Modèles de bandit : une histoire bayésienne et fréquentiste

Emilie Kaufmann

Un modèle de bandit à K bras est un ensemble de K lois de probabilités $\nu_1, \dots, \nu_K$, appelées *bras*, avec lesquelles un agent, ignorant les caractéristiques de ces bras, interagit. A chaque instant t, ce dernier choisit un bras A_t et observe une récompense X_t tirée sous la loi du bras choisi : $X_t \sim \nu_{A_t}$. L'objectif de l'agent est d'adopter une stratégie de tirages des bras maximisant l'espérance des récompenses cumulées jusqu'à un horizon T, $\mathbb{E}\left[\sum_{t=1}^{T} X_t\right]$. Si on note μ_a l'espérance de ν_a, l'agent va donc chercher à jouer le plus souvent possible le meilleur bras, de moyenne $\mu^* = \text{argmax}_a\, \mu_a$, qu'il ne connaît pas a priori.

La stratégie de l'agent, appelée parfois *algorithme de bandit*, est séquentielle : le choix du bras A_{t+1} est basé sur les bras choisis et les récompenses observées précédemment, $A_1, \dots, A_t, X_1, \dots, X_t$. Une bonne stratégie exploite cette information de sorte à réaliser un compromis entre exploration (essayer les bras peu joués jusque là) et exploitation (favoriser les bras qui ont obtenu des bonnes performances jusque là). Dans cet article, nous évoquerons des stratégies réalisant ce compromis de manière *optimale*, dans un sens qui dépendra de la modélisation probabiliste choisie, fréquentiste ou bayésienne.

Si le nom des modèles de bandit fait référence à un casino où il s'agirait de découvrir la machine à sous, ou bandit manchot, qui a le bras le plus performant, ce cadre n'est en fait qu'un prête-nom et à l'origine, ces modèles ont été introduits dans le contexte des essais cliniques [23]. Pour un symptôme donné, un médecin a à sa disposition K traitements, de probabilité de succès $\mu_1, \dots, \mu_K$, inconnues au début de l'étude clinique. Il choisit d'allouer au t-ème patient de l'étude l'un des traitements, A_t, et observe ensuite $X_t = 1$ si le patient est guéri, $X_t = 0$ sinon, avec

$$\mathbb{P}(X_t = 1 | A_t = a) = \mu_a.$$

Dans cette situation, les bras modélisent l'efficacité des traitements et produisent des récompenses binaires ; ce sont donc les lois de Bernoulli. Maximiser l'espérance des récompenses revient à maximiser le nombre moyen de patients guéris lors de l'étude. D'autres types d'applications motivent aujourd'hui l'étude des modèles de bandit, comme les systèmes de recommandation, où il s'agit de choisir des contenus à présenter à des utilisateurs [19], l'allocation adaptative de spectre dans les systèmes de radio cognitives [12] ou encore l'exploration optimiste d'un arbre minimax pour la résolution d'un jeu [17].

Pour certaines de ces applications, des modèles plus complexes ont parfois été proposés, mais dans le reste de cet article, nous allons considérer le modèle de bandit le plus simple, où les bras sont des lois de Bernoulli, notées $\mathcal{B}(\mu_1), \dots, \mathcal{B}(\mu_K)$ et où de plus les récompenses observées sont supposées indépendantes. Depuis les années 1930, ce modèle a été successivement étudié par plusieurs communautés, adoptant parfois une approche statistique dite bayésienne ou une approche dite fréquentiste. L'objectif de cet article est de présenter ces deux points de vue différents en statistique, et de voir comment dans le cas particulier de l'étude des modèles de bandit ils peuvent se compléter et s'influencer mutuellement. Une première partie de cet article est dédiée à l'historique de ces deux approches. Nous insisterons ensuite sur l'impact récent des stratégies utilisant des outils bayésiens pour la minimisation du regret (une mesure de performance fréquentiste présentée en section 3), qui sont au cœur de la thèse [13].

1 Deux modèles probabilistes

Le modèle de bandit à récompenses binaires que nous considérons est un cas particulier de modèle *paramétrique*, où la loi de chaque bras dépend d'un paramètre, ici sa moyenne $\mu_a \in [0,1]$. Dès lors, pour faire de l'inférence dans ces modèles, c'est-à-dire apprendre de l'information sur les paramètres des bras à partir d'observations de ceux-ci, deux approches concurrentes en statistiques peuvent être utilisées.

L'approche fréquentiste suppose que le paramètre μ_a de chaque bras est un paramètre inconnu, qui peut être inféré à l'aide d'estimateurs ponctuels, ou d'intervalles de confiances. Dans l'approche bayésienne, pour modéliser l'incertitude sur les paramètres des bras, on fait l'hypothèse que ceux-ci ont eux-mêmes été tirés aléatoirement, et sont des réalisations indépendantes d'une *loi a priori* π_0. L'information disponible sur μ_a après des observations de ce bras est alors encodée par la loi de la variable aléatoire μ_a conditionnellement à ces observations, appelée *loi a posteriori*.

