{F}_t$ defined above yields

$$\mathbb{E}_{\nu'}[N_1(t)]\mathrm{KL}(\nu_1', \nu_1) + \mathbb{E}_{\nu'}[N_2(t)]\mathrm{KL}(\nu_2', \nu_2) \geq d(\mathbb{P}_{\nu'}(A), \mathbb{P}_\nu(A)).$$

One has $p_t(\nu) = 1 - \mathbb{P}_\nu(A)$ and $p_t(\nu') = \mathbb{P}_{\nu'}(A)$. As $\mathcal{A}$ is consistent, for every $\epsilon > 0$ there exists $t_0(\epsilon)$ such that for all $t \geq t_0(\epsilon)$, $\mathbb{P}_{\nu'}(A) \leq \epsilon \leq \mathbb{P}_\nu(A)$. For $t \geq t_0(\epsilon)$,

$$\mathbb{E}_{\nu'}[N_1(t)]\mathrm{KL}(\nu_1', \nu_1) + \mathbb{E}_{\nu'}[N_2(t)]\mathrm{KL}(\nu_2', \nu_2) \geq d(\epsilon, 1 - p_t(\nu)) \geq (1-\epsilon)\log\frac{1-\epsilon}{p_t(\nu)} + \epsilon\log\epsilon.$$

Taking the limsup and letting ϵ go to zero, one can show that

$$\limsup_{t\to\infty} -\frac{1}{t}\log p_t(\nu) \leq \limsup_{t\to\infty}\left(\frac{\mathbb{E}_{\nu'}[N_1(t)]}{t}\mathrm{KL}(\nu_1', \nu_1) + \frac{\mathbb{E}_{\nu'}[N_2(t)]}{t}\mathrm{KL}(\nu_2', \nu_2)\right) \leq \max_{a=1,2}\mathrm{KL}(\nu_a', \nu_a).$$

The first statement of Theorem 5.13 follows by optimizing over the possible model ν' satisfying $\mu_1' < \mu_2'$ to make the right hand side of the inequality as small as possible.

If the sampling strategy of $\mathcal{A}$ is uniform, using that $\mathbb{E}_\nu[N_1] = \mathbb{E}[N_2] = \mathbb{E}[\tau]/2$, $\limsup -\frac{1}{t}\log p_t(\nu)$ is upper bounded by $(\mathrm{KL}(\nu_1', \nu_1) + \mathrm{KL}(\nu_2', \nu_2))/2$, and similarly optimizing over the choice of ν' gives the first statement of Theorem 5.14.

Fixed-confidence setting Assume the strategy $\mathcal{A}$ is δ-PAC. It therefore satisfies $\mathbb{P}_\nu(A) \geq 1 - \delta$ and $\mathbb{P}_{\nu'}(A) \leq \delta$. Applying Lemma 5.10 (with the stopping rule τ) and using again the monotonicity properties of $d(x, y)$, one obtains that

$$\mathbb{E}_\nu[N_1]\mathrm{KL}(\nu_1, \nu_1') + \mathbb{E}_\nu[N_2]\mathrm{KL}(\nu_2, \nu_2') \geq d(\delta, 1-\delta). \tag{5.28}$$

For $\delta \leq 0.15$, as already used in the proof of Theorem 5.11, one has $d(\delta, 1-\delta) \geq \log(1/(2\delta))$. Using moreover that $\tau = N_1 + N_2$, one has

$$\mathbb{E}_\nu[\tau] \geq \frac{1}{\max_a \mathrm{KL}(\nu_a, \nu_a')}\log\left(\frac{1}{2\delta}\right).$$

The second statement of Theorem 5.13 follows by optimizing over the possible model ν' satisfying $\mu_1' < \mu_2'$ to make the right hand side of the inequality as large as possible.

If $\mathcal{A}$ uses a uniform sampling strategy, using the fact that $\mathbb{E}_\nu[N_1] = \mathbb{E}[N_2] = \mathbb{E}[\tau]/2$ in Equation (5.28) similarly gives the second statement of Theorem 5.14.

5.4.2 The Gaussian Case

We study in this Section the class of two-armed Gaussian bandit models with known variances defined by (5.24), where σ_1 and σ_2 are fixed. In this case, we observed above that the lower bounds of Theorem 5.13 are similar, because $c^*(\nu) = c_*(\nu)$. We prove in this section that indeed

$$\kappa_C(\nu) = \kappa_B(\nu) = \frac{2(\sigma_1 + \sigma_2)^2}{(\mu_1 - \mu_2)^2}$$

by exhibiting strategies that reach these performance bounds. These strategies are based on the simple recommendation of the empirical best arm but use non-uniform sampling in cases where σ_1 and σ_2 differ. When $\sigma_1 = \sigma_2$ we provide in Theorem 5.15 an improved stopping rule that is δ-PAC but results in a significant reduction of the running time of fixed-confidence algorithms.

FIXED-BUDGET SETTING

We consider the simple family of *static strategies* that draw n_1 samples from arm 1 followed by $n_2 = t - n_1$ samples of arm 2, and then choose arm 1 if $\hat{\mu}_{1,n_1} < \hat{\mu}_{2,n_2}$, where $\hat{\mu}_{i,n_i}$ denotes the empirical mean of the n_i samples from arm i. Assume for instance that $\mu_1 > \mu_2$. Since $\hat{\mu}_{1,n_1} - \hat{\mu}_{2,n_2} - \mu_1 + \mu_2 \sim \mathcal{N}\left(0, \sigma_1^2/n_1 + \sigma_2^2/n_2\right)$, the probability of error of such a strategy is easily upper bounded as:

$$\mathbb{P}\left(\hat{\mu}_{1,n_1} < \hat{\mu}_{2,n_2}\right) \quad \leq \quad \exp\left(-\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^{-1} \frac{(\mu_1 - \mu_2)^2}{2}\right).$$

The right hand side is minimized when $n_1/(n_1 + n_2) = \sigma_1/(\sigma_1 + \sigma_2)$, and the static strategy drawing $n_1 = \lceil \sigma_1 t/(\sigma_1 + \sigma_2) \rceil$ times arm 1 is such that

$$\liminf_{t\to\infty} -\frac{1}{t} \log p_t(\nu) \geq \frac{(\mu_1 - \mu_2)^2}{2(\sigma_1 + \sigma_2)^2},$$

which matches the bound of Theorem 5.13 for Gaussian bandit models.

FIXED-CONFIDENCE SETTING

Equal Variances. We start with the simpler case $\sigma_1 = \sigma_2 = \sigma$. Thus, the quantity $I_*(\nu)$ introduced in Theorem 5.14 coincides with $c_*(\nu)$, which suggests that uniform sampling could be optimal. A uniform sampling strategy equivalently collects paired samples (X_s, Y_s) from both arms. The difference $X_s - Y_s$ is normally distributed with mean $\mu = \mu_1 - \mu_2$ and a δ-PAC algorithm is equivalent to a sequential test of $H_0 : (\mu < 0)$ versus $H_1 : (\mu > 0)$ such that both type I and type II error probabilities are bounded by δ. [Robbins, 1970] proposes the stopping rule

$$\tau = \inf\left\{t \in 2\mathbb{N}^* : \left|\sum_{s=1}^{t/2}(X_s - Y_s)\right| > \sqrt{2\sigma^2 t \beta(t,\delta)}\right\}, \quad \text{with} \quad \beta(t,\delta) = \frac{t+1}{t}\log\left(\frac{t+1}{2\delta}\right). \tag{5.29}$$

The recommendation rule chooses the empirically best arm at time τ. This procedure can be seen as an *elimination strategy*, in the sense of [Jennison et al., 1982]. The authors of this paper derive a lower bound on the sample complexity of any δ-PAC *elimination* strategy (whereas our lower bound applies to *any* δ-PAC algorithm) which is matched by Robbins' algorithm: the above stopping rule τ satisfies

$$\lim_{\delta\to 0} \frac{\mathbb{E}_\nu[\tau]}{\log(1/\delta)} = \frac{8\sigma^2}{(\mu_1 - \mu_2)^2}.$$

This value coincide with the lower bound on $\kappa_C(\nu)$ of Theorem 5.13 in the case of two-armed Gaussian distributions with similar known variance σ^2. This proves that in this case, Robbins' rule (5.29) is not only optimal among the class of elimination strategies, but also among the class of δ-PAC algorithm.

Any δ-PAC elimination strategy that uses a threshold function (or *exploration rate*) $\beta(t,\delta)$ smaller than Robbins' also matches our asymptotic lower bound, while stopping earlier than the latter. From a practical point of view, it is therefore interesting to exhibit smaller exploration rates that preserve the δ-PAC property. The failure probability of such an algorithm is upper bounded, for example when $\mu_1 < \mu_2$, by

$$\mathbb{P}_\nu\left(\exists k \in \mathbb{N} : \sum_{s=1}^{k} \frac{X_s - Y_s - (\mu_1 - \mu_2)}{\sqrt{2\sigma^2}} > \sqrt{2k\beta(2k,\delta)}\right) = \mathbb{P}\left(\exists k \in \mathbb{N} : S_k > \sqrt{2k\beta(2k,\delta)}\right) \tag{5.30}$$

where S_k is a sum of k i.i.d. variables of distribution $\mathcal{N}(0,1)$. [Robbins, 1970] obtains a non-explicit confidence region of risk at most δ by choosing $\beta(2k,\delta) = \log(\log(k)/\delta) + o(\log\log(k))$. The dependency in k is in some sense optimal, because the Law of Iterated Logarithm (LIL) states that $\limsup_{k\to\infty} S_k/\sqrt{2k\log\log(k)} = 1$ almost surely. Recently, [Jamieson et al., 2014] proposed an explicit confidence region inspired by the LIL. However, Lemma 1 of [Jamieson et al., 2014] cannot be used to upper bound (5.30) by δ and we provide in Appendix A a result derived independently (Lemma A.1 therein) that achieves this goal and can be used to obtain the following result (proved in Section 5.4.5).

Theorem 5.15. *For δ small enough, the elimination strategy with threshold $g(t,\delta) = \sqrt{2\sigma^2 t\beta(t,\delta)}$ is δ-PAC with*

$$\beta(t,\delta) = \log\frac{1}{\delta} + \frac{3}{4}\log\log\frac{1}{\delta} + \frac{3}{2}\log(1+\log(t/2)). \tag{5.31}$$

We refer to Section 5.4.4 for numerical simulations that illustrate the significant savings (in the average number of samples needed to reaching a decision) resulting from the use of the less conservative exploration rate allowed by Theorem 5.15.

Mismatched Variances. In the case where $\sigma_1 \neq \sigma_2$, we rely on the α-Elimination strategy, described in Algorithm 5 below. For $a = 1,2$, $\hat{\mu}_a(t)$ denotes the empirical mean of the samples gathered from arm a up to time t. The algorithm is based on a non-uniform sampling strategy governed by the parameter $\alpha \in (0,1)$ which ensures that, at the end of every round t, $N_1(t) = \lceil \alpha t \rceil$, $N_2(t) = t - \lceil \alpha t \rceil$ and $\hat{\mu}_1(t) - \hat{\mu}_2(t) \sim \mathcal{N}\left(\mu_1 - \mu_2, \sigma_t^2(\alpha)\right)$ (where $\sigma_t^2(\alpha)$ is defined at line 6 of Algorithm 5). The sampling schedule used here is thus deterministic.

Algorithm 5 α-Elimination

Require: Exploration function $\beta(t,\delta)$, parameter α.

1: *Initialization*: $\hat{\mu}_1(0) = \hat{\mu}_2(0) = 0$, $\sigma_0^2(\alpha) = 1$, $t = 0$
2: **while** $|\hat{\mu}_1(t) - \hat{\mu}_2(t)| \leq \sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)}$ **do**
3: $\quad t \leftarrow t+1$.
4: $\quad$ If $\lceil \alpha t \rceil = \lceil \alpha(t-1) \rceil$, $A_t \leftarrow 2$, else $A_t \leftarrow 1$
5: $\quad$ Observe $X_t \sim \nu_{A_t}$ and compute the empirical means $\hat{\mu}_1(t)$ and $\hat{\mu}_2(t)$
6: $\quad$ Compute $\sigma_t^2(\alpha) = \sigma_1^2/\lceil \alpha t \rceil + \sigma_2^2/(t - \lceil \alpha t \rceil)$
7: **end while**
8: **return** $\operatorname{argmax}_{a=1,2} \hat{\mu}_a(t)$

Theorem 5.16 shows that the $\sigma_1/(\sigma_1+\sigma_2)$-elimination algorithm, with a suitable exploration rate, is δ-PAC and matches the lower bound on $\mathbb{E}_\nu[\tau]$, at least asymptotically when $\delta \to 0$. Its proof can be found in Section 5.4.5.

Theorem 5.16. *If $\alpha = \sigma_1/(\sigma_1+\sigma_2)$, the α-elimination strategy using the exploration rate $\beta(t,\delta) = \log\frac{t}{\delta} + 2\log\log(6t)$ is δ-PAC on $\mathcal{M}$ and satisfies, for every $\nu \in \mathcal{M}$, for every $\epsilon > 0$,*

$$\mathbb{E}_\nu[\tau] \leq (1+\epsilon)\frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}\log\left(\frac{1}{\delta}\right) + \underset{\delta\to 0}{o_\epsilon}\left(\log\left(\frac{1}{\delta}\right)\right).$$

Remark 5.17. *When $\sigma_1 = \sigma_2$, $1/2$-elimination reduces, up to rounding effects, to the elimination procedure described in the previous paragraph, for which Theorem 5.15 suggests an exploration rate of order*

$\log(\log(t)/\delta)$. *As the feasibility of this exploration rate when $\sigma_1 \neq \sigma_2$ is yet to be established, we focus on Gaussian bandits with equal variances in the numerical experiments of Section 5.4.4.*

5.4.3 The Bernoulli Case

We consider in this section the class of Bernoulli bandit models defined by

$$\mathcal{M} = \{\nu = (\mathcal{B}(\mu_1), \mathcal{B}(\mu_2)) : (\mu_1, \mu_2) \in]0;1[^2, \mu_1 \neq \mu_2\},$$

where each arm can be equivalently parameterized by the natural parameter of the exponential family, $\theta_a = \log(\mu_a/(1-\mu_a))$. Recall the Kullback-Leibler divergence between two Bernoulli distributions can be indifferently considered as a function of the means, $d(\mu_1, \mu_2) = \mathrm{KL}(\mathcal{B}(\mu_1), \mathcal{B}(\mu_2))$, or of the natural parameters, $\mathrm{K}(\theta_1, \theta_2)$.

In this Section, we prove that $\kappa_C(\nu) > \kappa_B(\nu)$ for Bernoulli bandit models (Proposition 5.19). To do so, we first introduce a static strategy matching the lower bound of Theorem 5.13 in the fixed-budget setting (Proposition 5.18). This strategy is reminiscent of the algorithm exhibited for Gaussian bandits in Section 5.4.2 and uses parameter-dependent non uniform sampling. This strategy is not directly helpful in practice but we show that it can be closely approximated by an algorithm using uniform sampling. In the fixed-confidence setting we similarly conjecture that little can be gained from using a non-uniform sampling strategy and propose an algorithm based on a non-trivial stopping strategy that is believed to match the bound of Theorem 5.14.

FIXED-BUDGET SETTING

By carefully upper bounding the probability of error of a static strategy in the Bernoulli case, one can show the following result, proved in Section 5.6.3.

Proposition 5.18. *Let $\alpha(\theta_1, \theta_2)$ be defined by*

$$\alpha(\theta_1, \theta_2) = \frac{\theta^* - \theta_1}{\theta_2 - \theta_1} \quad \textit{where} \quad \mathrm{K}(\theta^*, \theta_1) = \mathrm{K}(\theta^*, \theta_2).$$

For all t, the static strategy that allocates $\lceil \alpha(\theta_1, \theta_2) t \rceil$ samples to arm 1, and recommends the empirical best arm, satisfies $p_t(\nu) \leq \exp(-tK^*(\theta_1, \theta_2))$.

This shows in particular that for every $\nu \in \mathcal{M}$ there exists a consistent static strategy such that

$$\liminf_{t\to\infty} -\frac{1}{t} \log p_t \geq \mathrm{K}^*(\theta_1, \theta_2), \quad \text{and hence that} \quad \kappa_B(\nu) = \frac{1}{\mathrm{K}^*(\theta_1, \theta_2)}.$$

By combining this observation with Theorem 5.13 and the fact that for Bernoulli distributions it can be shown that $\mathrm{K}_*(\theta_1, \theta_2) < \mathrm{K}^*(\theta_1, \theta_2)$, one obtains the following inequality.

Proposition 5.19. *For all $\nu \in \mathcal{M}$, $\kappa_C(\nu) > \kappa_B(\nu)$.*

Note that we have determined the complexity of the fixed-budget setting by exhibiting and algorithm that is of limited practical interest for Bernoulli bandit models. Indeed, the optimal static strategy defined in Proposition 5.18 requires the knowledge of the quantity $\alpha(\theta_1, \theta_2)$, that depends in general on the unknown means of the arms. So far, it is not known whether there exists a *universal* strategy, that would satisfy $p_t(\nu) \leq \exp(-\mathrm{K}^*(\theta_1, \theta_2)t)$ on *every* Bernoulli bandit model.

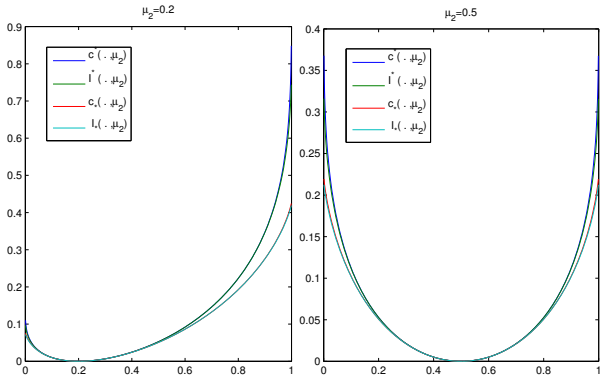

Figure 5.5: Comparison of different informational quantities for Bernoulli bandit models.

However, Lemma 5.23 in Section 5.6.3 shows that the strategy using uniform sampling and recommending the empirical best arm satisfies $p_t(\nu) \leq \exp(-I^*(\nu)t)$, where $I^*(\nu)$ is the quantity defined in Theorem 5.14 whose expression for Bernoulli distributions is

$$I^*(\nu) = I^*(\theta_1, \theta_2) = \frac{\mathrm{K}\left(\frac{\theta_1+\theta_2}{2}, \theta_1\right) + \mathrm{K}\left(\frac{\theta_1+\theta_2}{2}, \theta_2\right)}{2}.$$

Hence this simple strategy matches the bound of Theorem 5.14 in the fixed-budget setting (see Remark TC). It can be moreover observed that $I^*(\nu)$ is very close to $c^*(\nu) = \mathrm{KL}^*(\theta_1, \theta_2)$, and thus the problem-dependent optimal strategy described in Proposition 5.18 can be approximated by a very simple, universal algorithm. This fact is illustrated in Figure 5.5, on which we represent the different informational function c_*,I_*,c^* and I^* (defined in Theorem 5.13 and 5.14) when the mean μ_1 varies, for two fixed values of μ_2. In can be observed that $c^*(\nu)$ and $c_*(\nu)$ are almost indistinguishable from $I^*(\nu)$ and $I_*(\nu)$ respectively, while there is a gap between $c^*(\nu)$ and $c_*(\nu)$.

FIXED-CONFIDENCE SETTING

As illustrated in Figure 5.5, $c_*(\nu)$ and $I_*(\nu)$ are also very close, thus there is a strong incentive to use uniform sampling in the fixed-confidence setting as well. Finding an algorithm sampling the arms uniformly and matching the bound of Theorem 5.14 is therefore a crucial matter. This boils down to determining a proper stopping rule. In all the algorithms studied so far, the stopping rule was based on the difference of the empirical means of the arms. For Bernoulli arms the 1/2-Elimination procedure described in Algorithm 5 can be used, as each distribution ν_a is bounded and therefore 1/4-subgaussian. More precisely, with $\beta(t, \delta)$ as in Theorem 5.15, the algorithm stopping at the first time t such that

$$\hat{\mu}_1(t) - \hat{\mu}_2(t) > \sqrt{2\beta(t, \delta)/t}$$

has its sample complexity bounded by $2/(\mu_1 - \mu_2)^2 \log(1/\delta) + o\left(\log(1/\delta)\right)$. The expressions of $I_*(\nu)$ as a function of the means of the arms is given by

$$I_*(\nu) = I_*(\mu_1, \mu_2) = \frac{d\left(\mu_1, \frac{\mu_1+\mu_2}{2}\right) + d\left(\mu_2, \frac{\mu_1+\mu_2}{2}\right)}{2}.$$

Pinsker's inequality implies that $I_*(\mu_1, \mu_2) > (\mu_1 - \mu_2)^2/2$ and this algorithm does not match the lower bound of Theorem 5.14 relative to the fixed-confidence setting. The approximation $I_*(\mu_1, \mu_2) = (\mu_1 -$

Algorithm 6 Sequential Generalized Likelihood Ratio Test (SGLRT)

Require: Exploration function $\beta(t,\delta)$.

Initialization: $\hat{\mu}_1(0) = \hat{\mu}_2(0) = 0$. $t = 0$.
while $(tI_*(\hat{\mu}_1(t), \hat{\mu}_2(t)) \leq \beta(t,\delta)) \bigcup (t = 1[2])$ **do**
  $t = t + 1$. $A_t = t[2]$.
  Observe $X_t \sim \nu_{A_t}$ and compute the empirical means $\hat{\mu}_1(t)$ and $\hat{\mu}_2(t)$.
end while
return $a = \underset{a=1,2}{\text{argmax}}\ \hat{\mu}_a(t)$.

$\mu_2)^2/(8\mu_1(1-\mu_1)) + o\left((\mu_1-\mu_2)^2\right)$ suggests that the loss with respect to the optimal error exponent is particularly significant when both means are close to 0 or 1.

To circumvent this drawback, we propose the SGLRT (for Sequential Generalized Likelihood Ratio Test) algorithm, described in Algorithm 6. The stopping rule is based on the distance between the empirical means of the arms, measured with the function I_*, and is related to the generalized likelihood ratio statistic for testing the equality of two Bernoulli proportions. To test $H_0 : (\mu_1 = \mu_2)$ against $H_1 : (\mu_1 \neq \mu_2)$ based on $t/2$ paired samples of the arms $W_s = (X_s, Y_s)$, the Generalized Likelihood Ratio Test (GLRT) rejects H_0 when

$$\exp(-tI_*(\hat{\mu}_{1,t/2}, \hat{\mu}_{2,t/2})) = \frac{\max_{\mu_1,\mu_2:\mu_1=\mu_2} L(W_1,\ldots,W_{t/2};\mu_1,\mu_2)}{\max_{\mu_1,\mu_2} L(W_1,\ldots,W_{t/2};\mu_1,\mu_2)} < z_\delta,$$

where $L(W_1,\ldots,W_{t/2};\mu_1,\mu_2)$ denote the likelihood of the observations given parameters μ_1 and μ_2. The equality in the previous display is a consequence of the rewriting

$$I_*(x,y) \quad = \quad H\left(\frac{x+y}{2}\right) - \frac{1}{2}\left[H\left(\frac{x}{2}\right) + H\left(\frac{y}{2}\right)\right],$$

where $H(x) = -x\log(x) - (1-x)\log(1-x)$ denotes the binary entropy function. Hence, Algorithm (6) can be interpreted as a sequential version of the GLRT with (varying) threshold $z_{t,\delta} = \exp(-\beta(t,\delta))$.

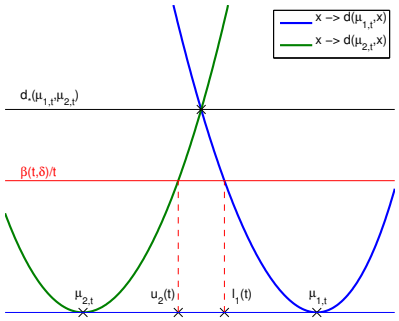

Figure 5.6: The KL-confidence intervals used by KL-LUCB are separated if and only if the threshold $\beta(t,\delta)/t$ is below $d^*(\hat{\mu}_{1,t}, \hat{\mu}_{2,t})$

Elements of analysis of the SGLRT. The SGLRT algorithm is also related to the KL-LUCB or KL-Racing algorithms described in Section 5.2. Indeed, for two armed-bandits, both algorithm use uniform sampling and stop at round t (having used $2t$ samples of the arms), when $l_1(t) > u_2(t)$ or $l_2(t) > u_1(t)$. Figure 5.6 can help convince oneself that this stopping condition is equivalent to stopping when $d_*(\hat{\mu}_{1,t}, \hat{\mu}_{2,t}) > \beta(t,\delta)/t$, where $\beta(t,\delta)$ is the exploration rate used by KL-LUCB.