Plaçons nous dans le contexte des modèles de bandit, et notons $N_a(t)$ le nombre d'observations du bras a à l'instant t, et $S_a(t)$ leur somme. Dans le cadre fréquentiste, le paramètre μ_a est usuellement estimé par sa moyenne empirique $\hat{\mu}_a(t) := S_a(t)/N_a(t)$, qui constitue l'estimateur du maximum de vraisemblance. Dans le cadre bayésien, si la loi a priori π_0 est uniforme sur l'intervalle $[0,1]$, ce que nous supposerons tout au long de cet article, la formule de Bayes permet de montrer que la loi a posteriori de μ_a est une loi Beta de paramètres $S_a(t)+1$ et $N_a(t)-S_a(t)+1$. On notera $\pi_a(t) := \mathrm{Beta}(S_a(t)+1, N_a(t)-S_a(t)+1)$. La figure ci-dessous permet de visualiser la densité des loi a posteriori Beta (qui ont une allure gaussienne), ainsi que la mise à jour de celles-ci avec une nouvelle observation.

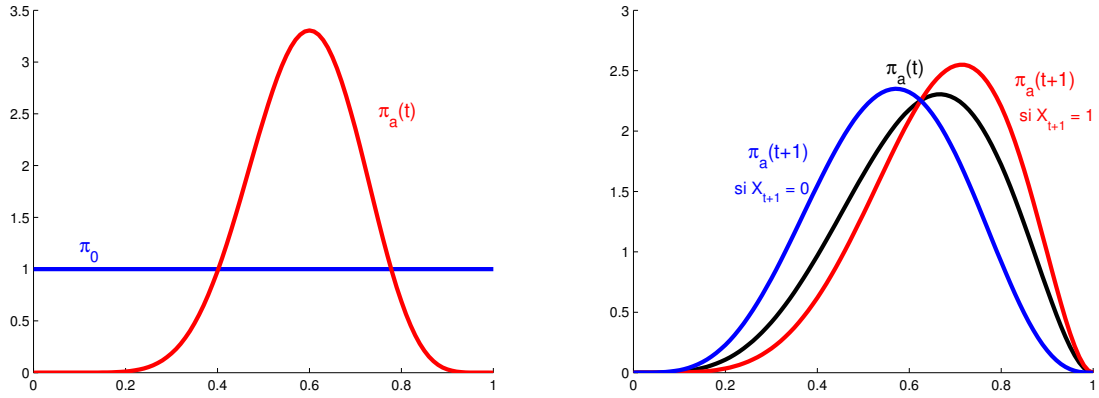

FIGURE 1 – La figure de gauche présente la loi a priori $\pi_0 = \mathcal{U}([0,1])$ et la loi a posteriori $\pi_a(t) = \mathrm{Beta}(10,7)$. Sur celle de droite, étant donné la loi a posteriori $\pi_a(t)$, on voit son évolution au temps $t+1$ si le bras a est choisi à l'instant $t+1$ selon que $X_{t+1} = 0$ ou $X_{t+1} = 1$.

Les approches fréquentiste et bayésienne correspondent à deux modèles probabilistes différents, le modèle fréquentiste dépendant du vecteur des paramètres des moyennes $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$, et le modèle bayésien dépendant de la loi a priori π_0. On peut donc chercher à maximiser l'espérance de la somme des récompenses sous chacun de ces deux modèles, notée respectivement

$$\mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} X_t\right] \quad \text{et} \quad \mathbb{E}^{\pi_0}\left[\sum_{t=1}^{T} X_t\right] = \mathbb{E}_{\mu_a \overset{i.i.d.}{\sim} \pi_0}\left[\mathbb{E}_{\boldsymbol{\mu}}\left[\sum_{t=1}^{T} X_t\right]\right].$$

Il est important de noter que l'on peut dissocier ces deux *mesures de performance* fréquentiste et bayésienne

des *outils* associés à ces deux modélisations permettant de construire des algorithmes de bandit : les intervalles de confiance d'une part, et les loi a posteriori d'autre part. Dans la suite de l'article nous présentons d'abord des stratégies bayésiennes et fréquentistes destinées à maximiser les récompenses sous le modèle associé. Mais nous verrons pour finir que l'utilisation d'algorithmes bayésiens peut également être très pertinente dans un cadre fréquentiste.

2 Stratégies bayésiennes

On peut considérer que le problème de bandit a d'abord été étudié par la communauté bayésienne, car le premier algorithme de bandit mentionné dans la littérature est un algorithme bayésien proposé par Thompson en 1933 [23]. Cet algorithme a ensuite été quelque peu oublié, mais le problème de bandit bayésien a été de nouveau étudié à partir des années 1950 [5, 8, 4]. Cette période correspond au développement de la programmation dynamique par Bellmann [3], et la maximisation des récompenses dans un modèle de bandit en est un exemple d'application.

Dans un modèle de bandit bayésien, à un instant donné t, l'historique du jeu est résumé par les loi a posteriori sur chacun des bras, formant l'état courant $\Pi(t) = (\pi_1(t), \dots, \pi_K(t))$. Celui-ci est représenté sur la figure 2, dans un modèle de bandit à 5 bras.

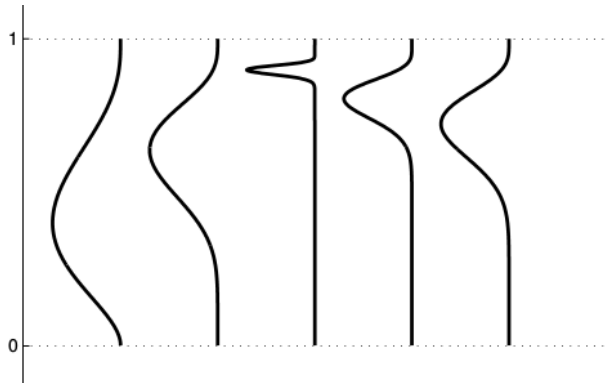

FIGURE 2 – Une version normalisée de la densité $\pi_a(t) = \mathrm{Beta}(S_a(t)+1, N_a(t)-S_a(t)+1)$ de chaque bras est représentée verticalement. Plus la loi a posteriori est concentrée, plus le bras a été joué.

Chaque loi Beta étant paramétrée par deux entiers indiquant le nombre de 1 et de 0 observés, l'état peut aussi être représenté par une matrice de taille $K \times 2$ dont la ligne a indique les paramètres de l'a posteriori courant sur μ_a. Dans un état Π, lorsque le bras a est tiré et la récompense $x \in \{0,1\}$ est observée, on note $\Pi^{a,x}$ le nouvel état obtenu, dans lequel seule la loi a posteriori du bras a est mise à jour en prenant en compte cette nouvelle récompense. Si $\Pi = (\mathrm{Beta}(1,2), \mathrm{Beta}(5,1), \mathrm{Beta}(0,2))$, on écrira

$$\Pi = \begin{pmatrix} 1 & 2 \\ 5 & 1 \\ 0 & 2 \end{pmatrix}, \text{ et on a } \Pi^{2,1} = \begin{pmatrix} 1 & 2 \\ 6 & 1 \\ 0 & 2 \end{pmatrix} \text{ et } \Pi^{2,0} = \begin{pmatrix} 1 & 2 \\ 5 & 2 \\ 0 & 2 \end{pmatrix}.$$

Etant donnée une stratégie séquentielle de tirage des bras $\mathcal{A} = (A_t)$, nous pouvons introduire la *fonction valeur* associée. Pour r un entier inférieur à T, elle est définie par l'expression suivante, où la

suite des récompenses (X_t) est obtenue sous la stratégie $\mathcal{A}$:

$$V_{\mathcal{A}}(\Pi, r) = \mathbb{E}_{\boldsymbol{\mu}\sim\Pi}\left[\sum_{t=1}^{r} X_t\right].$$

Résoudre le problème de bandit bayésien revient à trouver une stratégie $\mathcal{A}$ maximisant

$$\mathbb{E}^{\pi_0}\left[\sum_{t=1}^{T} X_t\right] = V_{\mathcal{A}}(\Pi_0, T), \quad \text{avec} \quad \Pi_0 = (\pi_0, \ldots, \pi_0).$$

Or, on peut montrer qu'il existe une telle stratégie optimale, dont la fonction valeur $V^*(\Pi_0, T) = \max_{\mathcal{A}} V_{\mathcal{A}}(\Pi_0, T)$ satisfait l'équation de programmation dynamique suivante :

$$V^*(\Pi, r) \quad = \quad \max_{a=1\ldots K}\left(\mathbb{E}_{\mu_a\sim\Pi_a}[\mu_a] + \mathbb{E}_{\substack{X\sim\mathcal{B}(\mu_a)\\ \mu_a\sim\Pi_a}}\left[V^*\left(\Pi^{a,X}, r-1\right)\right]\right). \tag{1}$$

La stratégie optimale est alors $A_{t+1} = g^*(\Pi(t), T-t)$, où $g^*(\Pi, r)$ est le bras qui réalise le maximum dans (1). En principe cette solution peut donc se calculer par récurrence en utilisant (1), mais la grande taille de l'espace d'état (inclus dans $\{0, \ldots, T\}^{K\times 2}$) ne permet ce calcul que pour des horizons T très petits, et un nombre limité de bras.

Une révolution survient dans les années 1970 avec les travaux de Gittins [10], qui vont rendre pratique le calcul de la politique optimale dans un cadre un peu différent où l'objectif est de maximiser

$$\mathbb{E}^{\pi_0}\left[\sum_{t=1}^{\infty} \alpha^{t-1} X_t\right],$$

la somme des récompenses jusqu'à un horizon infini, actualisées par un coefficient $\alpha \in]0, 1[$. Gittins montre en effet que la politique optimale se réduit à une *politique d'indices*, de la forme

$$A_{t+1} = \operatorname*{argmax}_{a=1\ldots K} \mathrm{I}_\alpha(\pi_a(t)).$$

Ainsi pour chaque bras, il suffit de calculer un indice, appelé par la suite *indice de Gittins*, qui dépend des observations passées de ce bras uniquement, à travers l'a posteriori courant $\pi_a(t)$.