From Figure 5.5, $I_*(x,y)$ mostly coincides with $d_*(x,y)$ and more precisely $I_*(x,y) < d_*(x,y)$. Using all this, one can upper bound the probability of error of the SGLRT, for example when $\mu_1 < \mu_2$:

$$\begin{aligned}
&\mathbb{P}_\nu\left(\exists t \in 2\mathbb{N}^* : \hat{\mu}_{1,t/2} > \hat{\mu}_{2,t/2}, tI_*(\hat{\mu}_{1,t/2}, \hat{\mu}_{2,t/2}) > \beta(t,\delta)\right) \\
&\qquad \le \mathbb{P}_\nu\left(\exists t \in 2\mathbb{N}^* : \hat{\mu}_{1,t/2} > \hat{\mu}_{2,t/2}, (t/2)d_*(\hat{\mu}_{1,t/2}, \hat{\mu}_{2,t/2}) > (\beta(t,\delta)/2)\right) \\
&\qquad = \mathbb{P}_\nu\left(\exists s \in \mathbb{N}^* : \hat{\mu}_{1,s} > \hat{\mu}_{2,s}, sd_*(\hat{\mu}_{1,s}, \hat{\mu}_{2,s}) > (\beta(2s,\delta)/2)\right)
\end{aligned}$$

This is an upper bound of the probability of error of KL-LUCB using the exploration rate $\tilde{\beta}(t,\delta) = \beta(2t,\delta)/2$ and proceeding as in Theorem 5.1, one can prove the following result.

Lemma 5.20. *With the exploration rate*

$$\beta(t,\delta) = 2\log\left(\frac{t(\log(3t))^2}{\delta}\right)$$

the SGLRT algorithm is δ-PAC.

For this exploration rate, we were able to obtain the following asymptotic guarantee on the stopping time τ of Algorithm 6, using Lemma 5.21 below (proved in Section 5.6.4):

$$\forall \alpha > 0, \quad \limsup_{\delta\to\infty} \frac{\tau}{\log(1/\delta)} \le \frac{2(1+\alpha)}{I_*(\mu_1,\mu_2)} \ a.s.$$

Still from Lemma 5.21, it follows that with an exploration rate $\beta(t,\delta) = \log((\log(t)+1)/\delta)$ —for which the SGLRT algorithm is not provably δ-PAC—, one has

$$\forall \alpha > 0, \quad \limsup_{\delta\to\infty} \frac{\tau}{\log(1/\delta)} \le \frac{(1+\alpha)}{I_*(\mu_1,\mu_2)} \ a.s..$$

No upper bound on $\mathbb{E}_\nu[\tau]/\log(1/\delta)$ can be deduced from this result, but it provides an intuition on which stopping rule to use. By analogy with the result of Theorem 5.15 we conjecture that the use of and exploration rate of order $\log(\log(t)/\delta)$ should also lead to a δ-PAC algorithm. This conjecture is supported by the numerical experiments reported in Section 5.4.4 below.

Lemma 5.21. *Let f and g be two continuous function such that $f(\mu_1,\mu_2) \neq 0$ and $g(t) = o(t^r)$ for all $r > 1$. For every $\alpha > 0$ the strategy using uniform sampling and the stopping rule*

$$\tau = \inf\left\{t \in 2\mathbb{N}^* : t\, f(\hat{\mu}_{1,t/2}, \hat{\mu}_{2,t/2}) \ge \log\frac{g(t)}{\delta}\right\} \text{ satisfies } \mathbb{P}_\nu\left(\limsup_{\delta\to 0} \frac{\tau}{\log(1/\delta)} \le \frac{1+\alpha}{f(\mu_1,\mu_2)}\right) = 1.$$

5.4.4 Numerical experiments

The goal of this Section is twofold: to compare results obtained in the fixed-budget and fixed-confidence settings and to illustrate the improvement resulting from the adoption of the reduced exploration rate of Theorem 5.15.

In Figure 5.7, we consider two Gaussian bandit models with known common variance: the 'easy' one is $\{\mathcal{N}(0.5, 0.25), \mathcal{N}(0, 0.25)\}$, corresponding to $\kappa_C = \kappa_B = \kappa = 8$, on the left; and the 'difficult' one is $\{\mathcal{N}(0.01, 0.25), \mathcal{N}(0, 0.25)\}$, that is $\kappa = 2 \times 10^4$, on the right. In the fixed-budget setting, stars ('*') report the probability of error $p_n(\nu)$ as a function of n. In the fixed-confidence setting, we plot both the empirical probability of error by circles ('O') and the specified maximal error probability δ by crosses ('X') as a function of the empirical average of the running times. Note the logarithmic scale used for the probabilities on the y-axis. All results are averaged on $N = 10^6$ independent Monte Carlo replications. For comparison purposes, a plain line represents the theoretical rate $x \mapsto \exp(-x(1/\kappa))$ which is a straight line on the log scale.

In the fixed-confidence setting, we report results for algorithms of the form (5.29) with $g(t, \delta) = \sqrt{2\sigma^2 t \beta(t, \delta)}$ for three different exploration rates $\beta(t, \delta)$. The exploration rate we consider are: the provably-PAC rate of Robbins' algorithm $\log(t/\delta)$ (large blue symbols), the conjectured 'optimal' exploration rate $\log((\log(t) + 1)/\delta)$, almost provably δ-PAC according to Theorem 5.15 (bold green symbols), and the rate $\log(1/\delta)$, which would be appropriate if we were to perform the stopping test only at a single pre-specified time (orange symbols). For each algorithm, the log probability of error is approximately a linear function of the number of samples, with a slope close to $-1/\kappa$, where κ is the complexity. We can visualize the gain in sample complexity achieved by smaller exploration rates, but while the rate $\log((\log(t) + 1)/\delta)$ appears to guarantee the desired probability of error across all problems, the use of $\log(1/\delta)$ seems too risky, as one can see that the probability of error becomes larger than δ on difficult problems. To illustrate the gain in sample complexity when the means of the arms are known, we add in red the SPRT algorithm mentioned in the introduction of Section 5.4 along with the theoretical relation between the probability of error and the expected number of samples, materialized as a dashed line. The SPRT stops for t such that $|(\mu_1 - \mu_2)(S_{1,t/2} - S_{2,t/2})| > \log(1/\delta)$.

Robbins' algorithm is δ-PAC and matches the complexity (which is illustrated by the slope of the measures), though in practice the use of the exploration rate $\log((\log(t) + 1)/\delta)$ leads to huge gain in terms of number of samples used. It is important to keep in mind that running times play the same role as error exponents and hence the threefold increase of average running times observed on the rightmost

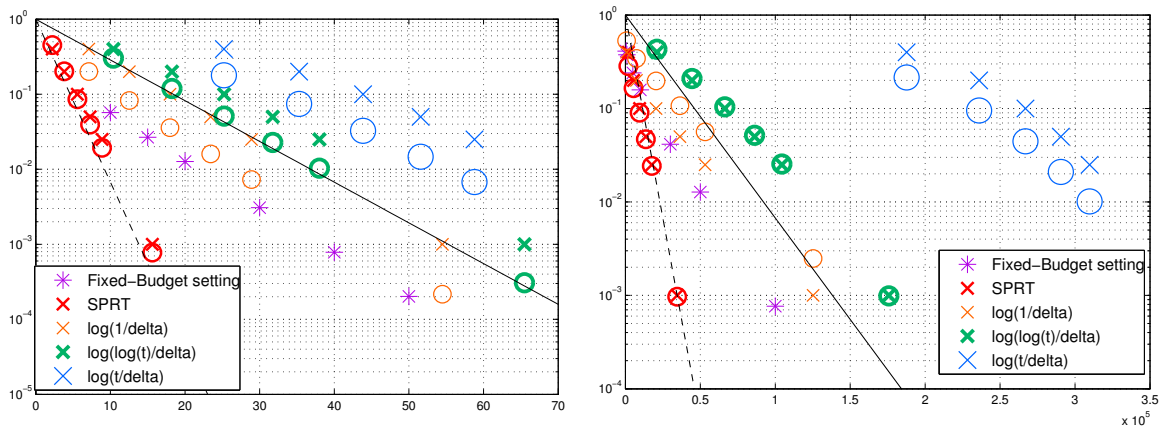

Figure 5.7: Experimental results for Gaussian bandit models

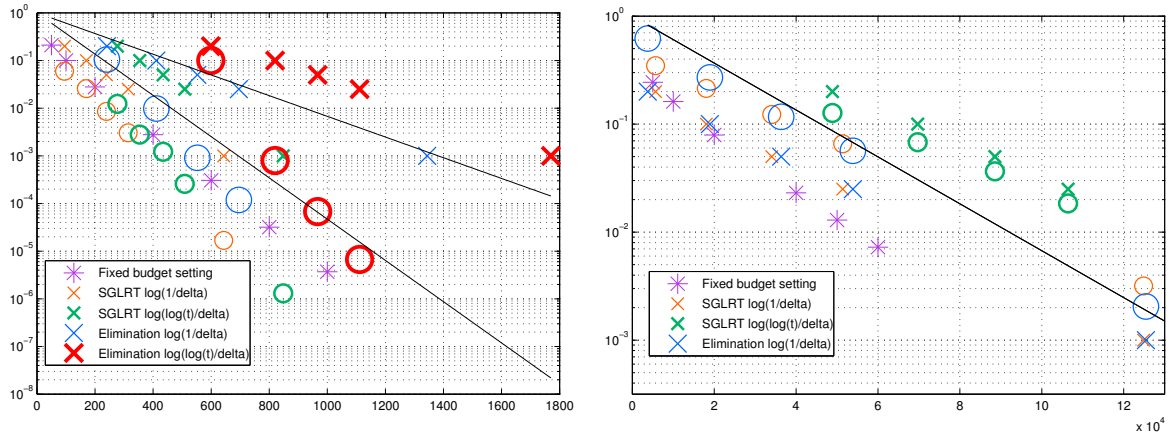

Figure 5.8: Results for Bernoulli bandit models: $0.2 - 0.1$ (left) and $0.51 - 0.5$ (right).

plot of Figure 5.7 when using $\beta(t,\delta) = \log(t/\delta)$ is really prohibitive.

In Figure 5.8, we compare on two Bernoulli bandit models the performance of the SGLRT algorithm (Algorithm 6) using two different exploration rates, $\log(1/\delta)$ and $\log((\log(t)+1)/\delta)$, to the algorithm that stops when the difference of empirical means exceeds the threshold $\sqrt{2\beta(t,\delta)/t}$ (for the same exploration rates), that we refer to as 'Elimination'. Plain lines also materialize the theoretical optimal rate $x \mapsto \exp(-x/\kappa_C(\nu))$ and the rate attained by the Elimination algorithm $x \mapsto \exp(-x/\kappa')$, where $\kappa' = 2/(\mu_1 - \mu_2)^2$. On the bandit model $0.51 - 0.5$ (right) theses two rates are very close and SGLRT mostly coincides with Elimination, but on the bandit model $0.2 - 0.1$ (left) the practical gain of the use of a more sophisticated stopping strategy is well illustrated. Besides, our experiments show that SGLRT using $\log((\log(t)+1)/\delta)$ is δ-PAC on both the (relatively) easy and difficult problems we consider, unlike the other algorithms considered.

If one compares on each problem the results for the fixed-budget setting to those for the best δ-PAC algorithm (or conjectured δ-PAC for the SGLRT algorithm for Bernoulli bandits), in green, one can see that to obtain the same probability of error, the fixed-confidence algorithm needs an average number of samples of order at least twice larger than the deterministic number of samples required by the fixed-budget setting algorithm. This remark should be related to the fact that a δ-PAC algorithm is designed to be uniformly good across all problems, whereas consistency is a weak requirement in the fixed-budget setting: any strategy that draws both arm infinitely often and recommends the empirical best is consistent. Figure 5.7 shows that when the values of μ_1 and μ_2 are unknown, the sequential version of the test is no more preferable to its batch counterpart and can even become much worse if the exploration rate $\beta(t,\delta)$ is chosen too conservatively. This observation should be mitigated by the fact that the sequential (or fixed-confidence) approach is adaptive with respect to the difficulty of the problem whereas it is impossible to predict the efficiency of a batch (or fixed-budget) experiment without some prior knowledge regarding the problem under consideration.

5.4.5 Proof of Theorem 5.15 and Theorem 5.16

Proof of Theorem 5.15. According to (5.30) it boils down to finding an exploration rate such that $\mathbb{P}(\exists t \in \mathbb{N}^* : S_t > \sqrt{2\sigma^2 t\beta(t,\delta)}) \leq \delta$, where $S_t = X_1 + \cdots + X_t$ is a sum of i.i.d. normal random variable.

Let $\beta(t,\delta)$ be of the form $\beta(t,\delta) = \log\frac{1}{\delta} + c\log\log\frac{1}{\delta} + d\log\log(et)$, for some constants $c > 0$ and

$d > 1$. Lemma A.1 given in Appendix A yields

$$\mathbb{P}\Big(\exists t \in \mathbb{N} : S_t > \sqrt{2\sigma^2 t\beta(t,\delta)}\Big) \leq \zeta\Big(d\big(1 - \frac{1}{2(z + c\log z)}\big)\Big)\frac{\sqrt{e}}{(2\sqrt{2})^d}\frac{(\sqrt{z + c\log z} + \sqrt{8})^d}{z^c}\delta,$$

where $z := \log\frac{1}{\delta} > 0$. To upper bound the above probability by δ, at least for large values of z (which corresponds to small values of δ), it suffices to choose the parameters c and d such that

$$\sqrt{e}\,\zeta\Big(d\big(1 - \frac{1}{2(z + c\log z)}\big)\Big)\frac{1}{(2\sqrt{2})^d}\frac{(\sqrt{z + c\log z} + 2\sqrt{2})^d}{z^c} \leq 1.$$

For $c = d/2$, the left hand side tends to $\sqrt{e}\zeta(d)/(2\sqrt{2})^d$ when z goes to infinity, which is smaller than 1 for $d \geq 1.47$. Thus, for δ small enough, the desired inequality holds for $d = 3/2$ and $c = 3/4$, which corresponds to the exploration rate of Theorem 5.15.

Proof of Theorem 5.16. Let $\alpha = \sigma_1/(\sigma_1 + \sigma_2)$. We first prove that with the exploration rate $\beta(t,\delta) = \log(t/\delta) + 2\log\log(6t)$ the algorithm is δ-PAC. Assume that $\mu_1 > \mu_2$ and recall $\tau = \inf\{t \in \mathbb{N} : |d_t| > \sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)}\}$. The probability of error of the α-elimination strategy is upper bounded by

$$\begin{aligned}
\mathbb{P}_\nu\left(d_\tau \leq -\sqrt{2\sigma_\tau^2(\alpha)\beta(\tau,\delta)}\right) &\leq \mathbb{P}_\nu\left(d_\tau - (\mu_1 - \mu_2) \leq -\sqrt{2\sigma_\tau^2(\alpha)\beta(\tau,\delta)}\right)\\
&\leq \mathbb{P}_\nu\left(\exists t \in \mathbb{N}^* : d_t - (\mu_1 - \mu_2) < -\sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)}\right)\\
&\leq \sum_{t=1}^{\infty}\exp\left(-\beta(t,\delta)\right),
\end{aligned}$$

by an union bound and Chernoff bound applied to $d_t - (\mu_1 - \mu_2) \sim \mathcal{N}\left(0, \sigma_t^2(\alpha)\right)$. The choice of $\beta(t,\delta)$ mentioned above ensures that the series in the right hand side is upper bounded by δ, which shows the algorithm is δ-PAC:

$$\sum_{t=1}^{\infty} e^{-\beta(t,\delta)} \leq \delta\sum_{t=1}^{\infty}\frac{1}{t(\log(6t))^2} \leq \delta\left(\frac{1}{(\log 6)^2} + \int_1^\infty \frac{dt}{t(\log(6t))^2}\right) = \delta\left(\frac{1}{(\log 6)^2} + \frac{1}{\log(6)}\right) \leq \delta.$$

To upper bound the expected sample complexity, we start by upper bounding the probability that τ exceeds some deterministic time T:

$$\begin{aligned}
\mathbb{P}_\nu(\tau \geq T) &\leq \mathbb{P}_\nu\left(\forall t = 1\dots T,\ d_t \leq \sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)}\right) \leq \mathbb{P}_\nu\left(d_T \leq \sqrt{2\sigma_T^2(\alpha)\beta(T,\delta)}\right)\\
&= \mathbb{P}_\nu\left(d_T - (\mu_1 - \mu_2) \leq -\left[(\mu_1 - \mu_2) - \sqrt{2\sigma_T^2(\alpha)\beta(T,\delta)}\right]\right)\\
&\leq \exp\left(-\frac{1}{2\sigma_T^2(\alpha)}\left[(\mu_1 - \mu_2) - \sqrt{2\sigma_T^2(\alpha)\beta(T,\delta)}\right]^2\right).
\end{aligned}$$

The last inequality follows from Chernoff bound and holds for T such that $(\mu_1 - \mu_2) > \sqrt{2\sigma_T^2(\alpha)\beta(T,\delta)}$. Now, for $\gamma \in]0,1[$ we introduce

$$T_\gamma^* := \inf\left\{t_0 \in \mathbb{N} : \forall t \geq t_0, (\mu_1 - \mu_2) - \sqrt{2\sigma_t^2(\alpha)\beta(t,\delta)} > \gamma(\mu_1 - \mu_2)\right\}.$$

This quantity is well defined as $\sigma_t^2(\alpha)\beta(t,\delta)$ go to zero when t goes to infinity. Then,

$$\begin{aligned}
\mathbb{E}_\nu[\tau] &\leq T_\gamma^* + \sum_{T=T_\gamma^*+1} \mathbb{P}(\tau \geq T) \\
&\leq T_\gamma^* + \sum_{T=T_\gamma^*+1} \exp\left(-\frac{1}{2\sigma_T^2(\alpha)}\left[(\mu_1-\mu_2) - \sqrt{2\sigma_T^2(\alpha)\beta(T,\delta)}\right]^2\right) \\
&\leq T_\gamma^* + \sum_{T=T_\gamma^*+1}^{\infty} \exp\left(-\frac{1}{2\sigma_T^2(\alpha)}\gamma^2(\mu_1-\mu_2)^2\right).
\end{aligned}$$

For all $t \in \mathbb{N}^*$, it is easy to show that the following upper bound on $\sigma_t^2(\alpha)$ holds:

$$\forall t \in \mathbb{N},\ \sigma_t^2(\alpha) \leq \frac{(\sigma_1+\sigma_2)^2}{t} \times \frac{t - \frac{\sigma_1}{\sigma_2}}{t - \frac{\sigma_1}{\sigma_2} - 1}. \tag{5.32}$$

Using the bound (5.32), one has

$$\begin{aligned}
\mathbb{E}_\nu[\tau] &\leq T_\gamma^* + \int_0^\infty \exp\left(-\frac{t}{2(\sigma_1+\sigma_2)^2}\frac{t-\frac{\sigma_1}{\sigma_2}-1}{t-\frac{\sigma_1}{\sigma_2}}\gamma^2(\mu_1-\mu_2)^2\right)dt \\
&\leq T_\gamma^* + \frac{2(\sigma_1+\sigma_2)^2}{\gamma^2(\mu_1-\mu_2)^2}\exp\left(\frac{\gamma^2(\mu_1-\mu_2)^2}{2(\sigma_1+\sigma_2)^2}\right).
\end{aligned}$$

We now give an upper bound on T_γ^*. Let $r \in [0, e/2-1]$. There exists $N_0(r)$ such that for $t \geq N_0(r)$, $\beta(t,\delta) \leq \log(t^{1+r}/\delta)$. Using also (5.32), one gets $T_\gamma^* = \max(N_0(t), \tilde{T}_\gamma)$, where

$$\tilde{T}_\gamma = \inf\left\{t_0 \in \mathbb{N} : \forall t \geq t_0, \frac{(\mu_1-\mu_2)^2}{2(\sigma_1+\sigma_2)^2}(1-\gamma)^2 t > \frac{t-\frac{\sigma_1}{\sigma_2}-1}{t-\frac{\sigma_1}{\sigma_2}}\log\frac{t^{1+r}}{\delta}\right\}.$$

If $t > (1+\gamma\frac{\sigma_1}{\sigma_2})/\gamma$ one has $(t-\frac{\sigma_1}{\sigma_2}-1)/(t-\frac{\sigma_1}{\sigma_2}) \leq (1-\gamma)^{-1}$. Thus $\tilde{T}_\gamma = \max((1+\gamma\frac{\sigma_1}{\sigma_2})/\gamma, T'_\gamma)$, with

$$T'_\gamma = \inf\left\{t_0 \in \mathbb{N} : \forall t \geq t_0, \exp\left(\frac{(\mu_1-\mu_2)^2}{2(\sigma_1+\sigma_2)^2}(1-\gamma)^3 t\right) \geq \frac{t^{1+r}}{\delta}\right\}.$$

Applying Lemma 5.22 with $\eta = \delta$, $s = 1+r$ and $\beta = (1-\gamma)^3(\mu_1-\mu_2)^2/(2(\sigma_1+\sigma_2)^2)$ leads to

$$T'_\gamma \leq \frac{(1+r)}{(1-\gamma)^3} \times \frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}\left[\log\frac{1}{\delta} + \log\log\frac{1}{\delta}\right] + R(\mu_1,\mu_2,\sigma_1,\sigma_2,\gamma,r),$$

with

$$R(\mu_1,\mu_2,\sigma_1,\sigma_2,\gamma,r) = \frac{1+r}{(1-\gamma)^3}\frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}\left[1 + (1+r)\log\left(\frac{2(\sigma_1+\sigma_2)^2}{(1-\gamma)^3(\mu_1-\mu_2)^2}\right)\right].$$

Now for $\epsilon > 0$ fixed, choosing r and γ small enough leads to

$$\mathbb{E}_\nu[\tau] \leq (1+\epsilon)\frac{2(\sigma_1+\sigma_2)^2}{(\mu_1-\mu_2)^2}\left[\log\frac{1}{\delta} + \log\log\frac{1}{\delta}\right] + \mathcal{C}(\mu_1,\mu_2,\sigma_1,\sigma_2,\epsilon),$$

where $\mathcal{C}$ is a constant independent of δ. This concludes the proof.

5.5 Conclusions and future work

In this chapter, we made progress towards understanding the complexity of best arms identification in bandit models. For two-armed bandits, we obtained complete results, identifying the complexity of both the fixed-budget and fixed-confidence settings in important parametric families of distributions. With the example of Bernoulli bandits, we especially show that these two complexities are not always equal.

In the fixed-confidence setting, we also provided in the general case (m best arm identification among $K > 2$ arms) new lower bounds as well as improved algorithms. The KL-LUCB and KL-Racing algorithms perform well in practice, but there is a small gap between the (informational) upper bound obtained for KL-UCB and the (informational) lower bound derived, that should be investigated further. As future work we also plan to examine whether the refined exploration rate that can be used in elimination algorithm for two-armed (sub)gaussian bandits could be incorporated to the KL-LUCB or KL-Racing algorithms, and if better theoretical guarantees for the SGLRT algorithm can be obtained.

We acknowledge that we were not able to characterize optimal algorithms in the fixed budget setting when there are more than two arms. We presented a first attempt to derive new lower bounds in the paper [Kaufmann et al., 2014b], but there is still room for improvements in this direction.

5.6 Elements of proof

5.6.1 A useful technical lemma

The following lemma is useful in Sections 5.2 and 5.4 to obtain upper bounds on the number of sample used by algorithms for the fixed-confidence setting.