Dans le cadre qui nous intéresse où l'horizon est fini et où il n'y a pas d'actualisation, on peut aussi définir des *indices de Gittins à horizon fini*, $\mathrm{I}(\pi, r)$ qui dépendent du temps restant r plutôt que du paramètre α. Le lecteur intéressé pourra se référer à [9, 13] pour l'expression des indices de Gittins, avec actualisation ou à horizon fini. On peut alors définir la politique d'indices associée :

$$A_{t+1} = \operatorname*{argmax}_{a=1\ldots K} \mathrm{I}(\pi_a(t), T-t).$$

Malheureusement le théorème de Gittins ne s'applique pas en dehors du cadre actualisé [4], et cette stratégie ne coincide donc pas avec la solution de (1). Toutefois, des simulations numériques pour des courts horizons où la stratégie optimale est calculable montrent que cette politique de Gittins semble être une bonne approximation de la politique optimale [13]. D'un point de vue numérique, le calcul de ces indices de Gittins requiert la résolution de plusieurs équations de programmation dynamique, mais sur des espaces d'état bien plus petits, liés à un seul bras. Cela reste néanmoins complexe, en particulier lorsque l'horizon T est très grand, et le calcul efficace de ces indices est toujours un champ d'investigation [20]. Mais lorsque les horizons considérés sont courts, il peut être une bonne idée de mettre en œuvre la politique d'indices associée pour résoudre un problème de bandit.

3 Stratégies fréquentistes

Nous avons vu qu'il existe une solution exacte au problème de maximisation des récompenses dans un modèle de bandit bayésien, mais les méthodes présentées liées à son calcul ou son approximation sont coûteuses numériquement. Dans la cadre fréquentiste, il n'y aura pas de solution exacte, mais nous allons être en mesure de définir une notion d'optimalité *asymptotique*, et nous verrons une famille de stratégies, peu coûteuses à mettre en œuvre, qui s'approchent de l'optimalité.

Le problème de bandit a été introduit dans un cadre fréquentiste par Robbins en 1952 [21], et a ensuite été popularisé avec l'article fondateur de Lai et Robbins en 1985 [18], qui donne une caractérisation précise des bonnes stratégies d'un point de vue fréquentiste.

Tout d'abord, remarquons que si les moyennes des bras étaient connues, la stratégie optimale consisterait à ne tirer que le bras de moyenne $\mu^* = \max_a \mu_a$, et maximiser l'espérance de la somme des récompenses est équivalent à minimiser le *regret*, qui est définit comme l'écart de performance par rapport à cette stratégie oracle :

$$R_{\boldsymbol{\mu}}(T) := \mathbb{E}\left[\mu^* T - \sum_{t=1}^{T} X_t\right].$$

Le regret peut se réécrire en fonction du nombre moyen de tirages de chaque bras,

$$R_{\boldsymbol{\mu}}(T) = \sum_{a=1}^{K} (\mu^* - \mu_a)\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]. \tag{2}$$

Une bonne stratégie doit donc tirer aussi peu que possible les bras sous-optimaux (tels que $\mu^* > \mu_a$), et le résultat de Lai et Robbins nous donne une borne inférieure asymptotique sur ce nombre de tirages. Il dit que toute stratégie uniformément efficace, c'est-à-dire qui a un regret faible sur l'ensemble des problèmes de bandit ($\forall \boldsymbol{\mu} \in [0,1]^K, \forall \alpha \in]0,1],\ R_{\boldsymbol{\mu}}(T) = o(T^\alpha)$) vérifie, pour tout $\boldsymbol{\mu}$,

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

où $d(\mu_a, \mu^*)$ est la divergence de Kullback-Leibler entre les distributions $\mathcal{B}(\mu_a)$ et $\mathcal{B}(\mu^*)$, donnée par $d(x,y) = x\log(x/y) + (1-x)\log((1-x)/(1-y))$. Une borne inférieure logarithmique sur le regret s'obtient alors en utilisant (2).

Cette borne inférieure permet de définir la notion de stratégie *asymptotiquement optimale* comme une stratégie pour laquelle l'espérance du nombre de tirages d'un bras sous-optimal a est *majorée* par $\log(T)/d(\mu_a, \mu^*)$, au moins asymptotiquement. La littérature s'est alors attachée à exhiber des stratégies asymptotiquement optimales, aussi explicites que possible. La majorité des stratégies proposées sont des stratégies dites optimistes : elles reposent sur des intervalles de confiances pour chacune des moyennes construits à partir des observations disponibles, qui sont illustrés sur la figure 3. Cette figure est à mettre en regard de la figure 2, qui représente l'information utilisée par un algorithme bayésien.

Ces intervalles de confiance contiennent tous les modèles de bandits que l'on juge statistiquement possibles à l'instant courant. L'approche optimiste consiste alors à agir comme si on était dans le meilleur modèle possible, dans lequel chaque μ_a serait égal au sommet de son intervalle de confiance. On choisit alors

$$A_{t+1} = \operatorname*{argmax}_{a=1\ldots K} \mathrm{UCB}_a(t),$$

où l'indice $\mathrm{UCB}_a(t)$ est une borne de confiance supérieure sur la moyenne μ_a (Upper Confidence Bound), qui dépend donc des observations passées du bras a uniquement. On retrouve une politique d'indices, qui fait donc écho à ce qui avait été proposé dans la littérature bayésienne.

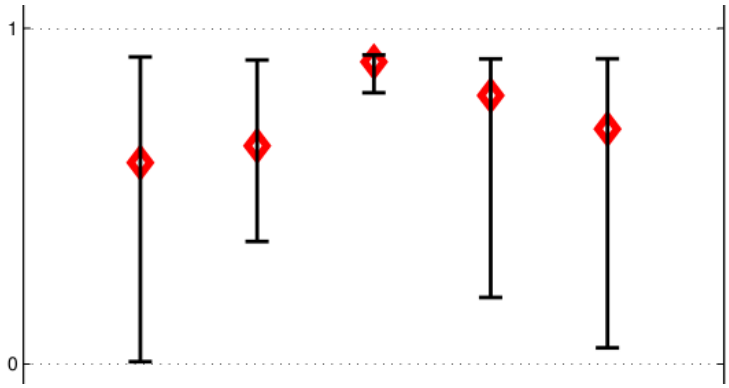

FIGURE 3 – Intervalles de confiance construits a l'aide des observations jusqu'à $T = 500$. La taille de ces intervalles est décroissante avec le nombre d'observations du bras. Les losanges rouges représentent les moyennes (inconnues) des bras.