Lemma 5.22. *For every $\beta, \eta > 0$ and $s \in [0, e/2]$, the following implication is true:*

$$x_0 = \frac{s}{\beta}\log\left(\frac{e\log\left(1/(\beta^s\eta)\right)}{\beta^s\eta}\right) \quad \Rightarrow \quad \forall x \geq x_0, \;\; e^{\beta x} \geq \frac{x^s}{\eta}.$$

Proof Lemma 5.22 easily follows from the fact that for any $s, \eta > 0$,

$$x_0 = s\log\left(\frac{e\log\left(\frac{1}{\eta}\right)}{\eta}\right) \quad \Rightarrow \quad \forall x \geq x_0, \;\; e^{x} \geq \frac{x^s}{\eta}$$

Indeed, it suffices to apply this statement to $x = x\beta$ and $\eta = \eta\beta^s$. The mapping $x \mapsto e^x - x^s/\eta$ is increasing when $x \geq s$. As $x_0 \geq s$, it suffices to prove that x_0 defined above satisfies $e^{x_0} \geq x_0^s/\eta$.

$$\begin{aligned}\log\left(\frac{x_0^s}{\eta}\right) &= s\log\left(s\log\left(\frac{e\log\frac{1}{\eta}}{\eta}\right)\right) + \log\frac{1}{\eta} = s\left(\log(s) + \log\left[\log\frac{1}{\eta} + \log\left(e\log\frac{1}{\eta}\right)\right]\right) + \log\frac{1}{\eta} \\ &\leq s\left(\log(s) + \log\left[2\log\frac{1}{\eta}\right]\right) + \log\frac{1}{\eta}\end{aligned}$$

where we use that for all y, $\log(y) \leq \frac{1}{e}y$. Then

$$\log\left(\frac{x_0^s}{\eta}\right) \leq s\left(\log(s) + \log(2) + \log\log\frac{1}{\eta} + \log\frac{1}{\eta}\right).$$

For $s \leq \frac{e}{2}$, $\log(s) + \log(2) \leq 1$, hence

$$\log\left(\frac{x_0^s}{\eta}\right) \leq s\left(1 + \log\log\frac{1}{\eta} + \log\frac{1}{\eta}\right) = s\log\left(\frac{e\log\left(\frac{1}{\eta}\right)}{\eta}\right) = x_0,$$

which is equivalent to $e^{x_0} \geq \frac{x_0^s}{\eta}$ and concludes the proof.

5.6.2 Proof of Lemma 5.4

The quantity we have to bound in order to prove Lemma 5.4 is

$$A := \sum_{u=\lceil C_1\gamma/d(\mu_a,c)\rceil+1}^{T} \mathbb{P}\left(ud(\hat{\mu}_{a,u}, c) \leq \gamma\right).$$

This sum also arises in the analysis of the KL-UCB algorithm and is precisely upper-bounded by [Cappé et al., 2013] in Appendix A.2, for the choice $C_1 = 1$. However, in order to obtain an exponential decay in γ, we have to adapt their method to the choice $C_1 > 1$. Introducing

$$d^+(x,c) = d(x,c)\mathbb{1}_{(x<c)} \text{ and } d^-(x,c) = d(x,c)\mathbb{1}_{(x>c)},$$

we use:

$$A \leq \sum_{u=n_1(a,c,\gamma)+1}^{T} \mathbb{P}\left(ud^+(\hat{\mu}_{a,u}, c) \leq \gamma\right) \text{ for } \mu_a < c, \text{ and}$$

$$A \leq \sum_{u=n_1(a,c,\gamma)+1}^{T} \mathbb{P}\left(ud^-(\hat{\mu}_{a,u}, c) \leq \gamma\right) \text{ for } \mu_a > c,$$

with $n_1(a,c,\gamma) = \left\lceil \frac{C_1\gamma}{d(\mu_a,c)} \right\rceil$. We now introduce notation that will be useful in the rest of the proof. The two mappings

$$\begin{array}{rccc} d^+ : & [0,c] & \longrightarrow & [0, d(0,c)] \\ & x & \mapsto & d(x,c) \end{array} \qquad \begin{array}{rccc} d^- : & [c,1] & \longrightarrow & [0, d(1,c)] \\ & x & \mapsto & d(x,c) \end{array}$$

are bijective and monotone. Then, for $\alpha \in [0, d(\mu_a, c)]$, the quantity $s^*_\alpha(\mu_a, c)$ is well-defined by:

$$d(s^*_\alpha(\mu_a,c), c) = \alpha \quad \text{and} \quad s^*_\alpha(\mu_a,c) \in (\mu_a, c).$$

With this new notation, one has, for $a \in (\mathcal{S}^*_m)^c$:

$$\mathbb{P}\left(ud^+(\hat{\mu}_{a,u}, c) \leq \gamma)\right) = \mathbb{P}\left(d^+(\hat{\mu}_{a,u}, c) \leq \frac{\gamma}{u}\right) = \mathbb{P}\left(\hat{\mu}_{a,u} \geq s^*_{\frac{\gamma}{u}}(\mu_a, c)\right).$$

And for $a \in \mathcal{S}^*_m$:

$$\mathbb{P}\left(ud^-(\hat{\mu}_{a,u}, c) \leq \gamma)\right) = \mathbb{P}\left(\hat{\mu}_{a,u} \leq s^*_{\frac{\gamma}{u}}(\mu_a, c)\right).$$

Using Chernoff's concentration inequality and a comparison with an integral yields in both cases:

$$A \leq \sum_{u=n_1(a,c,\gamma)+1}^{T} \exp\left(-ud\left(s^*_{\frac{\gamma}{u}}(\mu_a,c), \mu_a\right)\right) \leq \int_{n_1(a,c,\gamma)}^{\infty} \exp\left(-ud\left(s^*_{\frac{\gamma}{u}}(\mu_a,c), \mu_a\right)\right) du.$$

With the change of variable $u = \gamma v$, one has:

$$A \leq \underbrace{\gamma \int_{\frac{C_1}{d(\mu_a,c)}}^{\infty} \exp\left(-\gamma v d\left(s^*_{\frac{1}{v}}(\mu_a,c),\mu_a\right)\right) dv}_{m_\gamma}. \tag{5.33}$$

An asymptotic equivalent. This last integral takes the form

$$\int_{\frac{C_1}{d(\mu_a,c)}}^{\infty} \exp\left(-\gamma\phi(v)\right) \quad \text{with} \quad \phi(v) = v d\left(s^*_{\frac{1}{v}}(\mu_a,c),\mu_a\right)$$

and ϕ is increasing. We can use the Laplace method for approximating the integral when γ goes to infinity.

$$\phi'(v) = d\left(s^*_{\frac{1}{v}}(\mu_a,c),\mu_a\right) - \frac{1}{v}\frac{d'\left(s^*_{\frac{1}{v}}(\mu_a,c),\mu_a\right)}{d'\left(s^*_{\frac{1}{v}}(\mu_a,c),c\right)} \geq 0.$$

And $\phi'\left(\frac{C_1}{d(\mu_a,c)}\right) = 0$ iff $C_1 = 1$. If $C_1 > 1$ the following equivalent holds:

$$\int_{\frac{C_1}{d(\mu_a,c)}}^{\infty} \exp\left(-\gamma\phi(v)\right) \underset{\gamma\to\infty}{\sim} \frac{\exp\left(-\gamma\phi\left(\frac{C_1}{d(\mu_a,c)}\right)\right)}{\gamma\phi'\left(\frac{C_1}{d(\mu_a,c)}\right)}.$$

Noting that $s^*_{\frac{d(\mu_a,c)}{C_1}}(\mu_a,c) = s_{C_1}(\mu_a,c)$, we get

$$m_\gamma \underset{\gamma\to\infty}{\sim} \frac{\exp(-\gamma F_{C_1}(\mu_a,c))}{\phi'\left(\frac{C_1}{d(\mu_a,c)}\right)} \quad \text{with} \quad F_{C_1}(\mu_a,c) = \frac{C_1 d\left(s_{C_1}(\mu_a,c),\mu_a\right)}{d(\mu_a,c)}.$$

And $\phi'\left(\frac{C_1}{d(\mu_a,c)}\right)$ can be written as

$$\phi'\left(\frac{C_1}{d(\mu_a,c)}\right) = \frac{d(\mu_a,c)}{C_1}\left(F_{C_1}(\mu_a,c) - \frac{d'(s_{C_1}(\mu_a,c),\mu_a)}{d'(s_{C_1}(\mu_a,c),c)}\right).$$

This asymptotic equivalent shows that, starting from (5.33), we cannot improve the constant $F_{C_1}(\mu_a,c)$ in the exponential with a bigger (and maybe non problem-dependent) one. If $C_1 = 1$ the same reasoning holds, but the Laplace equivalent is different and leads to:

$$m_\gamma \underset{\gamma\to\infty}{\sim} \sqrt{\gamma}\sqrt{\frac{\pi}{-2\phi''\left(\frac{1}{d(\mu_a,c)}\right)}},$$

which does not exhibit an exponential decay.

An 'optimal' bound of the probability. We now give a non-asymptotic upper bound of (5.33) involving the optimal rate $F_{C_1}(\mu_a,c)$ in the exponential. If $v \geq \frac{C_1}{d(\mu_a,c)}$, $s^*_{\frac{1}{v}}(\mu_a,c) \geq s^*_{\frac{d(\mu_a,c)}{C_1}}(\mu_a,c)$ and we can use this bound in the integral in (5.33) to get:

$$A \leq \int_{\frac{C_1}{d(\mu_a,c)}}^{\infty} \exp\left(-u d\left(s_{C_1}(\mu_a,c),\mu_a\right)\right) du = \frac{\exp\left(-F_{C_1}(\mu_a,c)\gamma\right)}{d\left(s_{C_1}(\mu_a,c),\mu_a\right)}.$$

5.6.3 Proof of Proposition 5.18

Bounding the probability of error of a static strategy using n_1 samples from arm 1 and n_2 samples from arm 2 relies on the following lemma.

Lemma 5.23. *Let $(X_{1,t})_{t\in\mathbb{N}}$ and $(X_{2,t})_{t\in\mathbb{N}}$ be two independent i.i.d sequences, such that $X_{1,1} \sim \nu_{\theta_1}$ and $X_{2,1} \sim \nu_{\theta_2}$ belong to an exponential family. Assume that $\mu(\theta_1) > \mu(\theta_2)$. Then*

$$\mathbb{P}\left(\frac{1}{n_1}\sum_{t=1}^{n_1} X_{1,t} < \frac{1}{n_2}\sum_{t=1}^{n_2} X_{2,t}\right) \le \exp(-(n_1+n_2)g_\alpha(\theta_1,\theta_2)),$$

where $\alpha = \frac{n_1}{n_1+n_2}$ and $g_\alpha(\theta_1,\theta_2) := \alpha \mathrm{K}(\alpha\theta_1 + (1-\alpha)\theta_2, \theta_1) + (1-\alpha)\mathrm{K}(\alpha\theta_1 + (1-\alpha)\theta_2, \theta_2)$.

The function $\alpha \mapsto g_\alpha(\theta_1,\theta_2)$, can be maximized analytically, and the value α^* that realizes the maximum is given by

$$\begin{aligned} \mathrm{K}(\alpha^*\theta_1 + (1-\alpha^*)\theta_2, \theta_1) &= \mathrm{K}(\alpha^*\theta_1 + (1-\alpha^*)\theta_2, \theta_2) \\ \alpha^*\theta_1 + (1-\alpha^*)\theta_1 &= \theta^* \\ \alpha^* &= \frac{\theta^* - \theta_2}{\theta_1 - \theta_2} \end{aligned}$$

where θ^* is defined by $\mathrm{K}(\theta^*, \theta_1) = \mathrm{K}(\theta^*, \theta_2) = \mathrm{K}^*(\theta_1, \theta_2)$. More interestingly, the associated rate is such that

$$g_{\alpha^*}(\theta_1,\theta_2) = \alpha^*\mathrm{K}(\theta^*,\theta_1) + (1-\alpha^*)\mathrm{K}(\theta^*,\theta_2) = \mathrm{K}^*(\theta_1,\theta_2),$$

which leads to Proposition 5.18.

Remark 5.24. *When $\mu_1 > \mu_2$, applying Lemma 5.23 with $n_1 = n_2 = t/2$ yields*

$$\mathbb{P}\left(\hat{\mu}_{1,t/2} < \mu_{2,t/2}\right) \le \exp\left(-\frac{\mathrm{K}\left(\theta_1, \frac{\theta_1+\theta_2}{2}\right) + \mathrm{K}\left(\theta_2, \frac{\theta_1+\theta_2}{2}\right)}{2} t\right) = \exp\Big(-I_*(\nu)t\Big),$$

which shows that the strategy based on uniform sampling that recommends the empirical best arm matches the lower bound of Theorem 5.13 for the fixed-budget setting.

Proof of Lemma 5.23. The i.i.d. sequences $(X_{1,t})_{t\in\mathbb{N}}$ and $(X_{2,t})_{t\in\mathbb{N}}$ have respective densities f_{θ_1} and f_{θ_2} where $f_\theta(x) = \exp(\theta x - b(\theta))$ and $\mu(\theta_1) = \mu_1, \mu(\theta_2) = \mu_2$. α is such that $n_1 = \alpha n$ and $n_2 = (1-\alpha)n$. One can write

$$\mathbb{P}\left(\frac{1}{n_1}\sum_{t=1}^{n_1} X_{1,t} - \frac{1}{n_2}\sum_{t=1}^{n_2} X_{2,t} < 0\right) = \mathbb{P}\left(\alpha\sum_{t=1}^{n_2} X_{2,t} - (1-\alpha)\sum_{t=1}^{n_1} X_{1,t} \ge 0\right).$$

For every $\lambda > 0$, multiplying by λ, taking the exponential of the two sides and using Markov's inequality (this technique is often referred to as Chernoff's method), one gets

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n_1}\sum_{t=1}^{n_1} X_{1,t} - \frac{1}{n_2}\sum_{t=1}^{n_2} X_{2,t} < 0\right) &\le \left(\mathbb{E}_\nu[e^{\lambda\alpha X_{2,1}}]\right)^{(1-\alpha)n} \left(\mathbb{E}_\nu[e^{\lambda(1-\alpha)X_{1,1}}]\right)^{\alpha n} \\ &= \exp\Bigg(n \underbrace{\left[(1-\alpha)\phi_{X_{2,1}}(\lambda\alpha) + \alpha\phi_{X_{1,1}}(-(1-\alpha)\lambda)\right]}_{G_\alpha(\lambda)}\Bigg) \end{aligned}$$

with $\phi_X(\lambda) = \log \mathbb{E}_\nu[e^{\lambda X}]$ for any random variable X. If $X \sim f_\theta$ a direct computation gives $\phi_X(\lambda) = b(\lambda+\theta) - b(\theta)$. Therefore the function $G_\alpha(\lambda)$ introduced above rewrites

$$G_\alpha(\lambda) = (1-\alpha)(b(\lambda\alpha+\theta_2) - b(\theta_2)) + \alpha(b(\theta_1 - (1-\alpha)\lambda) - b(\theta_1)).$$

Using that $b'(x) = \mu(x)$, we can compute the derivative of G and see that this function as a unique minimum in λ^* given by

$$\mu(\theta_1 - (1-\alpha)\lambda^*) = \mu(\theta_2 + \alpha\lambda^*) \quad \Leftrightarrow \quad \theta_1 - (1-\alpha)\lambda^* = \theta_2 + \alpha\lambda^* \quad \Leftrightarrow \quad \lambda^* = \theta_1 - \theta_2,$$

using that $\theta \mapsto \mu(\theta)$ is one-to-one. One can also show that

$$G(\lambda^*) \quad = \quad (1-\alpha)[b(\alpha\theta_1 + (1-\alpha)\theta_2) - b(\theta_2)] + \alpha[b(\alpha\theta_1 + (1-\alpha)\theta_2) - b(\theta_1)].$$

Using the expression of the KL-divergence between ν_{θ_1} and ν_{θ_2} as a function of the natural parameters: $\mathrm{K}(\theta_1,\theta_2) = \mu(\theta_1)(\theta_1-\theta_2) - b(\theta_1) + b(\theta_2)$, one can also show that

$$\begin{aligned}
&\alpha\mathrm{K}(\alpha\theta_1 + (1-\alpha)\theta_2, \theta_1) \\
&\qquad = -\alpha(1-\alpha)\mu(\alpha\theta_1 + (1-\alpha)\theta_2)(\theta_1-\theta_2) + \alpha[-b(\alpha\theta_1 + (1-\alpha)\theta_2) + b(\theta_1)] \\
&(1-\alpha)\mathrm{K}(\alpha\theta_1 + (1-\alpha)\theta_2, \theta_2) \\
&\qquad = \alpha(1-\alpha)\mu(\alpha\theta_1 + (1-\alpha)\theta_2)(\theta_1-\theta_2) + (1-\alpha)[-b(\alpha\theta_1 + (1-\alpha)\theta_2) + b(\theta_2)]
\end{aligned}$$

Summing these two equalities leads to

$$G(\lambda^*) = -\left[\alpha\mathrm{K}(\alpha\theta_1 + (1-\alpha)\theta_2, \theta_1) + (1-\alpha)\mathrm{K}(\alpha\theta_1 + (1-\alpha)\theta_2, \theta_2)\right] = -g_\alpha(\theta_1,\theta_2).$$

Hence the inequality $\mathbb{P}\left(\frac{1}{n_1}\sum_{t=1}^{n_1} X_{1,t} < \frac{1}{n_2}\sum_{t=1}^{n_2} X_{2,t}\right) \le \exp(nG(\lambda^*))$ concludes the proof.

5.6.4 Proof of Lemma 5.21.

We fix $\alpha > 0$ and introduce

$$\sigma = \max\left\{t \in 2\mathbb{N}^* : f(\hat{\mu}_{1,t/2}, \hat{\mu}_{2,t/2}) \le \frac{f(\mu_1,\mu_2)}{1+\alpha/2}\right\}.$$

By the law of large numbers, $\mathbb{P}(\sigma < +\infty) = 1$. Hence, $\lim_{n\to\infty}\mathbb{P}(\sigma \le n) = 1$ and for every $\alpha \in]0,1[$ there exists $N(\alpha,\alpha,\mu_1,\mu_2)$ such that $\mathbb{P}(\sigma \le N(\alpha,\alpha,\mu_1,\mu_2)) \ge 1-\alpha$. Therefore, introducing the event

$$E_\alpha \quad = \quad \left(\forall t \ge N(\alpha,\alpha,\mu_1,\mu_2), f(\hat{\mu}_{1,t/2}, \hat{\mu}_{2,t/2}) > \frac{f(\mu_1,\mu_2)}{1+\alpha/2}\right), \quad \text{one has} \quad \mathbb{P}(E_\alpha) \ge 1-\alpha.$$

On the event E_α,

$$\begin{aligned}
\tau &\le \max\left(N(\alpha,\alpha,\mu_1,\mu_2); \inf\left\{t \in \mathbb{N} : t\frac{f(\mu_1,\mu_2)}{1+\alpha/2} \ge \log\left(\frac{g(t)}{\delta}\right)\right\}\right) \\
\tau &\le N(\alpha,\alpha,\mu_1,\mu_2) + \inf\left\{t \in \mathbb{N} : t\frac{f(\mu_1,\mu_2)}{1+\alpha/2} \ge \log\left(\frac{g(t)}{\delta}\right)\right\}
\end{aligned}$$

We can use Lemma 5.22 to bound the right term in the right hand side, which shows that there exists a constant $C(\alpha, \mu_1, \mu_2)$ independent of δ such that

$$\tau \leq N(\alpha, \alpha, \mu_1, \mu_2) + \frac{1+\alpha}{f(\mu_1, \mu_2)} \left[\log \frac{1}{\delta} + \log \log \frac{1}{\delta} \right] + C(\alpha, \mu_1, \mu_2)$$

Thus we proved that for all $\alpha > 0$,

$$\mathbb{P}\left(\limsup_{\delta \to 0} \frac{\tau}{\log(1/\delta)} \leq \frac{1+\alpha}{f(\mu_1, \mu_2)} \right) \geq 1 - \alpha.$$

This concludes the proof.

Conclusion and perspectives

In this thesis, two different bandit problems have been studied: reward maximization and best arm(s) identification in bandit models. For the former, whose complexity is well-known, we have proposed and/or analysed algorithms based on Bayesian ideas that are optimal with respect to the (frequentist) regret. For the latter, we have introduced two complexity notions in the fixed-budget and fixed-confidence settings. We have provided new lower bounds on these complexities as well as improved algorithms matching these lower bounds in particular cases of two-armed bandits.

In both frameworks, we have been focused on obtaining distribution-dependent performance guarantees that feature information-theoretic quantities. A first comment is that the information quantities that appears in the complexity of regret minimization (Kullback-Leibler divergence) and best arm identification (Chernoff information) differ. A possible interpretation of this fact is that the error events in these two problems are different. Assume there are two arms and arm 1 is the best. Algorithms for regret minimization are such that arm 1 has been drawn a lot, so its mean is well estimated: one can consider that the s^{th} draw of arm 2 occurs when $(\hat{\mu}_{2,s} > \mu_1)$. The probability of this event involves Kullback-Leibler divergence. On the other hand, in best arm identification, the arms are drawn in a more comparable way, and a typical error occurs when the empirical means $\hat{\mu}_{1,s}$ and $\hat{\mu}_{2,s}$ are somewhere in the middle of the interval $[\mu_2, \mu_1]$ and in a reversed order. The probability of this event involves Chernoff information. In the fixed confidence setting, the lower bound derived in Chapter 5 when there are two arms involves another information-theoretic quantity, $\mathrm{K}_*(\theta_1, \theta_2)$, a Chernoff information in which the role of the arguments are reversed. It would be interesting to be able to interpret this quantity.

The general algorithms proposed for m best arms identification transpose recent improvements, related to the use of confidence intervals based on the Kullback-Leibler divergence, from regret minimization to the pure-exploration framework. As we demonstrated in this thesis the interest of using Bayesian algorithms for regret minimization, a natural question is: could Bayesian algorithms also be used for best arm(s) identification? Finding an heuristic like Thompson Sampling adapted for this different objective is not obvious. But for Bayes-UCB, I suspect that the Racing and LUCB algorithms using Bayesian confidence regions of the form $[l_a(t), u_a(t)]$ with $l_a(t) = Q(\frac{\delta}{Ct}, \pi_a^{t-1})$ and $u_a(t) = Q(1 - \frac{\delta}{Ct}, \pi_a^{t-1})$ can be shown to be δ-PAC for some constant C. At least, such an algorithm can be easily shown to be δ-PAC under the Bayesian modeling. However, it is not clear how to give an upper bound on the sample complexity $\mathbb{E}[\tau]$ in this Bayesian framework (that is, when the expectation includes an average over a prior distribution). In the same way we considered the Bayesian optimal strategy for the reward maximization objective, it would also be interesting to describe a Bayesian optimal strategy for best arm(s) identification. The papers [Naghshvar and Javidi, 2013, Chandrasekaran and Karp, 2014] consider particular cases with specific prior distributions.

Bandit models with correlated arms, like linear (contextual) bandit models have also only been studied in the rewards maximization framework. [Hoffman et al., 2014] are the first to study (single) best arm

identification in a linear bandit model, and to propose a Bayesian algorithm for this task. More precisely, each arm a (among K) is normally distributed with mean $x_a^T\theta$ and known variance σ^2, where $x_a \in \mathbb{R}^d$ is some features vector for am a and $\theta \in \mathbb{R}^d$ is an unknown parameter shared by all arms. Assuming a Gaussian prior distribution $\mathcal{N}\left(0, \eta^2 \mathrm{I}_d\right)$ on θ, the proposed algorithm uses upper and lower confidence bounds of the form $\mu_a(t) \pm \beta\sigma_a(t)$, where $\mu_a(t)$ and $\sigma_a(t)$ are respectively the mean and variance of the posterior distribution on $\mu_a := x_a^T\theta$ at time t. The algorithm is an adaptation of the UGapE algorithm of [Gabillon et al., 2012] using Bayesian confidence regions, for which the authors provide an upper bound on the probability of error in the fixed-budget setting. A first possible extension would be to consider the fixed-confidence setting, an using the LUCB algorithm with the Bayesian confidence regions described above should work. However, this algorithm would probably share the same shortcoming as the algorithms proposed by [Hoffman et al., 2014] for the fixed-budget setting, whose complexity term features a sum over all arms of an inverse squared gap $(\mu^* - \mu_a)^2$. In other words, the algorithm does not seem to take advantage of the correlation between arms, as it achieves the same performance as if the arms were regarded as independent. Thus, the right complexity of best arm identification in a linear bandit model is still to be investigated.