Evidemment, la performance d'une stratégie optimiste (ou de type UCB) dépend de la manière dont les intervalles de confiance sont construits et Auer et al. proposent avec l'algorithme UCB1 [2] l'utilisation d'intervalles de confiance basés sur l'inégalité de Hoeffding :

$$\mathrm{UCB}_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{2\log(t)}{N_a(t)}},$$

où $\hat{\mu}_a(t) = S_a(t)/N_a(t)$ est la moyenne empirique des observations. Le second terme s'interprète comme un bonus d'exploration qui diminue quand le bras est tiré et augmente (grâce au $\log(t)$) quand le bras n'est pas tiré, forçant ainsi le tirage de tous les bras. En plus de proposer une stratégie simple et explicite, les auteurs proposent *une analyse à temps fini*, montrant que pour tout bras sous optimal,

$$\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq \frac{8}{(\mu^* - \mu_a)^2}\log(T) + 1 + \frac{\pi^2}{3}.$$

Comme $d(\mu_a, \mu^*) > 2(\mu^* - \mu_a)^2$, cette inégalité de permet pas de montrer que UCB1 est asymptotiquement optimal, et il ne l'est pas.

Des raffinements dans la construction des intervalles de confiance ont été successivement proposés jusqu'à l'introduction de l'algorithme KL-UCB [6]. Cet algorithme est basé sur l'indice

$$u_a(t) = \max\{q \in [0,1] : N_a(t)d(\hat{\mu}_a(t), q) \leq \log(t)\}, \tag{3}$$

où d est la fonction de divergence qui intervient dans la borne de Lai et Robbins. A l'aide de l'inégalité de Chernoff, et d'autres outils pour gérer le fait que le nombre d'observations $N_a(t)$ est lui-même aléatoire, on peut montrer que $\mathbb{P}_{\boldsymbol{\mu}}(\mu_a \leq u_a(t)) \gtrsim 1 - 1/t$, justifiant bien qu'il s'agit d'une borne de confiance supérieure. Cet indice n'a pas d'expression explicite mais une approximation peut être calculée simplement en utilisant la convexité de l'application $q \mapsto d(\hat{\mu}_a(t), q)$, comme illustré sur la figure 4

Finalement, une analyse à temps fini montre que KL-UCB vérifie, pour tout bras sous-optimal a

$$\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq \frac{1}{d(\mu_a, \mu^*)}\log(T) + O(\sqrt{\log(T)}),$$

prouvant l'optimalité asymptotique de cette stratégie.

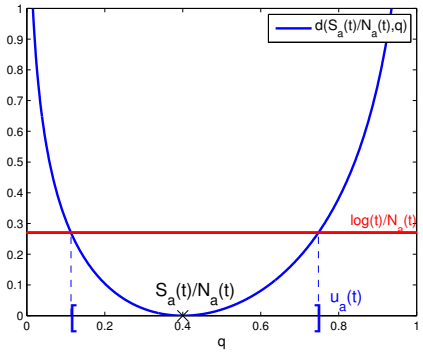

FIGURE 4 – Calcul de l'indice $u_a(t)$ utilisé par l'algorithme KL-UCB

4 Des outils bayésiens pour la minimisation du regret

Le but de cette section, qui décrit des travaux menés dans la thèse [13], est de présenter des algorithmes utilisant des outils bayésiens mais qui sont asymptotiquement optimaux du point de vue du regret, comme l'algorithme KL-UCB. Par rapport à ce dernier, les algorithmes Bayes-UCB et Thompson Sampling que nous décrivons maintenant sont plus simples d'utilisation, très efficaces empiriquement, et peuvent être aisément généralisés à d'autres types de distributions.

4.1 Bayes-UCB

Nous proposons une première manière simple d'exploiter les lois a posteriori sur les moyennes des bras, qui repose sur le principe d'optimisme, suggéré par la littérature fréquentiste. Pour chaque bras, on peut définir une région de la forme $[0, Q_a]$ à laquelle μ_a appartient avec forte probabilité sous la loi a posteriori, ce qui revient à choisir pour Q_a un quantile de la loi $\pi_a(t)$. L'action optimiste consiste alors à choisir le bras qui maximise ce quantile. Plus précisément, l'algorithme Bayes-UCB, illustré sur la figure 5, est défini par

$$A_{t+1} = \underset{a=1\ldots K}{\operatorname{argmax}}\, q_a(t) \quad \text{où} \quad q_a(t) := Q\left(\pi_a(t), 1 - \frac{1}{t}\right),$$

avec $Q(\pi, \alpha)$ le quantile d'ordre α de la distribution π, défini par $\mathbb{P}_{X\sim\pi}(X \leq Q(\pi, \alpha)) = \alpha$.