[Hoffman et al., 2014] also mention that their algorithm can be more generally applied to Gaussian Process optimization with a discretized space. In Gaussian Process Optimization, we have to find the maximum of a function $f : \mathcal{X} \to \mathbb{R}$, where f is assumed to be drawn from a Gaussian process. This optimization task naturally generalizes best arm identification in a (Bayesian) multi-armed bandit model. However, despite the optimization objective, the proposed algorithms are often analysed in terms of regret and not optimization error, or simple regret. Besides, in the benchmark of Bayesian optimization, we find the GP-UCB algorithm of [Srinivas et al., 2010], which is inspired by algorithms minimizing regret in classical bandits (and not algorithms for pure-exploration). It would be interesting to see whether transposing ideas from the best arm identification literature, like the use of upper *and* lower confidence bounds, could yield improved algorithms for Gaussian Process optimization. The algorithm proposed by [Hoffman et al., 2014] indeed seems to improve over GP-UCB. Other authors, like [Contal et al., 2013] also start to consider the use of pure-exploration tools for Gaussian Process optimization.

Interesting bandit problems that have not been studied in this thesis, neither from the rewards maximization nor from the pure-exploration perspective, are combinatorial bandit problems. Combinatorial bandit problems have been studied by [Cesa-Bianchi and Lugosi, 2012] in an adversarial setting. In such problems, arms are edges on a graph and a set of configurations $\mathcal{M}$ in this graph is available (e.g. subgraphs with m edges, spanning trees, matchings...). When an agent chooses a configuration, he observes some function of the rewards of each edge in this configuration (for example the sum of rewards, or the reward of each edge). Combinatorial bandit problems have seldom been considered in a stochastic setting, in which each arm produces i.i.d. rewards. It has been studied for example by [Lelarge et al., 2013] with an application to spectrum allocation in a wireless network. Still in a stochastic setting, the identification of the best configuration (without considering regret) is also a challenging task. For example it is not obvious how to design an algorithm based on eliminations (that is, a generalization of the Racing algorithm), as each individual arm belongs to several configurations and eliminating the worst arm in the best configuration would be problematic. Finally, it could also be investigated whether adopting a Bayesian approach is possible for these more complex combinatorial bandit problems.

Appendix A

Self normalized deviation inequalities

In several places in this thesis, we needed deviation inequalities for 'self normalized quantities' of the form A_t/B_t, where both A_t and B_t are random. For example we need to control the empirical mean of rewards collected from an arm up to a given time t, $S_a(t)/N_a(t)$, where both the number of draws up to time t, $N_a(t)$ and the sum of observations , $S_a(t)$, are random. Such deviation inequalities can be obtained by controlling the quantity the quantity $S_{a,s}/s$ uniformly for $s \in \{1, \dots, t\}$, or for $s \in \mathbb{N}^*$. We present deviation inequalities for similar quantities in this section.

A.1 Peeling trick versus mixtures method: the subgaussian case

In Chapter 5, in the proof of Theorem 5.15, a tight deviation inequality, uniform in t, is needed for the process $S_t/\sqrt{2t\sigma^2}$, where S_t is a sum of i.i.d random variables with distribution $\mathcal{N}\left(0, \sigma^2\right)$. We consider here the more general framework in which $S_t = X_1 + \cdots + X_t$ is a sum of independent increments that are σ^2-subgaussian, i.e. that satisfy, for every $\lambda \in \mathbb{R}$,

$$\phi_{X_i}(\lambda) := \log \mathbb{E}[\exp(\lambda X_i)] \leq \frac{\lambda^2\sigma^2}{2}.$$

A deviation inequality for $S_t/\sqrt{2t\sigma^2}$ can either be obtained using a so-called 'peeling-trick', as shown in Section A.1.1, or by the 'method of mixtures' that we present in Section A.1.2. Both methods rely in this case on the family of super-martingale $((W_t^\lambda)_{\lambda\in\mathbb{R}})$, indexed by $\lambda \in \mathbb{R}$, defined by

$$W_t^\lambda = \exp\left(\lambda S_t - \frac{\lambda^2\sigma^2}{2}t\right). \tag{A.1}$$

W_t^λ is a super-martingale with respect to the filtration $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$.

A.1.1 An 'optimal' confidence region obtained with the peeling-trick

For each $\lambda \in \mathbb{R}$, for each $u > 0$, the maximal inequality for super-martingales yields

$$\mathbb{P}\left(\bigcup_{t\geq 1}\left\{\lambda S_t - t\frac{\lambda^2\sigma^2}{2} > u\right\}\right) \leq \exp(-u). \tag{A.2}$$

Indeed, introducing the stopping time $N = \inf\{t \in \mathbb{N} : W_t^\lambda > e^u\}$, a maximal inequality yields

$$\mathbb{P}(N \leq n) = \mathbb{P}(\exists t \in [0, n] : W_t^\lambda > e^u) \leq e^{-u}\mathbb{E}[W_0^\lambda] \leq e^{-u}.$$

It follows that

$$\mathbb{P}(\exists t \in \mathbb{N} : W_t^\lambda > e^u) = \mathbb{P}(N < +\infty) = \lim_{n\to\infty} \mathbb{P}(N \le n) \le e^{-u}.$$

The peeling trick consists in partitioning the interval on which we want to control the process $S_t/\sqrt{2t\sigma^2}$ into 'slices' of exponentially growing size, and using inequality (A.2) for a well chosen value of λ on each slice. This idea is inspired by a proof of the Law of Iterated Logarithm given by [Neveu, 1972] and has been used for example by [Garivier and Moulines, 2011] and [Garivier and Cappé, 2011] to obtain deviation inequalities. It yields the following result.

Lemma A.1. *Let $S_t = X_1 + \cdots + X_t$ be a sum of independent, σ^2-subgaussian increments. Let $\zeta(u) = \sum_{k\ge1} k^{-u}$. For all $\beta > 1$ and $x \ge \frac{8}{(e-1)^2}$,*

$$\mathbb{P}\left(\exists t \in \mathbb{N}^* : \frac{S_t}{\sqrt{2\sigma^2 t}} > \sqrt{x + \beta \log\log(et)}\right) \le \sqrt{e}\,\zeta\Big(\beta\Big(1 - \frac{1}{2x}\Big)\Big)\Big(\frac{\sqrt{x}}{2\sqrt{2}} + 1\Big)^\beta \exp(-x).$$

Proof of Lemma A.1 We start by stating three technical lemmas, whose proofs are partly omitted.

Lemma A.2. *For every $\eta > 0$, every positive integer k, and every integer t such that $(1+\eta)^{k-1} \le t \le (1+\eta)^k$,*

$$\sqrt{\frac{(1+\eta)^{k-1/2}}{t}} + \sqrt{\frac{t}{(1+\eta)^{k-1/2}}} \le (1+\eta)^{1/4} + (1+\eta)^{-1/4}\,.$$

Lemma A.3. *For every $\eta > 0$,*

$$A(\eta) := \frac{4}{\left((1+\eta)^{1/4} + (1+\eta)^{-1/4}\right)^2} \ge 1 - \frac{\eta^2}{16}.$$

Lemma A.4. *Let t be such that $(1+\eta)^{k-1} \le t \le (1+\eta)^k$. Then, if $\lambda = \sigma^{-1}\sqrt{2zA(\eta)/(1+\eta)^{k-1/2}}$,*

$$\sigma\sqrt{2z} \ge \frac{A(\eta)z}{\lambda\sqrt{t}} + \frac{\lambda\sigma^2\sqrt{t}}{2}\,.$$

Proof of Lemma A.4:

$$\frac{A(\eta)z}{\lambda\sqrt{t}} + \frac{\lambda\sigma^2\sqrt{t}}{2} = \frac{\sigma\sqrt{2zA(\eta)}}{2}\left(\sqrt{\frac{(1+\eta)^{k-1/2}}{t}} + \sqrt{\frac{t}{(1+\eta)^{k-1/2}}}\right) \le \sigma\sqrt{2z}$$

according to Lemma A.2.

□

Let $\eta \in]0, e-1]$ to be defined later, and let $T_k^\eta = \mathbb{N} \cap \left[(1+\eta)^{k-1}, (1+\eta)^k\right[$.

$$\mathbb{P}\left(\bigcup_{t\ge1}\left\{\frac{S_t}{\sigma\sqrt{2t}} > \sqrt{x + \beta\log\log(et)}\right\}\right) \le \sum_{k=1}^\infty \mathbb{P}\left(\bigcup_{t\in T_k^\eta}\left\{\frac{S_t}{\sigma\sqrt{2t}} > \sqrt{x + \beta\log\log(et)}\right\}\right)$$
$$\le \sum_{k=1}^\infty \mathbb{P}\left(\bigcup_{t\in T_k^\eta}\left\{\frac{S_t}{\sigma\sqrt{2t}} > \sqrt{x + \beta\log\left(k\log(1+\eta)\right)}\right\}\right).$$

We use that $\eta \le e-1$ to obtain the last inequality since this condition implies

$$\log(\log(e(1+\eta)^{k-1}) \ge \log(k\log(1+\eta)).$$

For a positive integer k, let $z_k = x+\beta\log\left(k\log(1+\eta)\right)$ and $\lambda_k = \sigma^{-1}\sqrt{2z_kA(\eta)/(1+\eta)^{k-1/2}}$. Lemma A.4 shows that for every $t \in T_k^\eta$,

$$\left\{\frac{S_t}{\sigma\sqrt{2t}} > \sqrt{z_k}\right\} \subset \left\{\frac{S_t}{\sqrt{t}} > \frac{A(\eta)z_k}{\lambda_k\sqrt{t}} + \frac{\sigma^2\lambda_k\sqrt{t}}{2}\right\} .$$

Thus, using inequality (A.2),

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{t\in T_k^\eta}\left\{\frac{S_t}{\sigma\sqrt{2t}} > \sqrt{z_k}\right\}\right) &\le \mathbb{P}\left(\bigcup_{t\in T_k^\eta}\left\{\frac{S_t}{\sqrt{t}} > \frac{A(\eta)z_k}{\lambda_k\sqrt{t}} + \frac{\sigma^2\lambda_k\sqrt{t}}{2}\right\}\right) \\
&= \mathbb{P}\left(\bigcup_{t\in T_k^\eta}\left\{\lambda_k S_t - \frac{\sigma^2\lambda_k^2 t}{2} > A(\eta)z_k\right\}\right) \\
&\le \exp\left(-A(\eta)z_k\right) = \frac{\exp(-A(\eta)x)}{(k\log(1+\eta))^{\beta A(\eta)}} .
\end{aligned}$$

For x such that $x \ge \frac{8}{(e-1)^2}$, one chooses $\eta^2 = 8/x$ (which ensures $\eta \le e-1$). Using Lemma A.3, one obtains that $\exp(-A(\eta)x) \le \sqrt{e}\exp(-x)$. Moreover,

$$\frac{1}{\log(1+\eta)} \le \frac{1+\eta}{\eta} = \frac{\sqrt{x}}{2\sqrt{2}} + 1 .$$

Thus,

$$\mathbb{P}\left(\bigcup_{t\in T_k^\eta}\left\{\frac{S_t}{\sigma\sqrt{2t}} > \sqrt{z_k}\right\}\right) \le \frac{\sqrt{e}}{k^{\beta A(\eta)}}\left(\frac{\sqrt{x}}{2\sqrt{2}}+1\right)^{\beta A(\eta)}\exp(-x) \le \frac{\sqrt{e}}{k^{\beta A(\eta)}}\left(\frac{\sqrt{x}}{2\sqrt{2}}+1\right)^{\beta}\exp(-x)$$

and hence,

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{t\ge 1}\left\{\frac{S_t}{\sigma\sqrt{2t}} > \sqrt{x+\beta\log\log(et)}\right\}\right) &\le \sqrt{e}\zeta\left(\beta A(\eta)\right)\left(\frac{\sqrt{x}}{2\sqrt{2}}+1\right)^{\beta A(\eta)}\exp\left(-x\right) \\
&\le \sqrt{e}\zeta\left(\beta\left(1-\frac{1}{2x}\right)\right)\left(\frac{\sqrt{x}}{2\sqrt{2}}+1\right)^{\beta}\exp\left(-x\right) ,
\end{aligned}$$

using the lower bound on $A(\eta)$ given in Lemma A.3 and the fact that $A(\eta)$ is upper bounded by 1.

□

A.1.2 The mixtures method for martingales with subgausian increments

The so-called 'method of mixtures', introduced by [De La Pena et al., 2004], does not originally rely on the fact that W_t^λ defined in (A.1) is a super-martingale, but rather on the fact (which is a consequence of the super-martingale property in our case) that for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\lambda S_t - \frac{\lambda^2\sigma^2}{2}t\right)\right] \leq 1.$$

The method then consists in averaging this inequality over a Gaussian prior distribution. More precisely, assuming that $\lambda \sim \mathcal{N}\left(0, y^{-2}\right)$, one stil has under this new probabilistic model

$$\mathbb{E}\left[\exp\left(\lambda S_t - \frac{\lambda^2\sigma^2}{2}t\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda S_t - \frac{\lambda^2\sigma^2}{2}t\right)\middle|\lambda\right]\right] \leq 1.$$

Besides, a direct integration (over λ) gives a close form for the random variable

$$\mathbb{E}\left[\exp\left(\lambda S_t - \frac{\lambda^2\sigma^2}{2}t\right)\middle|\mathcal{F}_t\right],$$

whose expectation is then smaller than 1. Finally, the use of Markov inequality yields a deviation inequality. Corollary 12.5 of [De La Pena et al., 2009] could be applied directly to obtain a deviation inequality. However, in order to obtain an inequality that holds for all $t \in \mathbb{N}$, we follow [Abbasi-Yadkori et al., 2011] and consider randomly stopped super-martingales. This leads to the following result.

Lemma A.5. *Let $S_t = X_1 + \cdots + X_t$ be a sum of independent, σ^2-subgaussian increments. For all $y > 0$,*

$$\mathbb{P}\left(\exists t \in \mathbb{N}^* : \frac{S_t}{\sqrt{2\sigma^2 t}} > \sqrt{1+\frac{y}{t}}\sqrt{x + \frac{1}{2}\log\left(1+\frac{t}{y}\right)}\right) \leq e^{-x}$$

Proof of Lemma A.5 For every $\lambda \in \mathbb{R}$, the super-martingale W_t^λ satisfies $\mathbb{E}[W_t^\lambda] \leq 1$. Let τ be a stopping time with respect to $\mathcal{F}_t$. Using the same arguments as in Lemma 8 of [Abbasi-Yadkori et al., 2011], W_τ^λ is well defined and satisfies $\mathbb{E}[W_\tau^\lambda] \leq 1$.

Let $y > 0$ and assume that $\lambda \sim \mathcal{N}\left(0, y^{-2}\right)$. One still has $\mathbb{E}[W_\tau^\lambda] = \mathbb{E}[\mathbb{E}[W_\tau^\lambda|\lambda]] \leq 1$. Besides,

$$\begin{aligned}
\mathbb{E}[W_\tau^\lambda|\mathcal{F}_\tau] &= \int_{\mathbb{R}} \exp\left(\lambda S_\tau - \frac{\lambda^2\sigma^2}{2}\tau\right)\frac{y}{\sqrt{2\pi}}e^{-\frac{y^2}{2}\lambda^2}d\lambda \\
&= \frac{y}{\sqrt{2\pi}}\int_R \exp\left(-\frac{y^2+\sigma^2\tau}{2}\left[\left(\lambda - \frac{S_\tau}{y^2+\sigma^2\tau}\right)^2 - \frac{S_\tau^2}{(y^2+\sigma^2\tau)^2}\right]\right)d\lambda \\
&= \frac{y}{\sqrt{y^2+\sigma^2\tau}}\exp\left(\frac{S_\tau^2}{2(y^2+\sigma^2\tau)}\right).
\end{aligned}$$

Hence, $\mathbb{E}\left[\frac{y}{\sqrt{y^2+\sigma^2\tau}}\exp\left(\frac{S_\tau^2}{2(y^2+\sigma^2\tau)}\right)\right] \leq 1$ and Markov inequality yields

$$\begin{aligned}
\mathbb{P}\left(\frac{y}{\sqrt{y^2+\sigma^2\tau}}\exp\left(\frac{S_\tau^2}{2(y^2+\sigma^2\tau)}\right) > e^x\right) &\leq e^{-x} \\
\mathbb{P}\left(\frac{S_\tau}{\sqrt{2\sigma^2\tau}} > \sqrt{1+\frac{y^2}{\sigma^2\tau}}\sqrt{x+\frac{1}{2}\log\left(1+\frac{\sigma^2\tau}{y^2}\right)}\right) &\leq e^{-x}.
\end{aligned}$$

Now if τ is chosen to be the stopping time

$$\tau = \inf\left\{ t \in \mathbb{N} : \frac{S_t}{\sqrt{2\sigma^2 t}} > \sqrt{1+\frac{y^2}{\sigma^2 t}}\sqrt{x+\frac{1}{2}\log\left(1+\frac{\sigma^2 t}{y^2}\right)} \right\},$$

the probability in Lemma A.5 is upper bounded by

$$\mathbb{P}(\tau < +\infty) \leq \mathbb{P}\left(\frac{S_\tau}{\sqrt{2\sigma^2 \tau}} > \sqrt{1+\frac{y^2}{\sigma^2 \tau}}\sqrt{x+\frac{1}{2}\log\left(1+\frac{\sigma^2 \tau}{y^2}\right)} \right) \leq e^{-x},$$

which concludes the proof.

□

A.1.3 Comparison and generalization

At first sight, it may be difficult to compare Lemma A.1 and Lemma A.5. Writing Lemma A.1 with x replaced by $x + c\log(x)$ yields

$$\mathbb{P}\left(\exists t \in \mathbb{N}^* : \frac{S_t}{\sqrt{2\sigma^2 t}} > \sqrt{x + c\log(x) + \beta \log\log(et)} \right) \leq \underbrace{\sqrt{e}\,\zeta\Big(\beta\big(1-\frac{1}{2(x+c\log x)}\big)\Big)\Big(\frac{\sqrt{x+c\log x}}{2\sqrt{2}}+1\Big)^{\beta}\frac{1}{x^c}}_{F_{\beta,c}(x)} \exp(-x).$$

With a choice $c = \beta/2 + 1$ and $\beta > 1$, for x large enough, one has $F_{\beta,c}(x) \leq 1$. Hence, for every $\epsilon > 0$, for large values of x, one can write the following two inequalities, following from Lemma A.1 and Lemma A.5 respectively:

$$\mathbb{P}\left(\exists t \in \mathbb{N}^* : \frac{S_t}{\sqrt{2\sigma^2 t}} > \sqrt{x + 2\log(x) + (1+\epsilon)\log\log(et)} \right) \leq e^{-x} \tag{A.3}$$

$$\mathbb{P}\left(\exists t \in \mathbb{N}^* : \frac{S_t}{\sqrt{2\sigma^2 t}} > \sqrt{1+\frac{1}{t}}\sqrt{x+\frac{1}{2}\log\left(1+t\right)} \right) \leq e^{-x} \tag{A.4}$$

The dependency in x seems a bit worse in inequality (A.3) compared to (A.4). However, when considering the dependency in t, the deviation inequality (A.3) improves over (A.4) and is in some sense 'optimal' with respect to the Law of Iterated Logarithm in the Gaussian case, that states that

$$\limsup_{t\to\infty} \frac{S_t}{\sqrt{2\sigma^2 t}\sqrt{\log\log(t)}} = 1 \ a.s.$$

It would be interesting to be able to compare the deviation inequalities obtained with these two methods in different settings. However to prove the deviation inequality presented in Section A.3 and Section A.2 that were useful in this thesis in different contexts, we were not able to use both methods. Indeed, the method of mixtures can be generalized to deal with vector-valued martingales, as shown

by [Abbasi-Yadkori et al., 2011] with applications to linear contextual bandits, but the peeling-trick does not. Conversely, using a peeling trick, [Cappé et al., 2013] give an informational self-normalized deviation inequality useful in the analysis of KL-UCB, Bayes-UCB and Thompson Sampling, but it is not known yet how the method of mixtures could be applied in this setting.

Both methods can however be applied in a slightly more general setting to control a self-normalized process of the form $A_t^2/2B_t$, where A_t and $B_t \geq 0$ are such that, for all λ,

$$W_t^\lambda = \exp\left(\lambda A_t - \frac{\lambda^2}{2} B_t\right) \tag{A.5}$$

is a super-martingale. A straightforward adaptation of the proof of Lemma A.1 (using the slices $T_k^\eta = \{t \in \mathbb{N} : (1+\eta)^{k-1} \leq B_t \leq (1+\eta)^k\}$) and Lemma A.5 yields, for x small enough,

$$\begin{aligned}
\mathbb{P}\left(\exists t \in \mathbb{N}^* : \frac{A_t}{\sqrt{2B_t}} > \sqrt{x + 2\log(x) + (1+\epsilon)\log\log(eB_t)}\right) &\leq e^{-x} \\
\mathbb{P}\left(\exists t \in \mathbb{N}^* : \frac{A_t}{\sqrt{2B_t}} > \sqrt{1+\frac{1}{B_t}}\sqrt{x + \frac{1}{2}\log\left(1+B_t\right)}\right) &\leq e^{-x}
\end{aligned}$$

A.2 An informational deviation inequality

Let (X_i) be an i.i.d. sequence of random variables, whose log-moment generating function satisfy

$$\phi_{X_i}(\lambda) \leq \phi_Y(\lambda),$$

where the distribution of Y, ν_θ, belong to an exponential family and has mean μ. Then, if $S_t = X_1 + \cdots + X_t$, a natural super-martingale is

$$W_t^\lambda = \exp\left(\lambda S_t - t\phi_Y(\lambda)\right)$$

and the maximal inequality yields, for every $\lambda \in \mathbb{R}$, $u \in \mathbb{R}$,

$$\mathbb{P}\left(\exists s \in \mathbb{N}^* : \lambda S_s - s\phi_Y(\lambda) > u\right) \leq e^{-u}. \tag{A.6}$$

Lemma A.6 below (which corresponds to Lemma 1.9 in Chapter 1) can be obtained with a peeling-trick using the maximal inequality (A.6). Another key ingredient is the relationship between the log-moment generating function and the divergence d associated to the exponential family that follow from Lemma 1.4:

$$d(x,\mu) = \sup_{\lambda \in \mathbb{R}} \left(\lambda x - \phi_Y(\lambda)\right) \tag{A.7}$$

Lemma A.6. *Let (X_i) be a sequence of independent random variable such that $\phi_{X_i}(\lambda) \leq \phi_Y(\lambda)$ where $Y \sim \nu_\theta$ belongs to an exponential family with mean μ and associated divergence d. If $S_t = X_1 + \cdots + X_t$, one has*

$$\mathbb{P}\left(\exists s \in \{1,\ldots,t\} : s\, d^+\left(\frac{S_s}{s}, \mu\right) > \gamma\right) \leq e\left\lceil \gamma \log(t)\right\rceil \exp(-\gamma),$$

where $d^+(x,y) = d(x,y)\mathbb{1}_{(x<y)}$.

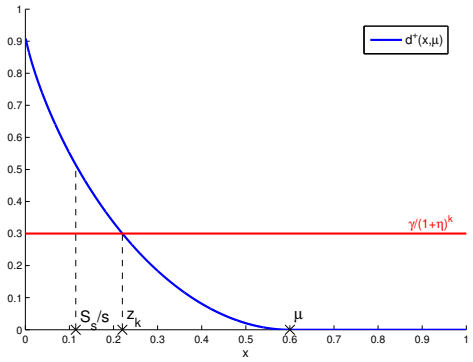

Figure A.1: Illustration of z_k in the Bernoulli case

Proof of Lemma A.6. Let $\eta > 0$ and for $k \geq 1$, let $T_k^\eta = \{s \in \mathbb{N} : (1+\eta)^{k-1} \leq s < (1+\eta)^k\}$. One has

$$\begin{aligned}\mathbb{P}\left(\exists s \in \{1,\dots,t\} : s\, d^+\left(\frac{S_s}{s},\mu\right) > \gamma\right) &\leq \sum_{k=1}^{\left\lceil \frac{\log(t)}{\log(1+\eta)}\right\rceil} \mathbb{P}\left(\exists s \in T_k^\eta : s\, d^+\left(\frac{S_s}{s},\mu\right) > \gamma\right) \\ &\leq \sum_{k=1}^{\left\lceil \frac{\log(t)}{\log(1+\eta)}\right\rceil} \mathbb{P}\Bigg(\underbrace{\exists s \in T_k^\eta : d^+\left(\frac{S_s}{s},\mu\right) > \frac{\gamma}{(1+\eta)^k}}_{A_k}\Bigg) \quad \text{(A.8)}\end{aligned}$$

The mapping $x \mapsto d^+(x,\mu)$ is nonincreasing. For k such that $\gamma/(1+\eta)^k > d(0,\mu)$, it follows that $\mathbb{P}(A_k) = 0$.

Now let k be such that $\gamma/(1+\eta)^k \leq d(0,\mu)$. There exists a unique $z_k < \mu$ such that $d^+(z_k,\mu) = \frac{\gamma}{(1+\eta)^k}$. On A_k, for $s \in T_k^\nu$, one has $S_s/s < z_k$ (see illustration in Figure A.1 in the Bernoulli case). From equality (A.7) there exists $\lambda_k \in \mathbb{R}$ such that

$$d^+(z_k,\mu) = \lambda_k z_k - \phi_Y(\lambda_k).$$

Moreover, it can be checked that, as $z_k < \mu$, $\lambda_k < 0$. Therefore, on A_k, for all $s \in T_k^\eta$,

$$\begin{aligned}\lambda_k \frac{S_s}{s} - \phi_Y(\lambda_k) &> \lambda_k z_k - \phi_Y(\lambda_k) \\ \lambda_k \frac{S_s}{s} - \phi_Y(\lambda_k) &> \frac{\gamma}{(1+\eta)^k} \\ \lambda_k S_s - \phi_Y(\lambda_k) s &> \frac{\gamma s}{(1+\eta)^k} \\ \lambda_k S_s - \phi_Y(\lambda_k) s &> \frac{\gamma}{(1+\eta)}.\end{aligned}$$

One can now upper bound $\mathbb{P}(A_k)$ using (A.6):

$$\mathbb{P}(A_k) \leq \mathbb{P}\left(\exists t \in T_k^\eta : \lambda_k S_s - \phi_Y(\lambda_k) s > \frac{\gamma}{1+\eta}\right) \leq \exp\left(-\frac{\gamma}{1+\eta}\right).$$

From (A.8), one gets

$$\mathbb{P}\left(\exists s \in \{1,\dots,t\} : s\, d^+\left(\frac{S_s}{s}, \mu\right) > \gamma\right) \leq \left\lceil \frac{\log(t)}{\log(1+\eta)} \right\rceil \exp\left(-\frac{\gamma}{1+\eta}\right)$$

Choosing η such that $1+\eta = \frac{\gamma}{\gamma-1}$ and using that $\log(\gamma/(1-\gamma)) \geq 1/\gamma$ yields the result.

□

Remark A.7. *It is not known whether the method of mixtures could be applied were, using the super-martingale W_t^λ. Indeed, assuming that $\lambda \sim \mathcal{N}\left(0, y^{-1}\right)$, there is no close form for*

$$\mathbb{E}[\exp\left(\lambda S_t - t\phi_Y(\lambda)\right)|\mathcal{F}_t].$$

Maybe a different 'prior distribution' for λ should be considered.

A.3 Deviation inequalities for vector-valued martingales

In Chapter 4, we present in Lemma 4.1 a confidence region for the parameter θ of a linear contextual bandit model. This confidence region follows from a deviation inequality for vector-valued martingales proposed by [Abbasi-Yadkori et al., 2011], that applies in the following more general framework.

For $(\mathcal{F}_t)$ a filtration, let $(X_t), (\eta_t)$ be two sequences of random variables such that $X_t \in \mathbb{R}^d$ is $\mathcal{F}_{t-1}$-measurable and $\eta_t \in \mathbb{R}$ is $\mathcal{F}_t$ mesurable and satisfies

$$\mathbb{E}[\eta_t|\mathcal{F}_{t-1}] = 0 \quad \text{and} \quad \forall \alpha \in \mathbb{R},\ \mathbb{E}[e^{\alpha\eta_t}|\mathcal{F}_{t-1}] \leq e^{\frac{\alpha^2\sigma^2}{2}}.$$

Let

$$S_t = \sum_{s=1}^{t} \eta_s X_s \quad \text{and} \quad V_t = \sum_{s=1}^{t} X_s X_s^T.$$

Then S_t is a martingale in $\mathbb{R}^d$ and for every $\lambda \in \mathbb{R}^d$, the sequence $(W_t^\lambda)_{t\in\mathbb{N}^*}$, with

$$W_t^\lambda = \exp\left(\lambda^T S_t - \frac{\sigma^2}{2}\|\lambda\|_{V_t}^2\right)$$

is a super-martingale, since

$$\begin{aligned}
\mathbb{E}\left[\frac{W_t^\lambda}{W_{t-1}^\lambda}\middle|\mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\exp\left(\lambda^T(\eta_t X_t) - \frac{\sigma^2}{2}(\|\lambda\|_{V_t}^2 - \|\lambda\|_{V_{t-1}}^2)\right)\middle|\mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\exp\left(\lambda^T(\eta_t X_t) - \frac{\sigma^2}{2}\lambda^T X_t X_t^T \lambda\right)\middle|\mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\exp((\lambda^T X_t)\eta_t)|\mathcal{F}_{t-1}\right]\exp\left(-\frac{\sigma^2}{2}(\lambda^T X_t)^2\right) \leq 1.
\end{aligned}$$

Applying the method of mixtures to this super-martingale, with a (vector-valued) Gaussian prior distribution for λ yields the following result.

Lemma A.8. *Under the above assumptions, for any matrix $V \in \mathcal{S}_n^{++}(\mathbb{R})$,*

$$\mathbb{P}\left(\exists t \in \mathbb{N} : \|S_t\|_{(V+V_t)^{-1}} > \sqrt{2\sigma^2\left(x + \frac{1}{2}\log\left(\frac{\det(V+V_t)}{\det(V)}\right)\right)}\right) \leq e^{-x}$$

Proof of Lemma A.8 Let τ be a stopping time. For all $\lambda \in \mathbb{R}^d$, $\mathbb{E}[W_\tau^\lambda] \leq 1$. Assuming that $\lambda \sim \mathcal{N}\left(0, \sigma^2 V\right)$, one still has $\mathbb{E}[W_\tau^\lambda] \leq 1$. Moreover, a direct integration and a bit of linear algebra (see [Abbasi-Yadkori et al., 2011]) yields

$$\mathbb{E}\left[W_\tau^\lambda | \mathcal{F}_\tau\right] = \sqrt{\frac{\det(V)}{\det(V+V_t)}}\exp\left(\frac{1}{2\sigma^2}\|S_t\|_{(V+V_t)^{-1}}\right).$$

Thus

$$\mathbb{E}\left[\sqrt{\frac{\det(V)}{\det(V+V_t)}}\exp\left(\frac{1}{2\sigma^2}\|S_t\|_{(V+V_t)^{-1}}\right)\right] \leq 1.$$

We conclude by using Markov inequality and choosing the stopping time τ as in the proof of Lemma A.5.

□

Remark A.9. *It is no known yet how to use a peeling-trick to obtain a result analogous to Lemma A.8. As W_t^λ is a super-martingale, for all $\lambda \in \mathbb{R}^d$*

$$\mathbb{P}\left(\forall t \in \mathbb{N}^*,\ \lambda^T S_t - \frac{\sigma^2}{2}\|\lambda\|_{V_t}^2 > u\right) \leq e^{-u},$$

but it is not clear how this inequality could be used (which 'slices' should be considered in the peeling, and how to choose interesting values of $\lambda \in \mathbb{R}^d$).

With the notation of this section, one could also use that, for any $x \in \mathbb{R}^d$, for any $\lambda \in \mathbb{R}$,

$$\tilde{W}_t^\lambda = \exp\left(\lambda x^T S_t - \frac{\sigma^2\lambda^2}{2}\|x\|_{V_t}^2\right)$$

is a super martingale. As explained in Section A.1.3, both the peeling-trick and mixtures method can be applied with $A_t = x^T S_t$ and $B_t = \sigma^2\|x\|_{V_t}^2$ in (A.5), leading to a deviation inequality of the form

$$\mathbb{P}\left(\exists t \in \mathbb{N} : x^T S_t > \|x\|_{V_t}\beta(t,\delta)\right) \leq \delta. \tag{A.9}$$

One can note that this inequality does not yield a deviation inequality of the form (4.9) for linear contextual bandits (see Section 4.3.3 in Chapter 4). Indeed, an inequality of the form (4.9) would follow from

$$\mathbb{P}\left(\exists t \in \mathbb{N} : x^T V_t^{-1} S_t > \|x\|_{V_t^{-1}}\beta(t,\delta)\right) \leq \delta,$$

which cannot be obtained using inequality (A.9).

Appendix B

Thompson Sampling for One-Dimensional Exponential Family Bandits

Abstract. *Thompson Sampling has been demonstrated in many complex bandit models, however the theoretical guarantees available for the parametric multi-armed bandit are still limited to the Bernoulli case. Here we extend them by proving asymptotic optimality of the algorithm using the Jeffreys prior for* 1*-dimensional exponential family bandits. Our proof builds on previous work, but also makes extensive use of closed forms for Kullback-Leibler divergence and Fisher information (through the Jeffreys prior) available in an exponential family. This allow us to give a finite time exponential concentration inequality for posterior distributions on exponential families that may be of interest in its own right. Moreover our analysis covers some distributions for which no optimistic algorithm has yet been proposed, including heavy-tailed exponential families.*

B.1 Introduction

K-armed bandit problems provide an elementary model for exploration-exploitation tradeoffs found at the heart of many online learning problems. In such problems, an agent is presented with K distributions (also called arms, or actions) $\{p_a\}_{a=1}^K$, from which she draws samples interpreted as rewards she wants to maximize. This objective induces a trade-off between choosing to sample a distribution that has already yielded high rewards, and choosing to sample a relatively unexplored distribution at the risk of loosing rewards in the short term. Here we make the assumption that the distributions, p_a, belong to a parametric family of distributions $\mathcal{P} = \{p(\cdot \mid \theta), \theta \in \Theta\}$ where $\Theta \subset \mathbb{R}$. The bandit model is described by a parameter $\theta_0 = (\theta_1, \dots, \theta_K)$ such that $p_a = p(\cdot \mid \theta_a)$. We introduce the mean function $\mu(\theta) = \mathbb{E}_{X \sim p(\cdot|\theta)}[X]$, and the optimal arm $\theta^* = \theta_{a^*}$ where $a^* = \operatorname{argmax}_a \mu(\theta_a)$.

An algorithm, $\mathcal{A}$, for a K-armed bandit problem is a (possibly randomised) method for choosing which arm a_t to sample from at time t, given a history of previous arm choices and obtained rewards, $\mathcal{H}_{t-1} := ((a_s, x_s))_{s=1}^{t-1}$: each reward x_s is drawn from the distribution p_{a_s}. The agent's goal is to design an algorithm with low regret:

$$\mathcal{R}(\mathcal{A}, t) = \mathcal{R}(\mathcal{A}, t)(\theta) := t\mu(\theta^*) - \mathbb{E}_{\mathcal{A}}\left[\sum_{s=1}^{t} x_s\right].$$

This quantity measures the expected performance of algorithm $\mathcal{A}$ compared to the expected performance

of an optimal algorithm given knowledge of the reward distributions, i.e. sampling always from the distribution with the highest expectation.

Since the early 2000s the "optimisim in the face of uncertainty" heuristic has been a popular approach to this problem, providing both simplicity of implementation and finite-time upper bounds on the regret (e.g. [Auer et al., 2002a, Cappé et al., 2013]). However in the last two years there has been renewed interest in the Thompson Sampling heuristic (TS). While this heuristic was first put forward to solve bandit problems eighty years ago in [Thompson, 1933], it was not until recently that theoretical analyses of its performance were achieved [Agrawal and Goyal, 2012, Agrawal and Goyal, 2013b, Kaufmann et al., 2012b, May et al., 2012]. In this paper we take a major step towards generalising these analyses to the same level of generality already achieved for "optimistic" algorithms.

Thompson Sampling Unlike optimistic algorithms which are often based on confidence intervals, the Thompson Sampling algorithm, denoted by $\mathcal{A}_{\pi_0}$ uses Bayesian tools and puts a prior distribution $\pi_{a,0} = \pi_0$ on each parameter θ_a. A posterior distribution, $\pi_{a,t}$, is then maintained according to the rewards observed in $\mathcal{H}_{t-1}$. At each time a sample $\theta_{a,t}$ is drawn from each posterior $\pi_{a,t}$ and then the algorithm chooses to sample $a_t = \arg\max_{a\in\{1,\dots,K\}}\{\mu(\theta_{a,t})\}$. Note that actions are sampled according to their posterior probabilities of being optimal.

Our contributions TS has proved to have impressive empirical performances, very close to those of state of the art algorithms such as DMED and KL-UCB [Kaufmann et al., 2012b, Honda and Takemura, 2010, Cappé et al., 2013]. Furthermore recent works [Kaufmann et al., 2012b, Agrawal and Goyal, 2013b] have shown that in the special case where each p_a is a Bernoulli distribution $\mathcal{B}(\theta_a)$, TS using a uniform prior over the arms is asymptotically optimal in the sense that it achieves the asymptotic lower bound on the regret provided by Lai and Robbins in [Lai and Robbins, 1985] (that holds for univariate parametric bandits). As explained in [Agrawal and Goyal, 2012, Agrawal and Goyal, 2013b], Thompson Sampling with uniform prior for Bernoulli rewards can be slightly adapted to deal with bounded rewards. However, there is no notion of asymptotic optimality for this non-parametric family of rewards. In this paper, we extend the optimality property that holds for Bernoulli distributions to more general families of parametric rewards, namely 1-dimensional exponential families if the algorithm uses the Jeffreys prior:

Theorem B.1. *Suppose that the reward distributions belong to a* 1*-dimensional canonical exponential family and let* π_J *denote the associated Jeffreys prior. Then,*

$$\lim_{T\to\infty} \frac{\mathcal{R}(\mathcal{A}_{\pi_J}, T)}{\ln T} = \sum_{a=1}^{K} \frac{\mu(\theta_{a^*}) - \mu(\theta_a)}{\mathrm{KL}(\theta_a, \theta_{a^*})}, \tag{B.1}$$

where $\mathrm{KL}(\theta, \theta') := KL(p_\theta, p'_\theta)$ *is the Kullback-Leibler divergence between* p_θ *and* p'_θ.

This theorem follows directly from Theorem B.2. In the proof of this result we provide in Theorem B.4 a finite-time, exponential concentration bound for posterior distributions of exponential family random variables, something that to the best of our knowledge is new to the literature and of interest in its own right. Our proof also exploits the connection between the Jeffreys prior, Fisher information and the Kullback-Leibler divergence in exponential families.

Related Work Another line of recent work has focused on distribution-independent bounds for Thompson Sampling. [Agrawal and Goyal, 2013b] establishes that $\mathcal{R}(\mathcal{A}_{\pi_U}, T) = O(\sqrt{KT\ln(T)})$ for Thompson Sampling for bounded rewards (with the classic uniform prior π_U on the underlying Bernoulli parameter). [Russo and Van Roy, 2014] go beyond the Bernoulli model, and give an upper bound on the Bayes risk (i.e. the regret averaged over the prior) independent of the prior distribution. For the parametric multi-armed bandit with K arms described above, their result states that the regret of Thompson Sampling using a prior π_0 is not too big when averaged over this same prior:

$$\mathbb{E}_{\theta\sim\pi_0^{\otimes K}}[\mathcal{R}(\mathcal{A}_{\pi_0}, T)(\theta)] \leq 4 + K + 4\sqrt{KT\log(T)}.$$

Building on the same ideas, [Bubeck and Liu, 2013] have improved this upper bound to $14\sqrt{KT}$. In our paper, we rather see the prior used by Thompson Sampling as a tool, and we want therefore to derive regret bounds for any given problem parametrized by θ that depend on this parameter.

[Russo and Van Roy, 2014] also use Thompson Sampling in more general models, like the linear bandit model. Their result is a bound on the Bayes risk that does not depend on the prior, whereas [Agrawal and Goyal, 2013b] gives a first bound on the regret in this model. Linear bandits consider a possibly infinite number of arms whose mean rewards are linearly related by a single, unknown coefficient vector. Once again, the analysis in [Agrawal and Goyal, 2013b] encounters the problem of describing the concentration of posterior distributions. However by using a conjugate normal prior, they can employ explicit concentration bounds available for Normal distributions to complete their argument.

Paper Structure In Section B.2 we describe important features of the one-dimensional canonical exponential families we consider, including closed-form expression for KL-divergences and the Jeffreys' prior. Section B.3 gives statements of the main results, and provides the proof of the regret bound. Section B.4 proves the posterior concentration result used in the proof of the regret bound.

B.2 Exponential Families and the Jeffreys Prior

A distribution is said to belong to a one-dimensional canonical exponential family if it has a density with respect to some reference measure ν of the form:

$$p(x \mid \theta) = A(x)\exp(T(x)\theta - F(\theta)), \tag{B.2}$$

where $\theta \in \Theta \subset \mathbb{R}$. T and A are some fixed functions that characterize the exponential family and $F(\theta) = \log\left(\int A(x)\exp\left[T(x)\theta\right]d\nu(x)\right)$. Θ is called the *parameter space*, $T(x)$ the *sufficient statistic*, and $F(\theta)$ the *normalisation function*. We make the classic assumption that F is twice differentiable with a continuous second derivative. It is well known [Wasserman, 2010] that:

$$\mathbb{E}_{X|\theta}(T(X)) = F'(\theta) \quad \text{and} \quad \mathrm{Var}_{X|\theta}[T(X)] = F''(\theta)$$

showing in particular that F is strictly convex. The mean function μ is differentiable and stricly increasing, since we can show that

$$\mu'(\theta) = \mathrm{Cov}_{X|\theta}(X, T(X)) > 0.$$

In particular, this shows that μ is one-to-one in θ.

KL-divergence in Exponential Families In an exponential family, a direct computation shows that the Kullback-Leibler divergence can be expressed as a Bregman divergence of the normalisation function, F:

$$\mathrm{KL}(\theta, \theta') = D_F^B(\theta', \theta) := F(\theta') - \left[F(\theta) + F'(\theta)(\theta' - \theta)\right]. \tag{B.3}$$

Jeffreys prior in Exponential Families In the Bayesian literature, a special "non-informative" prior, introduced by Jeffreys in [Jeffreys, 1946], is sometimes considered. This prior, called the Jeffreys prior, is invariant under re-parametrisation of the parameter space, and it can be shown to be proportional to the square-root of the Fisher information $I(\theta)$. In the special case of the canonical exponential family, the Fisher information takes the form $I(\theta) = F''(\theta)$, hence the Jeffreys prior for the model (B.2) is

$$\pi_J(\theta) \propto \sqrt{|F''(\theta)|}.$$

Under the Jeffreys prior, the posterior on θ after n observations is given by

$$p(\theta|y_1, \dots y_n) \propto \sqrt{F''(\theta)} \exp\left(\theta \sum_{i=1}^{n} T(y_i) - nF(\theta)\right) \tag{B.4}$$

When $\int_\Theta \sqrt{F''(\theta)}d\theta < +\infty$, the prior is called *proper*. However, stasticians often use priors which are not proper: the prior is called *improper* if $\int_\Theta \sqrt{F''(\theta)}d\theta = +\infty$ and any observation makes the corresponding posterior (B.4) integrable.

Some Intuition for choosing the Jeffreys Prior In the proof of our concentration result for posterior distributions (Theorem B.4) it will be crucial to lower bound the prior probability of an ϵ-sized KL-divergence ball around each of the parameters θ_a. Since the Fisher information $F''(\theta) = \lim_{\theta' \to \theta} K(\theta, \theta')/|\theta - \theta'|^2$, choosing a prior proportional to $F''(\theta)$ ensures that the prior measure of such balls are $\Omega(\sqrt{\epsilon})$.

Examples and Pseudocode Algorithm 7 presents pseudocode for Thompson Sampling with the Jeffreys prior for distributions parametrized by their natural parameter θ. But as the Jeffreys prior is invariant under reparametrization, if a distribution is parametrised by some parameter $\lambda \notin \theta$, the algorithm can use the Jeffreys prior $\propto \sqrt{I(\lambda)}$ on λ, drawing samples from the posterior on λ. Note that the posterior sampling step (in bold) is always tractable using, for example, a Hastings-Metropolis algorithm.

Some examples of common exponential family models are given in Figure B.1, together with the posterior distributions on the parameter λ that is used by TS with the Jeffreys prior. In addition to examples already studied in [Cappé et al., 2013] for which $T(x) = x$, we also give two examples of more general canonical exponential families, namely the Pareto distribution with known min value and unknown tail index λ, Pareto(x_m, λ), for which $T(x) = \log(x)$, and the Weibul distribution with known shape and unknown rate parameter, Weibull(k, λ), for which $T(x) = x^k$. These last two distributions are not covered even by the work in [Garivier and Cappé, 2011], and belong to the family of heavy-tailed distributions.

For the Bernoulli model, we note futher that the use of the Jeffreys prior is not covered by the previous analyses. These analyses make an extensive use of the uniform prior, through the fact that the coefficient of the Beta posteriors they consider have to be integers.

Algorithm 7 Thompson Sampling for Exponential Families with the Jeffreys prior

Require: F normalization function, T sufficient statistic, μ mean function
  for $t = 1 \dots K$ **do**
    Sample arm t and get rewards x_t
    $N_t = 1$, $S_t = T(x_t)$.
  end for
  for $t = K+1 \dots n$ **do**
    for $a = 1 \dots K$ **do**
      Sample $\theta_{a,t}$ **from** $\pi_{a,t} \propto \sqrt{F''(\theta)} \exp\left(\theta S_a - N_a F(\theta)\right)$
    end for
    Sample arm $A_t = \text{argmax}_a \mu(\theta_{a,t})$ and get reward x_t
    $S_{A_t} = S_{A_t} + T(x_t)$ $N_{A_t} = N_{A_t} + 1$
  end for

Name	Distribution	θ	Prior on λ	Posterior on λ
$\mathcal{B}(\lambda)$	$\lambda^x(1-\lambda)^{1-x}\delta_{0,1}$	$\log\left(\frac{\lambda}{1-\lambda}\right)$	Beta $\left(\frac{1}{2},\frac{1}{2}\right)$	Beta $\left(\frac{1}{2}+s,\frac{1}{2}+n-s\right)$
$\mathcal{N}(\lambda,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\lambda)^2}{2\sigma^2}}$	$\frac{\lambda}{\sigma^2}$	$\propto 1$	$\mathcal{N}\left(\frac{s}{n},\frac{\sigma^2}{n}\right)$
$\Gamma(k,\lambda)$	$\frac{\lambda^k}{\Gamma(k)}x^{k-1}e^{-\lambda x}1_{[0,+\infty[}(x)$	$-\lambda$	$\propto \frac{1}{\lambda}$	$\Gamma(kn,s)$
$\mathcal{P}(\lambda)$	$\frac{\lambda^x e^{-\lambda}}{x!}\delta_{\mathbb{N}}(x)$	$\log(\lambda)$	$\propto \frac{1}{\sqrt{\lambda}}$	$\Gamma\left(\frac{1}{2}+s,n\right)$
Pareto(x_m,λ)	$\frac{\lambda x_m^\lambda}{x^{\lambda+1}}1_{[x_m,+\infty[}(x)$	$-\lambda-1$	$\propto \frac{1}{\lambda}$	$\Gamma\left(n+1, s-n\log x_m\right)$
Weibull(k,λ)	$k\lambda(x\lambda)^{k-1}e^{-(\lambda x)^k}1_{[0,+\infty[}$	$-\lambda^k$	$\propto \frac{1}{\lambda^k}$	$\alpha\lambda^{(n-1)k}\exp(-\lambda^k s)$

Figure B.1: The posterior distribution after observations $y_1, \dots, y_n$ depends on n and $s = \sum_{i=1}^n T(y_i)$

B.3 Results and Proof of Regret Bound

An *exponential family K-armed bandit* is a K-armed bandit for which the reward distributions p_a are known to be elements of an exponential family of distributions $\mathcal{P}(\Theta)$. We denote by p_{θ_a} the distribution of arm a and its mean by $\mu_a = \mu(\theta_a)$.

Theorem B.2 (Regret Bound). *Assume that $\mu_1 > \mu_a$ for all $a \neq 1$, and that $\pi_{a,0}$ is taken to be the Jeffreys prior over Θ. Then for every $\epsilon > 0$ there exists a constant $\mathcal{C}(\epsilon, \mathcal{P})$ depending on ϵ and on the problem $\mathcal{P}$ such that the regret of Thompson Sampling using the Jeffreys prior satisfies*

$$\mathcal{R}(\mathcal{A}_{\pi_J}, T) \leq \frac{1+\epsilon}{1-\epsilon}\left(\sum_{a=2}^{K}\frac{(\mu_1-\mu_a)}{\mathrm{KL}(\theta_a,\theta_1)}\right)\ln(T) + \mathcal{C}(\epsilon,\mathcal{P}).$$

Proof: We give here the main argument of the proof of the regret bound, which proceed by bounding the expected number of draws of any suboptimal arm. Along the way we shall state concentration results whose proofs are postponed to later sections.