Dans les articles [15, 14], nous proposons une analyse fréquentiste de Bayes-UCB, et nous établissons en particulier le théorème suivant.

Théorème 1. *Pour tout $\epsilon > 0$, il existe des constantes C et D telles que pour tout bras sous-optimal a,*

$$\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq \frac{1+\epsilon}{d(\mu_a, \mu^*)} \log(T) + C\sqrt{\log(T)} + D.$$

Ce théorème permet de montrer que $\limsup_{T\to\infty} \mathbb{E}_{\boldsymbol{\mu}}[N_a(T)]/\log(T) \leq 1/d(\mu_a, \mu^*)$ et donc que Bayes-UCB est asymptotiquement optimal. Alors que l'algorithme KL-UCB utilise des intervalles de confiances construits avec la fonction de divergence d qui apparait dans la borne de Lai et Robbins, Bayes-UCB n'utilise pas directement cette fonction, et semble pourtant construire automatiquement les bonnes régions de confiance. La preuve du Théorème 1 est en effet basée sur le lemme suivant, qui montre

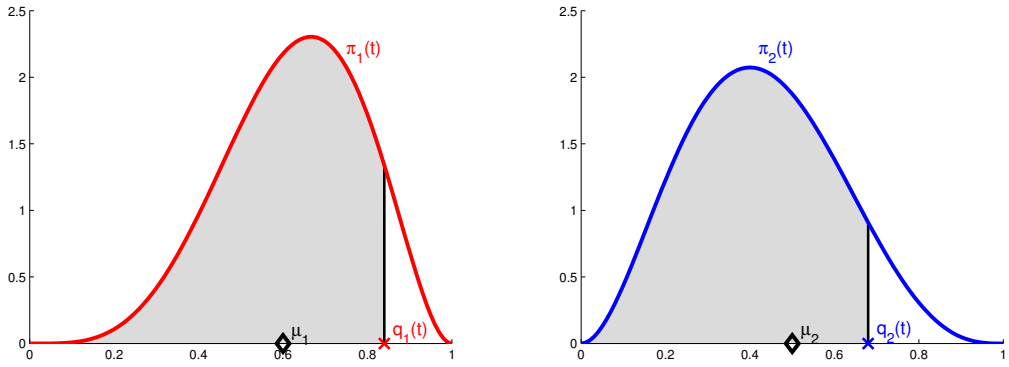

FIGURE 5 – Bayes-UCB calcule les quantiles d'ordre $1 - 1/t$ des distributions $\pi_1(t)$ et $\pi_2(t)$ et choisit ici le bras 1 car $q_1(t) \geq q_2(t)$

la proximité entre les indices utilisés par Bayes-UCB et par KL-UCB, qui nous permettra d'adapter les analyses existantes pour les stratégies optimistes.

Lemme 2. *L'indice $q_a(t)$ utilisé par Bayes-UCB vérifie $\tilde{u}_a(t) \leq q_a(t) \leq u_a(t)$, où $u_a(t)$ est l'indice (2) utilisé par KL-UCB, et $\tilde{u}_a(t)$ en est une version biaisée, définie par*

$$\tilde{u}_a(t) = \max\left\{q \in [0,1] : (N_a(t)+1)d\left(\frac{S_a(t)}{N_a(t)+1}, q\right) \leq \log\left(\frac{t}{N_a(t)+2}\right)\right\}.$$

Les lois a posteriori $\pi_a(t)$ étant des lois Beta, ce résultat provient d'un encadrement que l'on peut obtenir pour leurs quantiles, qui repose sur le contrôle suivant des queues de distribution :

$$\frac{e^{-(a+b-1)d\left(\frac{a-1}{a+b-1},x\right)}}{a+b} \leq \mathbb{P}_{X\sim\mathrm{Beta}(a,b)}(X \geq x) \leq e^{-(a+b-1)d\left(\frac{a-1}{a+b-1},x\right)}.$$

Ce dernier résultat est spécifique aux lois Beta (il provient en fait d'un lien qu'on peut établir entre lois Beta et lois binômiales, voir [15]). Pour obtenir une analyse de Bayes-UCB pour d'autres familles de distributions [14], un point crucial sera donc d'obtenir un contrôle similaire des queues des distributions a posteriori correspondantes.

4.2 L'échantillonnage de Thompson

Nous revenons maintenant sur la toute première approche introduite dans la littérature [23], qui proposait une manière très simple d'exploiter les lois a posteriori, assez différente du principe d'optimisme. L'idée de Thompson est d'utiliser une stratégie randomisée telle que la probabilité de choisir le bras a à un instant donné est égale à la probabilité a posteriori que ce bras soit optimal. L'échantillonnage de Thompson est usuellement mis en œuvre de la manière suivante. Pour chaque bras, on tire un échantillon $\theta_a(t)$ de la loi a posteriori $\pi_a(t)$, et on agit optimalement dans ce modèle échantillonné, ce qui revient à choisir le bras associé à l'échantillon le plus grand. Plus précisément

$$\begin{cases} \forall a \in \{1..K\}, \;\; \theta_a(t) \sim \pi_a(t) \\ A_{t+1} = \underset{a=1\ldots K}{\mathrm{argmax}}\, \theta_a(t). \end{cases}$$