Step 0: Notation We denote by $y_{a,s}$ the s-th observation of arm a and by $N_{a,t}$ the number of times arm a is chosen up to time t. $(y_{a,s})_{s\geq 1}$ is i.i.d. with distribution p_{θ_a}. Let $Y_a^u := (y_{a,s})_{1\leq s\leq u}$ be the vector of

first u observations from arm a. $Y_{a,t} := Y_a^{N_{a,t}}$ is therefore the vector of observations from arm a available at the beginning of round t. Recall that $\pi_{a,t}$, respectively $\pi_{a,0}$, is the posterior, respectively the prior, on θ_a at round t of the algorithm.

We define $L(\theta)$ to be such that $\mathbb{P}_{Y\sim p(|\theta)}(p(Y|\theta) \geq L(\theta)) \geq \frac{1}{2}$. Observations from arm a such that $p(y_{a,s}|\theta) \geq L(\theta_a)$ can therefore be seen as likely observations. For any $\delta_a > 0$, we introduce the event $\tilde{E}_{a,t} = \tilde{E}_{a,t}(\delta_a)$:

$$\tilde{E}_{a,t} = \left(\exists 1 \leq s' \leq N_{a,t} : p(y_{a,s'}|\theta_a) \geq L(\theta_a), \left| \frac{\sum_{s=1, s\neq s'}^{N_{a,t}} T(y_{a,s})}{N_{a,t} - 1} - F'(\theta_a) \right| \leq \delta_a \right). \tag{B.5}$$

For all $a \neq 1$ and Δ_a such that $\mu_a < \mu_a + \Delta_a < \mu_1$, we introduce

$$E_{a,t}^{\theta} = E_{a,t}^{\theta}(\Delta_a) := \left(\mu\left(\theta_{a,t}\right) \leq \mu_a + \Delta_a\right).$$

On $\tilde{E}_{a,t}$, the empirical sufficient statistic of arm a at round t is well concentrated around its mean and a 'likely' realization of arm a has been observed. On $E_{a,t}^{\theta}$, the mean of the distribution with parameter $\theta_{a,t}$ does not exceed by much the true mean, μ_a. δ_a and Δ_a will be carefully chosen at the end of the proof.

Step 1: Decomposition The idea of the proof is to decompose the probability of playing a suboptimal arm using the events given in Step 0, and that $\mathbb{E}[N_{a,T}] = \sum_{t=1}^{T} \mathbb{P}\left(a_t = a\right)$:

$$\mathbb{E}\left[N_{a,T}\right] = \underbrace{\sum_{t=1}^{T} \mathbb{P}\left(a_t = a, \tilde{E}_{a,t}, E_{a,t}^{\theta}\right)}_{(A)} + \underbrace{\sum_{t=1}^{T} \mathbb{P}\left(a_t = a, \tilde{E}_{a,t}, (E_{a,t}^{\theta})^c\right)}_{(B)} + \underbrace{\sum_{t=1}^{T} \mathbb{P}\left(a_t = a, \tilde{E}_{a,t}^c\right)}_{(C)}.$$

where E^c denotes the complement of event E. Term (C) is controlled by the concentration of the empirical sufficient statistic, and (B) is controlled by the tail probabilities of the posterior distribution. We give the needed concentration results in Step 2. When conditioned on the event that the optimal arm is played at least polynomially often, term (A) can be decomposed further, and then controled by the results from Step 2. Step 3 proves that the optimal arm is played this many times.

Step 2: Concentration Results We state here the two concentration results that are necessary to evaluate the probability of the above events.

Lemma B.3. *Let (y_s) be an i.i.d sequence of distribution $p(\cdot \mid \theta)$ and $\delta > 0$. Then*

$$\mathbb{P}\left(\left|\frac{1}{u}\sum_{s=1}^{u}[T(y_s) - F'(\theta)]\right| \geq \delta\right) \leq 2e^{-u\tilde{K}(\theta,\delta)},$$

where $\tilde{K}(\theta, \delta) = \min(K(\theta+g(\delta), \theta), K(\theta-h(\delta), \theta))$, with $g(\delta) > 0$ defined by $F'(\theta+g(\delta)) = F'(\theta)+\delta$ and $h(\delta) > 0$ defined by $F'(\theta - h(\delta)) = F'(\theta) - \delta$.

The two following inequalities that will be useful in the sequel can easily be deduced from Lemma B.3. Their proof is gathered in Appendix B.6 with that of Lemma B.3. For any arm a, for any $b \in]0, 1[$,

$$\sum_{t=1}^{T} \mathbb{P}(a_t = a, (\tilde{E}_{a,t}(\delta_a))^c) \leq \sum_{t=1}^{\infty} \left(\frac{1}{2}\right)^t + \sum_{t=1}^{\infty} 2te^{-(t-1)\tilde{K}(\theta_a,\delta_a)} \tag{B.6}$$

$$\sum_{t=1}^{T} \mathbb{P}((\tilde{E}_{a,t}(\delta_a))^c \cap N_{a,t} > t^b) \leq \sum_{t=1}^{\infty} t\left(\frac{1}{2}\right)^{t^b} + \sum_{t=1}^{\infty} 2t^2 e^{-(t^b-1)\tilde{K}(\theta_a,\delta_a)}, \tag{B.7}$$

The second result tells us that concentration of the empirical sufficient statistic around its mean implies concentration of the posterior distribution around the true parameter:

Theorem B.4 (Posterior Concentration). *Let $\pi_{a,0}$ be the Jeffreys prior. There exists constants $C_{1,a} = C_1(F,\theta_a) > 0$, $C_{2,a} = C_2(F,\theta_a,\Delta_a) > 0$, and $N(\theta_a,F)$ s.t., $\forall N_{a,t} \geq N(\theta_a,F)$,*

$$\mathbb{1}_{\tilde{E}_{a,t}}\mathbb{P}\big(\mu(\theta_{a,t}) > \mu(\theta_a) + \Delta_a | Y_{a,t}\big) \leq C_{1,a}e^{-(N_{a,t}-1)(1-\delta_a C_{2,a})\mathrm{KL}(\theta_a,\mu^{-1}(\mu_a+\Delta_a))+\ln(N_{a,t})}$$

whenever $\delta_a < 1$ and Δ_a are such that $1 - \delta_a C_{2,a}(\Delta_a) > 0$.

Step 3: Lower Bound the Number of Optimal Arm Plays with High Probability The main difficulty adressed in previous regret analyses for Thompson Sampling is the control of the number of draws of the optimal arm. We provide this control in the form of Proposition B.5 which is adapted from Proposition 1 in [Kaufmann et al., 2012b]. The proof of this result, an outline of which is given in Appendix B.9, explores in depth the randomised nature of Thompson Sampling. In particular, we show that the proof in [Kaufmann et al., 2012b] can be significantly simplified, but at the expense of no longer being able to describe the constant C_b explicitly:

Proposition B.5. *$\forall b \in (0,1)$, $\exists C_b(\pi,\mu_1,\mu_2,K) < \infty$ such that $\sum_{t=1}^{\infty}\mathbb{P}\left(N_{1,t} \leq t^b\right) \leq C_b$.*

Step 4: Bounding the Terms of the Decomposition Now we bound the terms of the decomposition as discussed in Step 1: An upper bound on term (C) is given in (B.6), whereas a bound on term (B) follows from Lemma B.6 below. Although the proof of this lemma is standard, and bears a strong similarity to Lemma 3 of [Agrawal and Goyal, 2013b], we provide it in Appendix B.8 for the sake of completeness.

Lemma B.6. *For all actions a and for all $\epsilon > 0$, $\exists\ N_\epsilon = N_\epsilon(\delta_a,\Delta_a,\theta_a) > 0$ such that*

$$(B) \leq \left[(1-\epsilon)(1-\delta_a C_{2,a})\mathrm{KL}(\theta_a,\mu^{-1}(\mu_a+\Delta_a))\right]^{-1}\ln(T) + \max\{N_\epsilon, N(\theta_a,F)\} + 1.$$

where $N_\epsilon = N_\epsilon(\delta_a,\Delta_a,\theta_a)$ is the smallest integer such that for all $n \geq N_\epsilon$

$$(n-1)^{-1}\ln(C_{1,a}n) < \epsilon(1-\delta_a C_{2,a})\mathrm{KL}(\theta_a,\mu^{-1}(\mu_a+\Delta_a)),$$

and $N(\theta_a,F)$ is the constant from Theorem B.4.

When we have seen enough observations on the optimal arm, term (A) also becomes a result about the concentration of the posterior and the empirical sufficient statistic, but this time for the optimal arm:

$$\begin{aligned}(A) &\leq \sum_{t=1}^{T}\mathbb{P}\left(a_t = a, \tilde{E}_{a,t}, E^{\theta}_{a,t}, N_{1,t} > t^b\right) + C_b \leq \sum_{t=1}^{T}\mathbb{P}\left(\mu(\theta_{1,t}) \leq \mu_1 - \Delta'_a, N_{1,t} > t^b\right) + C_b \\ &\leq \underbrace{\sum_{t=1}^{T}\mathbb{P}\left(\mu(\theta_{1,t}) \leq \mu_1 - \Delta'_a, \tilde{E}_{1,t}(\delta_1), N_{1,t} > t^b\right)}_{B'} + \underbrace{\sum_{t=1}^{T}\mathbb{P}\left(\tilde{E}^c_{1,t}(\delta_1) \cap N_{1,t} > t^b\right)}_{C'} + C_b \end{aligned} \tag{B.8}$$

where $\Delta'_a = \mu_1 - \mu_a - \Delta_a$ and $\delta_1 > 0$ remains to be chosen. The first inequality comes from Proposition B.5, and the second inequality comes from the following fact: if arm 1 is not chosen and arm a is such that $\mu(\theta_{a,t}) \leq \mu_a + \Delta_a$, then $\mu(\theta_{1,t}) \leq \mu_a + \Delta_a$. A bound on term (C') is given in (B.7) for $a = 1$ and δ_1.

In Theorem B.4, we bound the conditional probability that $\mu(\theta_{a,t})$ exceed the true mean. Following the same lines, we can also show that

$$\mathbb{P}\left(\mu(\theta_{1,t}) \leq \mu_1 - \Delta_a' | Y_{1,t}\right) \mathbb{1}_{\tilde{E}_{1,t}(\delta_1)} \leq C_{1,1} e^{-(N_{1,t}-1)(1-\delta_1 C_{2,1})\mathrm{KL}(\theta_1, \mu^{-1}(\mu_1-\Delta_a'))+\ln(N_{1,t})}.$$

For any $\Delta_a' > 0$, one can choose δ_1 such that $1-\delta_1 C_{1,1} > 0$. Then, with $N = N(\mathcal{P})$ such that the function $u \mapsto e^{-(u-1)(1-\delta_1 C_{2,1})\mathrm{KL}(\theta_1, \mu^{-1}(\mu_1-\Delta_a'))+\ln u}$ is decreasing for $u \geq N$, (B') is bounded by

$$N^{1/b} + \sum_{t=N^{1/b}+1}^{\infty} C_{1,1} e^{-(t^b-1)(1-\delta_1 C_{2,1})\mathrm{KL}(\theta_1, \mu^{-1}(\mu_1-\Delta_a'))+\ln(t^b)} < \infty.$$

Step 4: Choosing the Values δ_a and ϵ_a So far, we have shown that for any $\epsilon > 0$ and for any choice of $\delta_a > 0$ and $0 < \Delta_a < \mu_1 - \mu_a$ such that $1 - \delta_a C_{2,a} > 0$, there exists a constant $\mathcal{C}(\delta_a, \Delta_a, \epsilon, \mathcal{P})$ such that

$$\mathbb{E}[N_{a,T}] \leq \frac{\ln(T)}{(1-\delta_a C_{2,a}) K(\theta_a, \mu^{-1}(\mu_a + \Delta_a))(1-\epsilon)} + \mathcal{C}(\delta_a, \Delta_a, \epsilon, \mathcal{P})$$

The constant is of course increasing (dramatically) when δ_a goes to zero, Δ_a to $\mu_1 - \mu_a$, or ϵ to zero. But one can choose Δ_a close enough to $\mu_1 - \mu_a$ and δ_a small enough, such that

$$(1 - C_{2,a}(\Delta_a)\delta_a)\mathrm{KL}(\theta_a, \mu^{-1}(\mu_a + \Delta_a)) \geq \frac{\mathrm{KL}(\theta_a, \theta_1)}{(1+\epsilon)},$$

and this choice leads to

$$\mathbb{E}[N_{a,T}] \leq \frac{1+\epsilon}{1-\epsilon} \frac{\ln(T)}{\mathrm{KL}(\theta_a, \theta_1)} + \mathcal{C}(\delta_a, \Delta_a, \epsilon, \mathcal{P}).$$

Using that $\mathcal{R}(\mathcal{A}, T) = \sum_{a=2}^{K} (\mu_1 - \mu_a)\mathbb{E}_{\mathcal{A}}[N_{a,T}]$ for any algorithm $\mathcal{A}$ concludes the proof.

□

B.4 Posterior Concentration: Proof of Theorem B.4

For ease of notation, we drop the subscript a and let (y_s) be an i.i.d. sequence of distribution p_θ, with mean $\mu = \mu(\theta)$. Furthermore, by conditioning on the value of N_s, it is enough to bound $\mathbb{1}_{\tilde{E}_u} \mathbb{P}\left(\mu(\theta_u) \geq \mu + \Delta | Y^u\right)$ where $Y^u = (y_s)_{1 \leq s \leq u}$ and

$$\tilde{E}_u = \left(\exists 1 \leq s' \leq u : p(y_{s'}|\theta) \geq L(\theta), \left| \frac{\sum_{s=1, s\neq s'}^{u} T(y_s)}{u-1} - F'(\theta) \right| \leq \delta \right).$$

Step 1: Extracting a Kullback-Leibler Rate The argument rests on the following Lemma, whose proof can be found in Appendix B.7

Lemma B.7. *Let $\tilde{E}_u$ be the event defined by (B.5), and introduce $\Theta_{\theta,\Delta} := \{\theta' \in \Theta : \mu(\theta') \geq \mu(\theta) + \Delta\}$. The following inequality holds:*

$$\mathbb{1}_{\tilde{E}_u} \mathbb{P}\left(\mu(\theta_u) \geq \mu + \Delta | Y^u\right) \leq \frac{\int_{\theta' \in \Theta_{\theta,\Delta}} e^{-(u-1)(K[\theta,\theta'] - \delta|\theta-\theta'|)} \pi(\theta'|y_{s'}) d\theta'}{\int_{\theta' \in \Theta} e^{-(u-1)(K[\theta,\theta'] + \delta|\theta-\theta'|)} \pi(\theta'|y_{s'}) d\theta'}, \tag{B.9}$$

with $s' = \inf\{s \in \mathbb{N} : p(y_s|\theta) \geq L(\theta)\}$.

Step 2: Upper bounding the numerator of (B.9) We first note that on $\Theta_{\theta,\Delta}$ the leading term in the exponential is $K(\theta,\theta')$. Indeed, from (B.3) we know that

$$K(\theta,\theta')/|\theta-\theta'| = \left|F'(\theta) - (F(\theta)-F(\theta'))/(\theta-\theta')\right|$$

which, by strict convexity of F, is strictly increasing in $|\theta-\theta'|$ for any fixed θ. Now since μ is one-to-one and continuous, $\Theta^c_{\theta,\Delta}$ is an interval whose interior contains θ, and hence, on $\Theta_{\theta,\Delta}$,

$$\frac{K(\theta,\theta')}{|\theta-\theta'|} \geq \frac{F(\mu^{-1}(\mu+\Delta))-F(\theta)}{\mu^{-1}(\mu+\Delta)-\theta} - F'(\theta) := (C_2(F,\theta,\Delta))^{-1} > 0.$$

So for δ such that $1-\delta C_2 > 0$ we can bound the numerator of (B.9) by:

$$\begin{aligned}\int_{\theta'\in\Theta_{\theta,\Delta}} e^{-(u-1)(K(\theta,\theta')-\delta|\theta-\theta'|)}\pi(\theta'|y_{s'})d\theta' &\leq \int_{\theta'\in\Theta_{\theta,\Delta}} e^{-(u-1)K(\theta,\theta')(1-\delta C_2)}\pi(\theta'|y_{s'})d\theta' \\ \leq e^{-(u-1)(1-\delta C_2)\mathrm{KL}(\theta,\mu^{-1}(\mu+\Delta))}\int_{\Theta_{\theta,\Delta}}\pi(\theta'|y_{s'})d\theta' &\leq e^{-(u-1)(1-\delta C_2)\mathrm{KL}(\theta,\mu^{-1}(\mu+\Delta))} \qquad \text{(B.10)}\end{aligned}$$

where we have used that $\pi(\cdot|y_{s'})$ is a probability distribution, and that, since μ is increasing, $\mathrm{KL}(\theta,\mu^{-1}(\mu+\Delta)) = \inf_{\theta'\in\Theta_{\theta,\Delta}} K(\theta,\theta')$.

Step 3: Lower bounding the denominator of (B.9) To lower bound the denominator, we reduce the integral on the whole space Θ to a KL-ball, and use the structure of the prior to lower bound the measure of that KL-ball under the posterior obtained with the well-chosen observation $y_{s'}$. We introduce the following notation for KL balls: for any $x\in\Theta,\ \epsilon>0$, we define

$$B_\epsilon(x) := \left\{\theta'\in\Theta : K(x,\theta')\leq\epsilon\right\}.$$

We have $\frac{K(\theta,\theta')}{(\theta-\theta')^2} \to F''(\theta) \neq 0$ (since F is strictly convex). Therefore, there exists $N_1(\theta,F)$ such that for $u\geq N_1(\theta,F)$, on $B_{\frac{1}{u^2}}(\theta)$,

$$|\theta-\theta'| \leq \sqrt{2K(\theta,\theta')/F''(\theta)}.$$

Using this inequality we can then bound the denominator of (B.9) whenever $u\geq N_1(\theta,F)$ and $\delta<1$:

$$\begin{aligned}\int_{\theta'\in\Theta} e^{-(u-1)(K(\theta,\theta')+\delta|\theta-\theta'|)}\pi(\theta'|y_{s'})d\theta' &\geq \int_{\theta'\in B_{1/u^2}(\theta)} e^{-(u-1)(K(\theta,\theta')+\delta|\theta-\theta'|)}\pi(\theta'|y_{s'})d\theta' \\ \geq \int_{\theta'\in B_{1/u^2}(\theta)} e^{-(u-1)\left(K(\theta,\theta')+\delta\sqrt{\frac{2K(\theta,\theta')}{F''(\theta)}}\right)}\pi(\theta'|y_{s'})d\theta' &\geq \pi\left(B_{1/u^2}(\theta)|y_{s'}\right)e^{-\left(1+\sqrt{\frac{2}{F''(\theta)}}\right)}. \qquad \text{(B.11)}\end{aligned}$$

Finally we turn our attention to the quantity

$$\pi\left(B_{1/u^2}(\theta)|y_{s'}\right) = \frac{\int_{B_{1/u^2}(\theta)} p(y'_s|\theta')\pi_0(\theta')d\theta'}{\int_\Theta p(y'_s|\theta')\pi_0(\theta')d\theta'} = \frac{\int_{B_{1/u^2}(\theta)} p(y'_s|\theta')\sqrt{F''(\theta')}d\theta'}{\int_\Theta p(y'_s|\theta')\sqrt{F''(\theta')}d\theta'}. \qquad \text{(B.12)}$$

Now since the KL divergence is convex in the second argument, we can write $B_{1/u^2}(\theta) = (a,b)$. So, from the convexity of F we deduce that

$$\begin{aligned}\frac{1}{u^2} = K(\theta,b) = F(b) - \left[F(\theta)+(b-\theta)F'(\theta)\right] &= (b-\theta)\left[\frac{F(b)-F(\theta)}{(b-\theta)} - F'(\theta)\right] \\ \leq (b-\theta)\left[F'(b)-F'(\theta)\right] \leq (b-a)\left[F'(b)-F'(\theta)\right] &\leq (b-a)\left[F'(b)-F'(a)\right].\end{aligned}$$

As $p(y \mid \theta) \to 0$ as $y \to \pm\infty$, the set $\mathcal{C}(\theta) = \{y : p(y \mid \theta) \geq L(\theta)\}$ is compact. The map $y \mapsto \int_\Theta p(y|\theta')\sqrt{F''(\theta')}d\theta' < \infty$ is continuous on the compact $\mathcal{C}(\theta)$. Thus, it follows that

$$L'(\theta) = L'(\theta, F) := \sup_{y:p(y|\theta)>L(\theta)} \left\{ \int_\Theta p(y|\theta')\sqrt{F''(\theta')}d\theta' \right\} < \infty$$

is an upper bound on the denominator of (B.12).

Now by the continuity of F'', and the continuity of $(y, \theta) \mapsto p(y|\theta)$ in both coordinates, there exists an $N_2(\theta, F)$ such that for all $u \geq N_2(\theta, F)$

$$F''(\theta) \geq \frac{1}{2}\frac{F'(b) - F'(a)}{b-a} \text{ and } \left(p(y|\theta')\sqrt{F''(\theta')} \geq \frac{L(\theta)}{2}\sqrt{F''(\theta)},\ \forall \theta' \in B_{1/u^2}(\theta), y \in \mathcal{C}(\theta) \right).$$

Finally, for $u \geq N_2(\theta, F)$, we have a lower bound on the numerator of (B.12):

$$\int_{B_{1/u^2}(\theta)} p(y'_s|\theta')\sqrt{F''(\theta')}d\theta' \geq \frac{L(\theta)}{2}\sqrt{F''(\theta)}\int_a^b d\theta' = \frac{L(\theta)}{2}\sqrt{(F'(b) - F'(a))\,(b-a)} \geq \frac{L(\theta)}{2u}$$

Puting everything together, we get that there exist constants $C_2 = C_2(F, \theta, \Delta)$ and $N(\theta, F) = \max\{N_1, N_2\}$ such that for every $\delta < 1$ satisfying $1 - \delta C_2 > 0$, and for every $u \geq N$, one has

$$\mathbb{1}_{\tilde{E}_u}\mathbb{P}(\mu(\theta_u) \geq \mu(\theta) + \Delta | Y_u) \leq \frac{2e^{1+\sqrt{\frac{2}{F''(\theta)}}}L'(\theta)u}{L(\theta)} e^{-(u-1)(1-\delta C_2)\mathrm{KL}(\theta, \mu^{-1}(\mu+\Delta))}.$$

Remark B.8. *Note that when the prior is proper we do not need to introduce the observation $y_{s'}$, which significantly simplifies the argument. Indeed in this case, in (B.10) we can use π_0 in place of $\pi(\cdot|y_{s'})$ which is already a probability distribution. In particular, the quantity (B.12) is replaced by $\pi_0\left(B_{1/u^2}(\theta)\right)$, and so the constants L and L' are not needed.*

B.5 Conclusion

We have shown that choosing to use the Jeffreys prior in Thompson Sampling leads to an asymptotically optimal algorithm for bandit models whose rewards belong to a 1-dimensional canonical exponential family. The cornerstone of our proof is a finite time concentration bound for posterior distributions in exponential families, which, to the best of our knowledge, is new to the literature. With this result we built on previous analyses and avoided Bernoulli-specific arguments. Thompson Sampling with Jeffreys prior is now a provably competitive alternative to KL-UCB for exponential family bandits. Moreover our proof holds for slightly more general problems than those for which KL-UCB is provably optimal, including some heavy-tailed exponential family bandits.

Our arguments are potentially generalisable. Notably generalising to n-dimensional exponential family bandits requires only generalising Lemma B.3 and Step 3 in the proof of Theorem B.4. Our result is asymptotic, but the only stage where the constants are not explicitly derivable from knowledge of F, T, and θ_0 is in Lemma B.9. Future work will investigate these open problems. Another possible future direction lies the optimal choice of prior distribution. Our theoretical guarantees only hold for Jeffreys' prior, but a careful examination of our proof shows that the important property is to have, for every θ_a,

$$-\ln\left(\int_{(\theta':\mathrm{KL}(\theta_a,\theta')\leq n^{-2})} \pi_0(\theta')d\theta'\right) = o(n),$$

which could hold for prior distributions other than the Jeffreys prior.

B.6 Concentration of the Sufficient Statistics: Proof of Lemma B.3, and Inequalities (B.6) and (B.7)

Proof of Lemma B.3. The proof of Lemma B.3 follows from the classical Cramér-Chenoff technique (see [Boucheron et al., 2013]). For any $\lambda > 0$.

$$\begin{aligned} A :=& \mathbb{P}\left(\frac{1}{u}\sum_{i=1}^{u}[T(y_i) - F'(\theta)] \geq \delta\right) = \mathbb{P}\left(e^{\lambda\left(\sum_{i=1}^{u}[T(y_i)-F'(\theta)]\right)} \geq e^{\lambda u}\delta\right) \\ \leq& e^{-\lambda u\delta}\mathbb{E}\left[e^{\lambda\left(\sum_{i=1}^{u}[T(y_i)-F'(\theta)]\right)}\right] = e^{-u(\delta\lambda-\phi_a(\lambda))} \end{aligned}$$

where we have used the Markov inequality, and where

$$\phi_a(\lambda) := \ln \mathbb{E}_{X|\theta}\left[e^{\lambda(T(X)-F'(\theta))}\right] = F(\theta+\lambda) - F(\theta) - \lambda F'(\theta).$$

Now we optimize in λ by choosing $\lambda > 0$ that maximizes

$$\delta\lambda - \phi_a(\lambda) = \lambda(\delta + F'(\theta)) - F(\theta+\lambda) + F(\theta) := f(\lambda).$$

$f(\lambda)$ is differentiable in λ and its minimum, λ^*, satisfies $f'(\lambda^*) = 0$ i.e.

$$F'(\theta+\lambda^*) = \delta + F'(\theta).$$

(Note that $\lambda^* > 0$ since F' is increasing). Finally, we get

$$A \leq e^{-u((\delta+F'(\theta))\lambda^* - F(\theta+\lambda^*)+F(\theta))} = e^{-u(F'(\theta+\lambda^*)\lambda^* - F(\theta+\lambda^*)+F(\theta))} = e^{-uK(\theta+\lambda^*,\theta)}.$$

The same reasoning leads to the upper bound

$$\mathbb{P}\left(\frac{1}{u}\sum_{s=1}^{u}[T(y_s) - F'(\theta)] \leq -\delta\right) \leq e^{-u\mathrm{KL}(\theta-\nu^*,\theta)},$$

where ν^* is such that $F'(\theta-\nu^*) = F'(\theta) - \delta$.

□

For the proof of inequalities (B.6) and (B.7), we intoduce the notation $Y^u_{a,s'} = Y^s_a \backslash \{y_{a,s}\}$ (the first u observations of arms a exept observation $y_{a,s'}$). First note that we have $\tilde{E}^c_{a,t} \subseteq B_{a,N_{a,t}} \bigcup D_{a,N_{a,t}}$, with

$$\begin{aligned} B_{a,s} &= \left(\forall s' \in [1,s], p(y_{a,s'}|\theta_a) \leq L(\theta_a)\right), \\ D_{a,s} &= \left(\exists s' \in \{1,\dots s\} : \left|\frac{1}{s-1}\sum_{k=1,k\neq s'}^{s}(T(y_{a,k}) - F'(\theta_a))\right| \geq \delta_a\right). \end{aligned}$$

Indeed, we have used that for two sequences of event $F_{s'}$ and $G_{s'}$,

$$\left(\bigcup_{s'=1}^{s} F_{s'} \cap G_{s'}\right)^c = \bigcap_{s'\leq s} F^c_{s'} \cup G^c_{s'} \subset \bigcap_{s'\leq s} F^c_{s'} \cup \left(\bigcup_{s''\leq s} G^c_{s''}\right) = \left(\bigcap_{s'\leq s} F^c_{s'}\right) \cup \left(\bigcup_{s'\leq s} G^c_{s'}\right).$$

One then has

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{P}(a_t=a,\tilde{E}^c_{a,t}(\delta)) &\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=1}^{t}\mathbb{1}_{(a_t=a,N_{a,t}=s)}(\mathbb{1}_{B_{a,s}}+\mathbb{1}_{D_{a,s}})\right]\\
&\leq \mathbb{E}\left[\sum_{s=1}^{T}\mathbb{1}_{B_{a,s}}\right]+\mathbb{E}\left[\sum_{s=1}^{T}\mathbb{1}_{D_{a,s}}\right]\\
&\leq \sum_{s=1}^{T}\mathbb{P}\left(p(y_{a,1}|\theta_a)\leq L(\theta_a)\right)^s+\sum_{s=1}^{T}\sum_{s'=1}^{s}\mathbb{P}\left(\left|\frac{1}{s-1}\sum_{k=1,k\neq s'}^{s}(T(y_{a,k})-F'(\theta_a))\right|\geq\delta_a\right)\\
&\leq \sum_{s=1}^{\infty}\left(\frac{1}{2}\right)^s+\sum_{s=1}^{\infty}se^{-(s-1)\tilde{K}(\theta_a,\delta_a)},
\end{aligned}$$

where we use that the definition of $L(\theta)$ gives $\mathbb{P}\left(p(y_{a,1}|\theta_a)\leq L(\theta_a)\right)\leq\frac{1}{2}$. This leads to inequality (B.6). To proof (B.7), we write:

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{P}(\tilde{E}_{a,t}(\delta_a)^c\cap N_{a,t}>t^b) &\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{s=t^b}^{t}\mathbb{1}_{N_{a,t}=s}(\mathbb{1}_{B_{a,s}}+\mathbb{1}_{D_{a,s}})\right]\\
&\leq \sum_{t=1}^{T}\sum_{s=t^b}^{t}\mathbb{P}(p(y_{a,1}|\theta_a)\leq L(\theta_a))^s\\
&\qquad+\sum_{t=1}^{T}\sum_{s=t^b}^{t}\sum_{s'=1}^{s}\mathbb{P}\left(\left|\frac{1}{s-1}\sum_{k=1,k\neq s'}^{s}(T(y_{a,k})-F'(\theta_a))\right|\geq\delta_a\right)\\
&\leq \sum_{t=1}^{T}t\left(\frac{1}{2}\right)^{t^b}+\sum_{t=1}^{T}t^2\exp(-t^b\tilde{K}(\theta_a,\delta)).
\end{aligned}$$

B.7 Extracting the KL-divergence: Proof of Lemma B.7

We assume that the event $\tilde{E}_u$ holds, $s'\leq u$. So, on this event we have

$$\begin{aligned}
\mathbb{P}\left(\mu(\theta_u)\geq\mu+\Delta|Y^u\right) &= \frac{\int_{\theta'\in\Theta_{\theta,\Delta}}\prod_{s=1,s\neq s'}^{u}p(y_s\mid\theta')p(y_{s'}|\theta')\pi(\theta')d\theta'}{\int_{\theta'\in\Theta}\prod_{s=1,s\neq s'}^{u}p(y_s\mid\theta')p(y_{s'}|\theta')\pi(\theta')d\theta'}\\
&= \frac{\int_{\theta'\in\Theta_{\theta,\Delta}}\prod_{s=1,s\neq s'}^{u}\frac{p(y_s|\theta')}{p(y_s|\theta)}p(y_{s'}|\theta')\pi(\theta')d\theta'}{\int_{\theta'\in\Theta}\prod_{s=1,s\neq s'}^{u}\frac{p(y_s|\theta')}{p(y_s|\theta)}p(y_{s'}|\theta')\pi(\theta')d\theta'}\\
&= \frac{\int_{\theta'\in\Theta_{\theta,\Delta}}e^{-(u-1)K[Y'^u,\theta,\theta']}\pi(\theta'|y_{s'})d\theta'}{\int_{\theta'\in\Theta}e^{-(u-1)K[Y'^u,\theta,\theta']}\pi(\theta'|y_{s'})d\theta'}
\end{aligned}$$

where $\pi(\theta|y_{s'})$ denotes the posterior distribution on θ after observation $y_{s'}$ and

$$K[Y^u_{s'},\theta,\theta']:=\frac{1}{u-1}\sum_{s=1,s\neq s'}^{u}\ln\frac{p(y_s\mid\theta)}{p(y_s\mid\theta')}$$

denotes the empirical KL-divergence obtained from the observations $Y^u_{s'} = Y^u \setminus \{y_{s'}\}$. Introducing

$$r(Y^u_{s'}, \theta') = K[Y^u_{s'}, \theta, \theta'] - \mathbb{E}_{X|\theta}\left(\ln \frac{p(X \mid \theta)}{p(X \mid \theta')}\right),$$

we can rewrite

$$\mathbb{P}\left(\mu(\theta_u) \geq \mu + \Delta | Y^u\right) = \frac{\int_{\theta' \in \Theta_{\theta,\Delta}} e^{-(u-1)(K[\theta,\theta'] + r(Y'^u, \theta'))} \pi(\theta' | y_{s'}) d\theta'}{\int_{\theta' \in \Theta} e^{-(u-1)(K[\theta,\theta'] + r(Y'^u, \theta'))} \pi(\theta' | y_{s'}) d\theta'}.$$

Now, a direct computation show that

$$\left| r(Y'^u, \theta') \right| \leq |\theta - \theta'| \left| \frac{1}{u-1} \sum_{s=1, s \neq s'}^{u} \left[T(y_s) - F'(\theta) \right] \right|. \tag{B.13}$$

Indeed, for any $\theta, \theta' \in \Theta$

$$\ln \frac{p(y \mid \theta)}{p(y \mid \theta')} = T(y)(\theta - \theta') - [F(\theta) - F(\theta')],$$

and one also recalls that

$$K(\theta, \theta') = F'(\theta)(\theta - \theta') - [F(\theta) - F(\theta')]. \tag{B.14}$$

Hence

$$\begin{aligned} |r(Y^u_{s'}, \theta, \theta')| &= \left| \frac{1}{u-1} \sum_{s=1, s \neq s'}^{u} \left[\ln \frac{p(y_s \mid \theta)}{p(y_s \mid \theta')} - K(\theta, \theta') \right] \right| \\ &= \left| \frac{1}{u-1} \sum_{s=1, s \neq s'}^{u} \left[(T(x) - F'(\theta))(\theta - \theta') \right] \right| \leq \left| \frac{1}{u-1} \sum_{s=1, s \neq s'}^{u} \left[T(y_s) - \nabla F(\theta) \right] \right| |\theta' - \theta|. \end{aligned}$$

The inequality (B.13) leads to the result, using that on $\tilde{E}_u$,

$$\left| \frac{1}{u-1} \sum_{s=1, s \neq s'}^{u} \left[T(y_s) - F'(\theta) \right] \right| \leq \delta$$

B.8 Proof of Lemma B.6

From Theorem B.4 we know that, for $N_{a,t} \geq N(\theta_a, F)$,

$$\begin{aligned} \mathbb{1}_{\tilde{E}_{a,t}} \mathbb{P}((E^\theta_{a,t})^c \mid \mathcal{F}_t) &= \mathbb{1}_{\tilde{E}_{a,t}} \mathbb{P}((E^\theta_{a,t})^c \mid Y_{a,t}) \\ &\leq C_{1,a} e^{-(N_{a,t}-1)(1-\delta_a C_{2,a}) \mathrm{KL}(\theta_a, \mu^{-1}(\mu_a + \Delta_a)) + \ln N_{a,t}} \\ &\leq e^{-(N_{a,t}-1)\left((1-\delta_a C_{2,a}) \mathrm{KL}(\theta_a, \mu^{-1}(\mu_a + \Delta_a)) - \ln(C_{1,a} N_{a,t})/(N_{a,t}-1)\right)} \end{aligned}$$

Let $N_\epsilon = N_\epsilon(\delta_a, \Delta_a, \theta_a)$ be the smallest integer such that for all $n \geq N_\epsilon$

$$\frac{\ln(C_{1,a} n)}{n-1} < \epsilon (1 - \delta_a C_{2,a}) \mathrm{KL}(\theta_a, \mu^{-1}(\mu_a + \Delta_a)).$$

Defining

$$L_T := \frac{\ln T}{(1-\epsilon)(1-\delta_a C_{2,a})\mathrm{KL}(\theta_a, \mu^{-1}(\mu_a + \Delta_a))}$$

we have that for all t and T such that $N_{a,t} - 1 \geq \max(L_T, N_\epsilon, N(\theta_a, F))$,

$$\mathbb{1}_{\tilde{E}_{a,t}} \mathbb{P}(\mu(\theta_a(t) > \mu(\theta_a) + \Delta_a \mid \mathcal{F}_t) \leq \frac{1}{T}.$$

Let $\tau = \inf\{t \in \mathbb{N} \mid N_{a,t} \geq \max(L_T, N_\epsilon, N(\theta_a, F)) + 1\}$. τ is a stopping time with respect to $\mathcal{F}_t$. Then,

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{P}\left(a_t = a, (E_{a,t}^{\theta})^c, \tilde{E}_{a,t}\right) &\leq \mathbb{E}\left[\sum_{t=1}^{\tau} \mathbb{1}_{(a_t=a)}\right] + \mathbb{E}\left[\sum_{t=\tau+1}^{T} \mathbb{1}_{(a_t=a)} \mathbb{1}_{\tilde{E}_{a,t}} \mathbb{1}_{(E_{a,t}^{\theta})^c}\right] \\
&= \mathbb{E}[N_{a,\tau}] + \mathbb{E}\left[\sum_{t=\tau+1}^{T} \mathbb{1}_{(a_t=a)} \mathbb{1}_{\tilde{E}_{a,t}} \mathbb{P}\left((E_{a,t}^{\theta})^c \mid \mathcal{F}_t\right)\right] \\
&= \mathbb{E}[N_{a,\tau}] + \mathbb{E}\left[\sum_{t=\tau+1}^{T} \mathbb{1}_{(a_t=a)} \mathbb{1}_{\tilde{E}_{a,t}} \mathbb{P}\left(\mu(\theta_a(t) > \mu(\theta_a) + \Delta_a \mid Y_{a,t}\right)\right] \\
&\leq L_T + 1 + \max(N_\epsilon, N(\theta_a, F)) + \mathbb{E}\left[\sum_{t=\tau+1}^{T} \frac{1}{T}\right] \\
&\leq L_T + \max(N_\epsilon, N(\theta_a, F)) + 2.
\end{aligned}$$

B.9 Controling the Number of Optimal Plays: Outline Proof of Proposition B.5

The proof of this proposition is quite detailed, and essentially the same as the proof given for Proposition 1 in [Kaufmann et al., 2012b], which we will sometimes refer to. However, in generalising to the case of exponential family bandits we show how to avoid the need to explicity calculate posterior probabilities that lead to Lemma 4 in [Kaufmann et al., 2012b]. While simplifying the proof we loose the ability to specify the constants explicitly, and so the analysis becomes asymptotic, but holds for every $b \in]0, 1[$.

Sketch of the proof and key results Let τ_j be the occurrence of the j^{th} play of the optimal arm (with $\tau_0 := 0$). Let $\xi_j := (\tau_{j+1} - 1) - \tau_j$: this random variable measures the number of time steps between the j^{th} and the $(j+1)^{th}$ play of the optimal arm, and so $\sum_{a=2}^{K} N_{a,t} = \sum_{j=0}^{N_{1,t}} \xi_j$. We then upper bound $\mathbb{P}(N_{1,t} \leq t^b)$ as in [Kaufmann et al., 2012b]:

$$\mathbb{P}(N_{1,t} \leq t^b) \leq \mathbb{P}\left(\exists j \in \left\{0, .., \lfloor t^b \rfloor\right\} : \xi_j \geq t^{1-b} - 1\right) \leq \sum_{j=0}^{\lfloor t^b \rfloor} \mathbb{P}(\underbrace{\xi_j \geq t^{1-b} - 1}_{:=\mathcal{E}_j}) \tag{B.15}$$

We introduce the interval $\mathcal{I}_j = \{\tau_j, \tau_j + \lceil t^{1-b} - 1 \rceil\}$: on the event $\mathcal{E}_j$, $\mathcal{I}_j$ is included in $\{\tau_j, \tau_{j+1}\}$ and no draw of arm 1 occurs on $\mathcal{I}$. We also introduce for each arm $a \neq 1$ $d_a := \frac{\mu_1 - \mu_a}{2}$.

The idea of the rest of the analysis is based on the following remark. If on a subinterval $\mathcal{I} \subseteq [\tau_j, \tau_{j+1}[$ of size $f(t)$ arm 1 is not drawn and all the samples of the suboptimal arms fall below $\mu_2 + d_2 < \mu_1$, then for all $s \in \mathcal{I}$, $\mu(\theta_{1,s}) \leq \mu_2 + d_2$. On $\mathcal{I}$, the sequence $(\theta_{1,s})$ is i.i.d. with distribution π_{1,τ_j}, and hence,

$$\mathbb{P}(\forall s \in \mathcal{I},\ \mu(\theta_{1,s}) \leq \mu_2 + \delta) \leq \left(\mathbb{P}\left(\mu(\theta_{1,\tau_j}) \leq \mu_2 + \delta_2\right)\right)^{f(t)}$$

At this point, an asymptotic result, telling that the posterior on θ_1 concentrates to a Dirac in θ_1 (the Bernstein-Von-Mises theorem, see [Van der Vaart, 1998]) , leads to

$$\mathbb{P}(\mu(\theta_{1,\tau_j}) \leq \mu_2 + \delta_2) \underset{j\to\infty}{\to} 0.$$

Assuming that $\forall j$, $\mathbb{P}(\mu(\theta_{1,\tau_j}) \leq \mu_2 + \delta_2) \neq 1$, we have shown the following Lemma, which plays the role of an asymptotic counterpart for Lemma 3 in [Kaufmann et al., 2012b].

Lemma B.9. *There exists a constant $C = C(\pi_0) < 1$, such that for every (random) interval $\mathcal{I}$ included in $\mathcal{I}_j$ and for every positive function f, one has*

$$\mathbb{P}\left(\forall s \in \mathcal{I},\ \mu(\theta_{1,s}) \leq \mu_2 + \delta_2, \quad |\mathcal{I}| \geq f(t)\right) \leq C^{f(t)}.$$

Another key lemma is the following which generalizes Lemma 4 in [Kaufmann et al., 2012b]. The proof of this lemma is standard: it proceeds by conditioning on the event $\tilde{E}_{a,t}$ [1] and applying Theorem B.4, and Lemma B.3.

Lemma B.10. *For every $a \in A$, $\delta > 0$, there exist constants $C_a = C_a(\mu_a, \delta, F)$ and N such that for $t \geq N$,*

$$\mathbb{P}\left(\exists s \leq t, \exists a \neq 1 : \mu(\theta_{a,s}) > \mu_a + d_a, N_{a,s} > C_a \ln(t)\right) \leq \frac{2(K-1)}{t^2}.$$

The rest of the proof proceeds by finding a subinterval of $\mathcal{I}_j$ on which all the samples of all the suboptimal arms indeed fall below the corresponding thresholds $\mu_a + d_a$. This is done exactly as in [Kaufmann et al., 2012b] and we recall the main steps of the proof below. Before that, we need to introduce the notion of *saturated*, suboptimal action.

Definition B.11. *Let t be fixed. For any $a \neq 1$, an action a is said to be* saturated *at time s if it has been chosen at least $C_a \ln(t)$ times, i.e. $N_{a,t} \geq C_a \ln(t)$. We shall say that it is* unsaturated *otherwise. Furthermore at any time we call a choice of an unsaturated, suboptimal action an* interruption.

Step 1: Decomposition of $\mathcal{I}_j$ We want to study the process of saturation on the event $\mathcal{E}_j = \{\xi_j \geq t^{1-b} - 1\}$. We start by decomposing the interval $\mathcal{I}_j = \{\tau_j, \tau_j + \lceil t^{1-b} - 1 \rceil\}$ into K subintervals:

$$\mathcal{I}_{j,l} := \left\{\tau_j + \left\lceil \frac{(l-1)(t^{1-b}-1)}{K} \right\rceil, \tau_j + \left\lceil \frac{l(t^{1-b}-1)}{K} \right\rceil\right\},\ l = 1, \ldots, K.$$

Now for each interval $\mathcal{I}_{j,l}$, we introduce:

- $\mathcal{F}_{j,l}$: the event that by the end of the interval $\mathcal{I}_{j,l}$ at least l suboptimal actions are saturated;
- $n_{j,l}$: the number of interruptions during this interval.

We use the following decomposition to bound the probability of the event $\mathcal{E}_j$:

$$\mathbb{P}(\mathcal{E}_j) = \mathbb{P}(\mathcal{E}_j \cap \mathcal{F}_{j,K-1}) + \mathbb{P}(\mathcal{E}_j \cap \mathcal{F}^c_{j,K-1}) \tag{B.16}$$

Note that the quantities $\mathcal{E}_j$, $\mathcal{I}_{j,l}$, $\mathcal{F}_{j,l}$ and $n_{j,l}$ all depend on t, however we suppress this dependency for notational convenience. However, we keep in mind that we bound the different probabilities for $t \geq N$, so that Lemma B.10 applies.

1. Using $\tilde{E}_{a,t}$ in place of $E_{a,t}$ from [Kaufmann et al., 2012b] only changes slightly the constant C_a.

Step 2: Bounding $\mathbb{P}(\mathcal{E}_j \cap \mathcal{F}_{j,K-1})$ On the event $\mathcal{E}_j \cap \mathcal{F}_{j,K-1}$, only saturated suboptimal arms are drawn on the interval $\mathcal{I}_{j,K}$. Using Lemma B.10, we get

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_j \cap \mathcal{F}_{j,K-1}) \leq & \mathbb{P}(\{\exists s \in \mathcal{I}_{j,K}, a \neq 1 : \mu(\theta_{a,s}) > \mu_a + d_a\} \cap \mathcal{E}_j \cap \mathcal{F}_{j,K-1}) \\
& + \mathbb{P}(\{\forall s \in \mathcal{I}_{j,K}, a \neq 1 : \mu(\theta_{a,s}) \leq \mu_a + d_a\} \cap \mathcal{E}_j \cap \mathcal{F}_{j,K-1}) \\
\leq & \mathbb{P}(\exists s \leq t, a \neq 1 : \mu(\theta_{a,s}) > \mu_a + d_a, N_{a,t} > C_a \ln(t)) \\
& + \mathbb{P}(\{\forall s \in \mathcal{I}_{j,K}, a \neq 1 : \mu(\theta_{a,s}) > \mu_a + d_a\} \cap \mathcal{E}_j \cap \mathcal{F}_{j,K-1}) \\
\leq & \frac{2(K-1)}{t^2} + \mathbb{P}(\{\forall s \in \mathcal{I}_{j,K} : \mu(\theta_{1,s}) \leq \mu_2 + d_2\} \cap \mathcal{E}_j) \\
\leq & \frac{2(K-1)}{t^2} + C^{\frac{t^{1-b}-1}{K}}.
\end{aligned}$$

for $0 < C < 1$ as in Lemma B.9. The second last inequality comes from the fact that if arm 1 is not drawn, the sample $\theta_{1,s}$ must be smaller than some sample $\theta_{a,s}$ and therefore smaller than $\mu_2 + d_2$.

Step 3: Bounding $\mathbb{P}(\mathcal{E}_j \cap \mathcal{F}^c_{j,K-1})$ A similar argument to that employed in Step 2 can be used in an induction to show that for all $2 \leq l \leq K$, if t is larger than some deterministic constant $N_{\mu_1,\mu_2,b}$ specified in the base case,

$$\mathbb{P}(\mathcal{E}_j \cap \mathcal{F}^c_{j,l-1}) \leq (l-2)\left(\frac{2(K-1)}{t^2} + C^{\frac{t^{1-b}-1}{CK^2\ln(t)}}\right)$$

We refer the reader to [Kaufmann et al., 2012b] for a precise description of the induction. For $l = K$ we then get

$$\mathbb{P}(\mathcal{E}_j \cap \mathcal{F}^c_{j,K-1}) \leq (K-2)\left(\frac{2(K-1)}{t^2} + C^{\frac{t^{1-b}-1}{CK^2\ln(t)}}\right). \tag{B.17}$$

Step 4: Conclusion Putting Steps 2 and 3 together we obtain that for $t \geq N_0 := \max(N, N_{\mu_1,\mu_2,b})$,

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_j(t)) &\leq \frac{2(K-1)^2}{t^2} + C^{\frac{t^{1-b}-1}{K}} + (K-2)KC\ln(t)C^{\frac{t^{1-b}-1}{CK^2\ln(t)}}, \\
\mathbb{P}(N_{1,t} \leq t^b) &\leq \frac{2(K-1)^2}{t^{2-b}} + t^b C^{\frac{t^{1-b}-1}{K}} + (K-2)KCt^b\ln(t)C^{\frac{t^{1-b}-1}{CK^2\ln(t)}},
\end{aligned}$$

where we use B.15. It then follows that

$$\sum_{t=1}^{\infty} \mathbb{P}(N_{1,t} \leq t^b) \leq N_0 + \sum_{t=N_0+1}^{\infty} \mathbb{P}(\mathcal{E}_j) = C_b = C_b(\pi_0, \mu_1, \mu_2, K) < \infty.$$

Bibliography

[Abbasi-Yadkori et al., 2011] Abbasi-Yadkori, Y., D.Pál, and C.Szepesvári (2011). Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*.