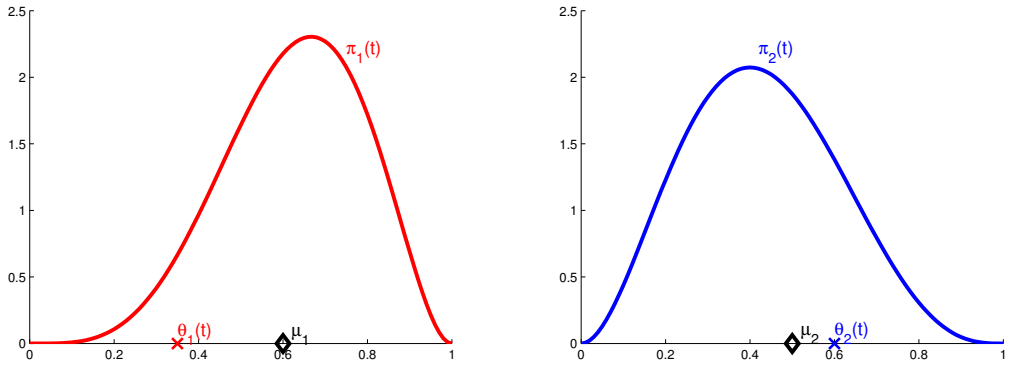

FIGURE 6 – L'algorithme tire des échantillons $\theta_1(t)$ et $\theta_2(t)$ sous chacune des lois $\pi_1(t)$ et $\pi_2(t)$, et choisit ici le bras 2 car $\theta_2(t) \geq \theta_1(t)$

L'algorithme est illustré sur la figure 6 dans un modèle de bandit à deux bras.

La force de cet algorithme réside dans son principe de base simple et générique (échantillonner un modèle de bandit possible sous la loi a posteriori, et d'agir optimalement dans ce modèle), pouvant potentiellement s'appliquer dans des modèles plus complexes. Pourtant cet algorithme, proposé en 1933, semble être tombé dans l'oubli pendant des décennies. Il a été redécouvert indépendamment par plusieurs personnes dans les années 2000 ([11, 22]), et est revenu sur le devant de la scène justement à cause de ses très bonnes performances pratiques dans des modèles de bandits plus complexes, utilisés pour des applications à la publicité en ligne [7].

Toutefois, les propriétés théoriques de l'échantillonnage de Thompson n'étaient pas encore bien comprises, et la première analyse du regret de cet algorithme a été proposée en 2012 par Agrawal et Goyal [1], donnant une borne de regret logarithmique. Ce résultat ne permettait toutefois pas de justifier l'optimalité asymptotique de l'approche, et dans l'article [16], nous avons proposé une analyse à temps fini de l'échantillonage de Thompson. Nous avons établi le théorème suivant, prouvant ainsi que l'échantillonnage de Thompson est également une stratégie asymptotiquement optimale au sens de la borne inférieure de Lai et Robbins.

Théorème 3. *Pour tout $\epsilon > 0$ et tout bras sous-optimal a, il existe des constantes C et D telles que l'échantillonnage de Thompson vérifie*

$$\mathbb{E}_{\boldsymbol{\mu}}[N_a(T)] \leq (1+\epsilon)\frac{\log(T)}{d(\mu_a, \mu^*)} + C\log\log(T) + D.$$

Nous avons cherché à proposer une preuve de ce résultat aussi proche que possible de l'analyse de Bayes-UCB. On peut en effet se ramener à cet algorithme en utilisant le fait que les échantillons tirés pour chaque bras sont, avec forte probabilité, inférieurs à certains quantiles bien choisis. Pour se ramener à cette analyse, un ingrédient supplémentaire a été nécessaire ; nous avons dû établir que l'échantillonnage de Thompson tire suffisamment souvent le bras optimal, a^*. Rappelons que tout bon algorithme de bandit doit en principe avoir tiré ce bras de l'ordre $t - O(\log(t))$ fois à l'instant t. Nous avons quantifié ce phénomène en montrant l'existence d'une constante $b \in (0,1)$ telle que la suite des probabilités $\mathbb{P}(N_{a^*}(t) < t^b)$ (indexée par t) décroit très vite. La preuve de ce résultat repose quand à elle sur la nature randomisée de l'algorithme.

5 Performances des stratégies présentées

Nous avons présenté dans cet article plusieurs algorithmes asymptotiquement optimaux du point de vue du regret, une mesure de performance fréquentiste, basés sur des outils fréquentistes (KL-UCB, basé sur des intervalles de confiances) et bayésiens (Bayes-UCB et Thompson Sampling exploitent des lois a posteriori sur les bras). Ces algorithmes ont donc le même comportement asymptotique, mais on peut raisonnablement se demander lesquels sont les plus performants pour un horizon fini T donné.

La figure 7 présente des courbes de regret pour les différents algorithmes, dont on peut noter le profil logarithmique, comme attendu. Pour les deux cas particuliers de modèles de bandit à deux bras qui y sont présentés, on remarque que le regret des deux algorithmes bayésiens est légèrement plus faible que celui de KL-UCB, et que l'algorithme UCB1 est clairement sous optimal. Du point de vue de la complexité numérique, Bayes-UCB et Thompson Sampling semblent également préférables à KL-UCB. En effet, ce dernier nécessite la résolution d'une problème d'optimisation pour le calcul de chaque indice (2), ce qui en pratique est plus long que le calcul d'un quantile ou l'obtention d'un échantillon d'une loi Beta.