[Agrawal, 1995] Agrawal, R. (1995). Sample mean based index policies with O(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078.

[Agrawal et al., 1989] Agrawal, R., Teneketzis, D., and Anantharam, V. (1989). Asymptotically Efficient Adaptive Allocation Schemes for Controlled i.i.d. Processes: Finite Parameter Space. *IEEE Transactions on Automatic Control*, 34(3):258–267.

[Agrawal and Goyal, 2012] Agrawal, S. and Goyal, N. (2012). Analysis of Thompson Sampling for the multi-armed bandit problem. In *Proceedings of the 25th Conference On Learning Theory*.

[Agrawal and Goyal, 2013a] Agrawal, S. and Goyal, N. (2013a). Further Optimal Regret Bounds for Thompson Sampling. In *Proceedings of the 16th Conference on Artificial Intelligence and Statistics*.

[Agrawal and Goyal, 2013b] Agrawal, S. and Goyal, N. (2013b). Thompson Sampling for Contextual Bandits with Linear Payoffs. In *International Conference on Machine Learning (ICML)*.

[Asmuth et al., 2009] Asmuth, J., Li, L., Littman, M., Nouri, A., and Wingate, D. (2009). A Bayesian sampling approach to exploration in reinforcement learning. In *Uncertainty in Artificial Intelligence (UAI)*.

[Audibert and Bubeck, 2010] Audibert, J.-Y. and Bubeck, S. (2010). Regret Bounds and Minimax Policies under Partial Monitoring. *Journal of Machine Learning Research*.

[Audibert et al., 2010] Audibert, J.-Y., Bubeck, S., and Munos, R. (2010). Best Arm Identification in Multi-armed Bandits. In *Proceedings of the 23rd Conference on Learning Theory*.

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

[Auer, 2002] Auer (2002). Using Confidence bounds for Exploration Exploitation trade-offs. *Journal of Machine Learning Research*, 3:397–422.

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

[Auer et al., 2002b] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. (2002b). The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 32(1):48–77.

[Bechhofer et al., 1968] Bechhofer, R., Kiefer, J., and Sobel, M. (1968). *Sequential identification and ranking procedures*. The University of Chicago Press.

[Bellman, 1954] Bellman, R. (1954). The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515.

[Bellman, 1956] Bellman, R. (1956). A problem in the sequential design of experiments. *The indian journal of statistics*, 16(3/4):221–229.

[Berry and Fristedt, 1985] Berry, D. and Fristedt, B. (1985). *Bandit Problems. Sequential allocation of experiments*. Chapman and Hall.

[Bickel and Doksum, 2001] Bickel, P. and Doksum, K. (2001). *Mathematical Statistics, Basic Ideas and Selected Topics*. Prentice Hall.

[Bishop, 2006] Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag New York.

[Boucheron et al., 2013] Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities. A non asymptotic theory of independence.* Oxford University Press.

[Bradt et al., 1956] Bradt, R., Johnson, S., and Karlin, S. (1956). On sequential designs for maximizing the sum of n observations. *Annals of Mathematical Statistics*, 27(4):1060–1074.

[Bubeck, 2010] Bubeck, S. (2010). *Jeux de bandits et fondation du clustering*. PhD thesis, Université de Lille 1.

[Bubeck and Cesa-Bianchi, 2012] Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Fondations and Trends in Machine Learning*, 5(1):1–122.

[Bubeck et al., 2012] Bubeck, S., Cesa-Bianchi, N., and Kakade, S. (2012). Towards Minimax Policies for Online Linear Opimization with Bandit Feedback. In *Proceedings of the 25th Conference On Learning Theory*.

[Bubeck and Liu, 2013] Bubeck, S. and Liu, C.-Y. (2013). Prior-free and prior-dependent regret bounds for Thompson Sampling. In *Advances in Neural Information Processing Systems*.

[Bubeck et al., 2011] Bubeck, S., Munos, R., and Stoltz, G. (2011). Pure Exploration in Finitely Armed and Continuous Armed Bandits. *Theoretical Computer Science 412, 1832-1852*, 412:1832–1852.

[Bubeck et al., 2013a] Bubeck, S., Perchet, V., and Rigollet, P. (2013a). Bounded regret in stochastic multi-armed bandits. In *Proceedings of the 26th Conference On Leaning Theory*.

[Bubeck et al., 2013b] Bubeck, S., Wang, T., and Viswanathan, N. (2013b). Multiple Identifications in multi-armed bandits. In *International Conference on Machine Learning (ICML)*.

[Burnetas and Katehakis, 1996] Burnetas, A. and Katehakis, M. (1996). Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142.

[Burnetas and Katehakis, 2003] Burnetas, A. and Katehakis, M. (2003). Asymptotic Bayes Analysis for the finite horizon one armed bandit problem. *Probability in the Engineering and Informational Sciences*, 17:53–82.

[Cappé et al., 2013] Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., and Stoltz, G. (2013). Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541.

[Cesa-Bianchi and Lugosi, 2006] Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning and Games*. Cambridge University Press.

[Cesa-Bianchi and Lugosi, 2012] Cesa-Bianchi, N. and Lugosi, G. (2012). Combinatorial Bandits. *Journal of Computer and System Sciences*, 78:1404–1422.

[Chandrasekaran and Karp, 2014] Chandrasekaran, K. and Karp, R. (2014). Finding a most biaised coin with fewest flips. In *Proceeding of the 27th Conference on Learning Theory*.

[Chang and Lai, 1987] Chang, F. and Lai, T. (1987). Optimal stopping and dynamic allocation. *Advances in Applied Probability*, 19:829–853.

[Chapelle and Li, 2011] Chapelle, O. and Li, L. (2011). An empirical evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems*.

[Chapelle et al., 2014] Chapelle, O., Manavoglu, E., and Rosales, R. (2014). Simple and scalable response prediction for display advertising. *Transactions on Intelligent Systems and Technology*.

[Chu et al., 2011] Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual Bandits with Linear Payoff Functions. In *Proceedings of the 14th Conference on Artificial Intelligence and Statistics*.

[Contal et al., 2013] Contal, E., Buffoni, D., and Vayatis, N. (2013). Parallel Gaussian Process Optimization with Upper Confidence Bounds and Pure Exploration. In *Proceedings of the European Conference on Machine Learning*.

[Cover and Thomas, 2006] Cover, T. and Thomas, J. (2006). *Elements of Information Theory (2nd Edition)*. Wiley.

[Dani et al., 2007] Dani, V., Hayes, T., and Kakade, S. (2007). The Price of Bandit Information in Online Optimization. In *Advances in Neural Information and Signal Processing*.

[Dani et al., 2008] Dani, V., Hayes, T., and Kakade, S. (2008). Stochastic Linear Optimization under Bandit Feedback. In *Advances in Neural Information and Signal Processing*, pages 355–366.

[De La Pena et al., 2004] De La Pena, V., Klass, M., and Lai, T. (2004). Self-Normalized Processes: Exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 32(3A):1902–1933.

[De La Pena et al., 2009] De La Pena, V., Lai, T., and Q., S. (2009). *Self-normalized processes. Limit Theory and Statistical applications*. Springer.

[Dembo and Zeitouni, 2010] Dembo, A. and Zeitouni, O. (2010). *Large Deviations Techniques and Applications, 2nd Edition*. Springer.

[Durrett, 2010] Durrett, R. (2010). *Probability: Theory and Examples*. Cambridge University Press.

[Even-Dar et al., 2006] Even-Dar, E., Mannor, S., and Mansour, Y. (2006). Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning Problems. *Journal of Machine Learning Research*, 7:1079–1105.

[Feldman, 1962] Feldman, D. (1962). Contributions to the "two-armed bandit". *The Annals of Mathematical Statistics*, 33(3):947–956.

[Filippi et al., 2010a] Filippi, S., Cappé, O., and Garivier, A. (2010a). Optimism in Reinforcement Learning and Kullback-Leibler Divergence. In *Allerton Conference on Communication, Control, and Computing*, Monticello, US.

[Filippi et al., 2010b] Filippi, S., Cappé, O., Garivier, A., and Szepesvári, C. (2010b). Parametric Bandits : The Generalized Linear case. In *Advances in Neural Information Processing Systems*.

[Fonteneau et al., 2013] Fonteneau, R., Korda, N., and Munos, R. (2013). An optimistic posterior sampling strategy for Bayesian reinforcement learning. In *Workshop on Bayesian Optimization, NIPS*.

[Frostig and Weiss, 1999] Frostig, E. and Weiss, G. (1999). Four proofs of Gittins' multiarmed bandit theorem. Technical report.

[Gabillon et al., 2012] Gabillon, V., Ghavamzadeh, M., and Lazaric, A. (2012). Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence. In *Advances in Neural Information Processing Systems*.

[Garivier, 2013] Garivier, A. (2013). Informational Confidence Bounds for Self-Normalized Averages and Applications. In *IEEE Information Theory Workshop*.

[Garivier and Cappé, 2011] Garivier, A. and Cappé, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Conference on Learning Theory*.

[Garivier and Moulines, 2011] Garivier, A. and Moulines, E. (2011). On Upper-Confidence Bound Policies for Switching Bandit Problems. In *Proceedings of the 22nd conference on Algorithmic Learning Theory*.

[Ginebra and Clayton, 1994] Ginebra, J. and Clayton, M. (1994). Small-sample frequentist properties of Bernoulli two-armed bandit Bayesian strategies. Technical report, University of Wisconsin.

[Ginebra and Clayton, 1999] Ginebra, J. and Clayton, M. (1999). Small-sample performance of Bernoulli two-armed bandit Bayesian strategies. *Journal of Statistical Planning and Inference*, 79(1):107–122.

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

[Gittins et al., 2011] Gittins, J., Glazebrook, K., and Weber, R. (2011). *Multi-armed bandit allocation indices (2nd Edition)*. Wiley.

[Gittins and Jones, 1974] Gittins, J. and Jones, D. (1974). A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics (proceedings of the 1972 European Meeting of Statisticians)*.

[Gopalan et al., 2014] Gopalan, A., Mannor, S., and Mansour, Y. (2014). Thompson Sampling for Complex Online Problems. In *International Conference on Machine Learning (ICML)*.

[Granmo, 2010] Granmo, O. (2010). Solving two-armed Bernoulli Bandit Problems using a Bayesian Learning Automaton. *International Journal of Intelligent Computing and Cybernetics*, 3(2):207–234.

[Graves and Lai, 1997] Graves, T. and Lai, T. (1997). Asymptotically Efficient adaptive choice of control laws in controlled markov chains. *SIAM Journal on Control and Optimization*, 35(3):715–743.

[Guha and Munagala, 2014] Guha, S. and Munagala, K. (2014). Stochastic Regret Minimization via Thompson Sampling. In *Proceedings of the 27th Conference On Learning Theory*.

[Heidrich-Meisner and Igel, 2009] Heidrich-Meisner, V. and Igel, C. (2009). Hoeffding and Bernstein Races for Selecting Policies in Evolutionary Direct Policy Search. In *International Conference on Machine Learning (ICML)*.

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

[Hoffman et al., 2014] Hoffman, M., Shahriari, B., and de Freitas, N. (2014). On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*.

[Honda and Takemura, 2010] Honda, J. and Takemura, A. (2010). An Asymptotically Optimal Bandit Algorithm for Bounded Support Models. In Kalai, T. and Mohri, M., editors, *Proceedings of the 23rd Conference on Learning Theory*.

[Honda and Takemura, 2014] Honda, J. and Takemura, A. (2014). Optimality of Thompson Sampling for Gaussian Bandits depends on priors. In *Proceedings of the 17th conference on Artificial Intelligence and Statistics*.

[Jaksch et al., 2010] Jaksch, T., Ortner, R., and Auer, P. (2010). Near-Optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11:1563–1600.

[Jamieson et al., 2014] Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014). lil'UCB: an Optimal Exploration Algorithm for Multi-Armed Bandits. In *Proceedings of the 27th Conference on Learning Theory*.

[Jeffreys, 1946] Jeffreys, H. (1946). An invariant form for prior probability in estimation problems. *Proceedings of the Royal Society of London, Serie A.*, 286:453–461.

[Jennison et al., 1982] Jennison, C., Johnstone, I. M., and Turnbull, B. W. (1982). Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. *Statistical Decision Theory and Related Topics III*, 2:55–86.

[Kalyanakrishnan and Stone, 2010] Kalyanakrishnan, S. and Stone, P. (2010). Efficient Selection in Multiple Bandit Arms: Theory and Practice. In *International Conference on Machine Learning (ICML)*.

[Kalyanakrishnan et al., 2012] Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. (2012). PAC subset selection in stochastic multi-armed bandits. In *International Conference on Machine Learning (ICML)*.

[Karnin et al., 2013] Karnin, Z., Koren, T., and Somekh, O. (2013). Almost optimal Exploration in multi-armed bandits. In *International Conference on Machine Learning (ICML)*.

[Katehakis and Robbins, 1995] Katehakis, M. and Robbins, H. (1995). Sequential choice from several populations. *Proceedings of the National Academy of Science*, 92:8584–8585.

[Kaufmann et al., 2012a] Kaufmann, E., Cappé, O., and Garivier, A. (2012a). On Bayesian Upper-Confidence Bounds for Bandit Problems. In *Proceedings of the 15th conference on Artificial Intelligence and Statistics*.

[Kaufmann et al., 2014a] Kaufmann, E., Cappé, O., and Garivier, A. (2014a). On the Complexity of A/B Testing. In *Proceedings of the 27th Conference On Learning Theory*.

[Kaufmann et al., 2014b] Kaufmann, E., Cappé, O., and Garivier, A. (2014b). On the Complexity of Best Arm Identification in Multi-Armed Bandit Models. *Submitted*, arXiv:1407.4443.

[Kaufmann and Kalyanakrishnan, 2013] Kaufmann, E. and Kalyanakrishnan, S. (2013). Information complexity in bandit subset selection. In *Proceeding of the 26th Conference On Learning Theory.*

[Kaufmann et al., 2012b] Kaufmann, E., Korda, N., and Munos, R. (2012b). Thompson Sampling : an Asymptotically Optimal Finite-Time Analysis. In *Proceedings of the 23rd conference on Algorithmic Learning Theory*.

[Kaufmann, 2014] Kaufmann, S. (2014). *Requiem pour Stanley*. Editions Lulu.

[Korda et al., 2013] Korda, N., Kaufmann, E., and Munos, R. (2013). Thompson Sampling for 1-dimensional Exponential family bandits. In *Advances in Neural Information Processing Systems*.

[Krause and Ong, 2011] Krause, A. and Ong, C. (2011). Contextual Gaussian Process Bandit Optimization. In *Advances in Neural Information Processing Systems*.

[Lai, 1987] Lai, T. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *Annals of Statistics*, 15(3):1091–1114.

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

[Laurent and Massart, 2000] Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338.

[Lelarge et al., 2013] Lelarge, M., Proutière, A., and Talebi, S. (2013). Spectrum Bandit Optimization. In *ITW*.

[Levin and Leu, 2008] Levin, B. and Leu, C. (2008). On a Conjecture of Bechhofer, Kiefer, and Sobel for the Levin-Robbins-Leu Binomial Subset Selection Procedures. *Sequential Analysis*, 27:106–125.

[Maillard et al., 2011] Maillard, O.-A., Munos, R., and Stoltz, G. (2011). A Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences. In *Proceedings of the 24th Conference On Learning Theory*.

[Mannor and Tsitsiklis, 2004] Mannor, S. and Tsitsiklis, J. (2004). The Sample Complexity of Exploration in the Multi-Armed Bandit Problem. *Journal of Machine Learning Research*, pages 623–648.

[Maron and Moore, 1997] Maron, O. and Moore, A. (1997). The Racing algorithm: Model selection for Lazy learners. *Artificial Intelligence Review*, 11(1-5):113–131.

[May et al., 2012] May, B., Korda, N., A., L., and D., L. (2012). Optimistic Bayesian sampling in contextual bandit problems. *Journal of Machine Learning Research*, 13:2069–2106.

[Mellor and Shapiro, 2013] Mellor, J. and Shapiro, J. (2013). Thompson Sampling in Switching Environments with Bayesian Online Change Point Detection. In *Proceeding of the 16th Conference on Artificial Intelligence and Statistics*.

[Mnih et al., 2008] Mnih, V., Szepesvári, C., and Audibert, J.-Y. (2008). Empirical Bernstein stopping. In *International Conference on Machine Learning (ICML)*.

[Naghshvar and Javidi, 2013] Naghshvar, M. and Javidi, T. (2013). Active sequential hypothesis testing. *Annals of Statistics*, 41(6):2703–2738.

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

[Nino-Mora, 2011] Nino-Mora, J. (2011). Computing a Classic Index for Finite-Horizon Bandits. *INFORMS Journal of Computing*, 23(2):254–267.

[Osband et al., 2013] Osband, I., Van Roy, B., and Russo, D. (2013). (More) Efficient Reinforcement Learning Via Posterior Sampling. In *Advances in Neural Information Processing Systems*.

[Paulson, 1964] Paulson, E. (1964). A sequential procedure for selecting the population with the largest mean from k normal populations. *Annals of Mathematical Statistics*, 35:174–180.

[Pavlidis et al., 2008] Pavlidis, N., Tasoulis, D., and Hand, D. (2008). Simulation studies of multi-armed bandits with covariates. In *10th Proceedings of the International Conference on Computer Modeling*.

[Puterman, 1994] Puterman, M. (1994). *Markov Decision Processes. Discrete Stochastic. Dynamic Programming*. Wiley.

[Robbins, 1952] Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

[Robbins, 1970] Robbins, H. (1970). Statistical Methods Related to the law of the iterated logarithm. *Annals of Mathematical Statistics*, 41(5):1397–1409.

[Rusmevichientong and Tsitsiklis, 2010] Rusmevichientong, P. and Tsitsiklis, J. (2010). Linearly Parameterized Bandits. *Mathematics of Operations Research*, 35(2):395–411.

[Russo and Van Roy, 2014] Russo, D. and Van Roy, B. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research (to appear)*.

[Salomon and Audibert, 2011] Salomon, A. and Audibert, J.-Y. (2011). Deviations of stochastic bandit regret. In *Proceedings of the 22nd conference on Algorithmic Learning Theory*.

[Schreck et al., 2013] Schreck, A., Fort, G., Le Corff, S., and Moulines, E. (2013). A shrinkage-thresholding Metropolis adjusted Langevin algorithm for Bayesian variable selection. *arXiv:1312.5658*.

[Scott, 2010] Scott, S. (2010). A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26:639–658.

[Siegmund, 1985] Siegmund, D. (1985). *Sequential Analysis*. Springer-Verlag.

[Sigaud and Buffet, 2008] Sigaud, O. and Buffet, O. (2008). *Processus Decisionnels de Markov et Intelligence artificielle*. Hermès.

[Srinivas et al., 2010] Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). Gaussian Process Optimization in the Bandit Setting : No Regret and Experimental Design. In *Proceedings of the International Conference on Machine Learning*.

[Srinivas et al., 2012] Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2012). Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265.

[Strens, 2000] Strens, M. (2000). A Bayesian Framework for Reinforcement Learning. In *ICML*.

[Thompson, 1933] Thompson, W. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294.

[Thompson, 1935] Thompson, W. (1935). On the theory of apportionment. *American Journal of Mathematics*, 57:450–456.

[Valko et al., 2013] Valko, M., Korda, N., Munos, R., and Cristinini, N. (2013). Finite-time analysis of kernelized contextual bandits. In *29th Conference on Uncertainty in Artificial Intelligence (UAI)*.

[Van der Vaart, 1998] Van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

[Wald, 1945] Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *Annals of Mathematical Statistics*, 16(2):117–186.

[Wasserman, 2010] Wasserman, L. (2010). *All of Statistics: A concise course in statistical inference*. Springer.

[Weber, 1992] Weber, R. (1992). On the Gittins index for multiarmed bandits. *Annals of Applied Probabilities*, 2(4):1024–1033.

Analyse de stratégies bayésiennes et fréquentistes pour l'allocation séquentielle de ressources

Emilie KAUFMANN

RESUME : Dans cette thèse, nous étudions des stratégies d'allocation séquentielle de ressources. Le modèle statistique adopté dans ce cadre est celui du bandit stochastique à plusieurs bras. Dans ce modèle, lorsqu'un agent tire un bras du bandit, il reçoit pour récompense une réalisation d'une distribution de probabilité associée au bras. Nous nous intéressons à deux problèmes de bandit différents : la maximisation de la somme des récompenses et l'identification des meilleurs bras (où l'agent cherche à identifier le ou les bras conduisant à la meilleure récompense moyenne, sans subir de perte lorsqu'il tire un «mauvais» bras). Nous nous attachons à proposer pour ces deux objectifs des stratégies de tirage des bras, aussi appelées algorithmes de bandit, que l'on peut qualifier d'optimales.

La maximisation des récompenses est équivalente à la minimisation d'une quantité appelée *regret*. Grâce à une borne inférieure asymptotique sur le regret d'une stratégie uniformément efficace établie par Lai et Robbins, on peut définir la notion d'algorithme asymptotiquement optimal comme un algorithme dont le regret atteint cette borne inférieure. Dans cette thèse, nous proposons pour deux algorithmes d'inspiration bayésienne, Bayes-UCB et Thompson Sampling, une analyse à temps fini dans le cadre des modèles de bandit à récompenses binaires, c'est-à-dire une majoration non asymptotique de leur regret. Cette majoration permet d'établir l'optimalité asymptotique des deux algorithmes.

Dans le cadre de l'identification des meilleurs bras, on peut chercher à déterminer le nombre total d'échantillons des bras nécessaires pour identifier, avec forte probabilité, le ou les meilleurs bras, sans la contrainte de maximiser la somme des observations. Nous définissons deux termes de complexité pour l'identification des meilleurs bras dans deux cadres considérés dans la littérature, qui correspondent à un budget fixé ou à un niveau de confiance fixé. Nous proposons de nouvelles bornes inférieures sur ces complexités, et nous analysons de nouveaux algorithmes, dont certains atteignent les bornes inférieures dans des cas particuliers de modèles de bandit à deux bras, et peuvent donc être qualifiés d'optimaux.

MOTS-CLEFS: Modèles de bandit, minimisation du regret, identification des meilleurs bras.

ABSTRACT: In this thesis, we study strategies for sequential resource allocation, under the so-called stochastic multi-armed bandit model. In this model, when an agent draws an arm, he receives as a reward a realization from a probability distribution associated to the arm. In this document, we consider two different bandit problems. In the reward maximization objective, the agent aims at maximizing the sum of rewards obtained during his interaction with the bandit, whereas in the best arm identification objective, his goal is to find the set of m best arms (i.e. arms with highest mean reward), without suffering a loss when drawing 'bad' arms. For these two objectives, we propose strategies, also called bandit algorithms, that are optimal (or close to optimal), in a sense precised below.

Maximizing the sum of rewards is equivalent to minimizing a quantity called *regret*. Thanks to an asymptotic lower bound on the regret of any uniformly efficient algorithm given by Lai and Robbins, one can define asymptotically optimal algorithms as algorithms whose regret reaches this lower bound. In this thesis, we propose, for two Bayesian algorithms, Bayes-UCB and Thompson Sampling, a finite-time analysis, that is a non-asymptotic upper bound on their regret, in the particular case of bandits with binary rewards. This upper bound allows to establish the asymptotic optimality of both algorithms.

In the best arm identification framework, a possible goal is to determine the number of samples of the arms needed to identify, with high probability, the set of m best arms. We define a notion of complexity for best arm identification in two different settings considered in the literature: the fixed-budget and fixed-confidence settings. We provide new lower bounds on these complexity terms and we analyse new algorithms, some of which reach the lower bound in particular cases of two-armed bandit models and are therefore optimal.

KEY-WORDS: Bandit models, regret minimization, best arm identification