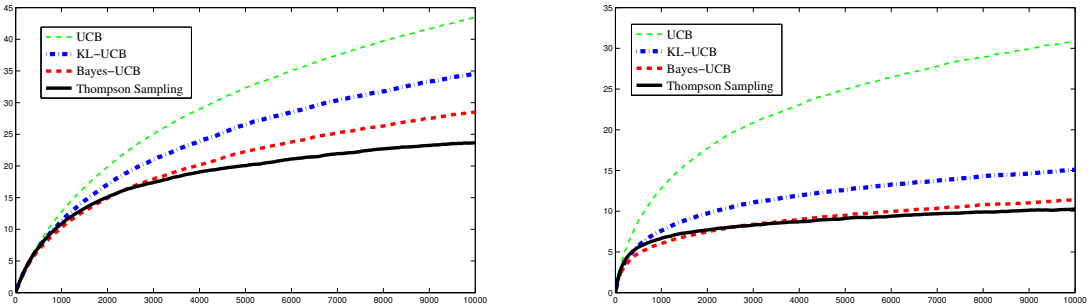

FIGURE 7 – $R_{\boldsymbol{\mu}}(T)$ en fonction de T pour $\boldsymbol{\mu} = [0.2\ 0.25]$ (gauche) et $\boldsymbol{\mu} = [0.8\ 0.9]$ (droite). Le regret est estimé à l'aide de $N = 20000$ simulations d'un jeu de bandit jusqu'à l'horizon $T = 10000$.

Enfin, l'intérêt principal des algorithmes bayésiens réside dans leur pouvoir de généralisation. Alors que l'algorithme KL-UCB est construit spécifiquement pour des bandits à récompenses binaires (il peut en fait se généraliser à d'autres types de distributions paramétrées par leur moyenne, en adaptant la divergence d utilisée, voir [6]), le principe de Bayes-UCB et Thompson Sampling est très général, est peut-être mis en œuvre dans des modèles plus complexes, dès lors qu'on est capable de définir une loi a priori sur le modèle pour lequel on peut échantillonner les lois a posteriori associées, ce qui permet aussi d'approximer les quantiles, si ceux-ci ne sont pas explicitement calculables. Dans le Chapitre 4 de [13], nous expliquons ainsi comment mettre en œuvre Bayes-UCB et Thompson Sampling dans des modèles dit contextuels, pertinents pour décrire les systèmes de recommandation. Toutefois, la compréhension théorique de ces algorithmes bayésiens dans des modèles structurés est encore très partielle, et son amélioration constitue une direction de recherche future.

Références

[1] S. Agrawal and N. Goyal. Analysis of Thompson Sampling for the multi-armed bandit problem. In *Proceedings of the 25th Conference On Learning Theory*, 2012.

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

[3] R. Bellman. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6) :503–515, 1954.

[4] D.A. Berry and B. Fristedt. *Bandit Problems. Sequential allocation of experiments*. Chapman and Hall, 1985.

[5] R.N Bradt, S.M. Johnson, and S. Karlin. On sequential designs for maximizing the sum of n observations. *Annals of Mathematical Statistics*, 27(4) :1060–1074, 1956.

[6] O. Cappé, A. Garivier, O-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3) :1516–1541, 2013.

[7] O. Chapelle and L. Li. An empirical evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems*, 2011.

[8] D. Feldman. Contributions to the "two-armed bandit". *The Annals of Mathematical Statistics*, 33(3) :947–956, 1962.

[9] J. Gittins, K. Glazebrook, and R. Weber. *Multi-armed bandit allocation indices (2nd Edition)*. Wiley, 2011.

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

[11] O.C. Granmo. Solving two-armed Bernoulli Bandit Problems using a Bayesian Learning Automaton. *International Journal of Intelligent Computing and Cybernetics*, 3(2) :207–234, 2010.

[12] W. Jouini, D. Ernst, C. Moy, and J. Palicot. Multi-armed bandit based policies for cognitive radio's decision making issues. In *ICC*, 2010.

[13] E. Kaufmann. *Analyse de stratégies bayésiennes et fréquentistes pour l'allocation séquentielle de ressources*. PhD thesis, 2014.

[14] E. Kaufmann. On bayesian index policies for sequential resource allocation. *Preprint arXiv :1601.01190*, 2016.

[15] E. Kaufmann, O. Cappé, and A. Garivier. On Bayesian Upper-Confidence Bounds for Bandit Problems. In *Proceedings of the 15th conference on Artificial Intelligence and Statistics*, 2012.

[16] E. Kaufmann, N. Korda, and R. Munos. Thompson Sampling : an Asymptotically Optimal Finite-Time Analysis. In *Proceedings of the 23rd conference on Algorithmic Learning Theory*, 2012.

[17] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *Proceedings of the 17th European Conference on Machine Learning*, 2006.

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

[19] L. Li, W. Chu, J. Langford, and R. Schapire. A contextual-bandit approach to personalized news article recommendation. In *WWW*, 2010.

[20] J. Nino-Mora. Computing a Classic Index for Finite-Horizon Bandits. *INFORMS Journal of Computing*, 23(2) :254–267, 2011.

[21] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5) :527–535, 1952.

[22] S.L. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26 :639–658, 2010.

[23] W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25 :285–294, 1933.