Analyse non asymptotique d'un test séquentiel de détection de rupture et application aux bandits non stationnaires

Lilian BESSON[1], Emilie KAUFMANN[2]

[1]CentraleSupélec (campus de Rennes) / IETR, équipe SCEE, Avenue de la Boulaie – CS 47601, 35576 Cesson-Sévigné, France

[2]CNRS & Université de Lille, Inria SequeL, CRIStAL (UMR 9189), 59000 Lille, France

Lilian.Besson@CentraleSupelec.fr, Emilie.Kaufmann@Univ-Lille.fr

Résumé – Nous étudions un test pour la détection séquentielle de rupture, basé sur le rapport de vraisemblance généralisé (GLR) et qui s'exprime en fonction de l'entropie relative binaire. Il s'applique à la détection de rupture sur la moyenne d'une distribution bornée, et nous obtenons un contôle non-asymptotique de sa probabilité de fausse alarme et de son délai de détection. Nous expliquons son utilisation pour la prise de décision séquentielle en proposant la stratégie de bandit GLR-klUCB, efficace dans des modèles de bandit stationnaires par morceaux.

Abstract – We study a strategy for online changepoint detection based on generalized likelihood ratios (GLR) and that can be expressed with the binary relative entropy. This test is used to detect a change in the mean of a bounded distribution, and we propose a non-asymptotic control of its false alarm probability and detection delay. We then explain how it can be useful for sequential decision making by proposing the GLR-klUCB bandit strategy, which is efficient in piece-wise stationary multi-armed bandit models.

1 Introduction

La détection séquentielle de rupture a fait l'objet d'études approfondies dans la communauté statistique, avec de nombreuses applications au traitement du signal [1]. Dans cet article, nous nous intéressons à la détection des rupture sur la moyenne d'une distribution de probabilité à support borné, pour laquelle nous proposons une *analyse à temps fini* d'un test séquentiel basé sur des rapports de vraisemblance généralisés.

Nous montrons que le test étudié, appelé B-GLRT, peut être appliqué à la *prise de décision séquentielle dans des environnements non stationnaires*, lorsqu'il est combiné à un algorithme de bandit à plusieurs bras. En nous appuyant sur les propriétés non-asymptotiques établies pour le test, nous montrons que la stratégie résultante, GLR-klUCB, a des garanties théoriques comparables à celles des meilleurs stratégies de l'état de l'art pour des problèmes de bandits stationnaires par morceaux. Nous proposons également une étude expérimentale montrant l'efficacité de cette stratégie activement adaptative.

Note. Cet article est une version courte de l'article [2].

2 Présentation du B-GLRT

Étant donnés des échantillons indépendants $X_1, X_2, \dots$ générés par des distributions à support dans $[0,1]$, nous cherchons à déterminer si tous les échantillons proviennent de distributions ayant une moyenne commune μ_0, ou s'il existe une *rupture* $\tau \in \mathbb{N}^*$ tel que $X_1, \dots, X_\tau$ ont une moyenne μ_0 et $X_{\tau+1}, X_{\tau+2}, \dots$ ont une moyenne différente, $\mu_1 \neq \mu_0$. Un détecteur séquentiel de rupture est un *temps d'arrêt* $\widehat{\tau}$ pour la filtration $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$, qui rejette l'hypothèse $\mathcal{H}_0 : (\exists \mu_0 \in [0,1] : \forall i \in \mathbb{N}^*, \mathbb{E}[X_i] = \mu_0)$ lorsque $(\widehat{\tau} < \infty)$.

L'utilisation en statistiques de rapports de vraisemblance généralisés (*generalized likelihood ratio*, GLR) pour des tests d'hypothèses composites remontent aux travaux de [3]. De tels outils ont été étudiés pour la détection de rupture par [4, 5]. Exploitant le fait qu'une distribution à support dans $[0,1]$ est $(1/4)$-sous gaussienne (*i.e.*, sa fonction génératrice des moments est dominée par celle d'une gaussienne de même moyenne et de variance $1/4$), le GLR gaussien, récemment étudié en profondeur par [6], peut être utilisé pour notre problème. Nous proposons plutôt d'exploiter le fait que les distributions de support $[0,1]$ sont dominées par les distributions de Bernoulli. Comme établi par [7], si ν est à support dans $[0,1]$, ν est *sous-Bernoulli* au sens suivant : ν vérifie $\mathbb{E}_{X\sim\nu}\left[e^{\lambda X}\right] \leq 1 - \mu + \mu e^{\lambda} = \mathbb{E}_{X\sim\mathcal{B}(\mu)}[e^{\lambda X}]$, où $\mathcal{B}(\mu)$ est la loi de Bernoulli de moyenne μ.

Si les échantillons (X_t) étaient tous tirés d'une distribution de Bernoulli, le problème considéré serait réduit à un test séquentiel paramétrique de $\mathcal{H}_0 : (\exists \mu_0 : \forall i \in \mathbb{N}^*, X_i \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\mu_0))$ contre l'alternative $\mathcal{H}_1 : (\exists \mu_0 \neq \mu_1, \tau > 1 : X_1, \dots, X_\tau \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\mu_0)$ et $X_{\tau+1}, \dots \overset{\text{i.i.d.}}{\sim} \mathcal{B}(\mu_1))$. La statistique GLR pour ce test basé sur un échantillon $\overline{X}$ de n observations est donnée par

$$\mathrm{GLR}(n) = \frac{\sup_{\mu_0,\mu_1,\tau<n} \ell(\overline{X};\mu_0,\mu_1,\tau)}{\sup_{\mu_0} \ell(\overline{X};\mu_0)},$$

où $\ell(\overline{X};\mu_0)$ (resp. $\ell(\overline{X};\mu_0,\mu_1,\tau)$) est la vraisemblance de l'échantillon sous un modèle dans $\mathcal{H}_0$ (resp. $\mathcal{H}_1$). Des valeurs élevées de cette statistique tendent à indiquer un rejet de $\mathcal{H}_0$.

En utilisant l'expression de la vraisemblance pour des distributions de Bernoulli, cette statistique peut s'exprimer en fonction de l'entropie relative binaire, définie sur $[0,1]^2$ par

$$\mathrm{kl}(x,y) = x\ln\big(x/y\big) + (1-x)\ln\big((1-x)/(1-y)\big),$$

et des moyennes glissantes $\widehat{\mu}_{k:k'}$, où $\widehat{\mu}_{k:k'} = \sum_{s=k}^{k'} X_s/(k'-k+1)$. On obtient $\log \mathrm{GLR}(n) = \sup_{s\in[2,n-1]}[s\,\mathrm{kl}\,(\widehat{\mu}_{1:s},\widehat{\mu}_{1:n}) + (n-s)\mathrm{kl}\,(\widehat{\mu}_{s+1:n},\widehat{\mu}_{1:n})]$. Ces calculs motivent la définition du détecteur séquentiel de rupture suivant.

Déf. 1. ***Le détecteur séquentiel de rupture B-GLRT*** *est le temps d'arrêt* $\widehat{\tau}_\delta = \inf\big\{n\in\mathbb{N}^* : \max_{s\in[2,n-1]}\big[s\,\mathrm{kl}\,(\widehat{\mu}_{1:s},\widehat{\mu}_{1:n}) + (n-s)\,\mathrm{kl}\,(\widehat{\mu}_{s+1:n},\widehat{\mu}_{1:n})\big] \geq \beta(n,\delta)\big\}$, *avec le seuil* $\beta(n,\delta)$.

L'analyse proposée dans la section suivante montre que le B-GLRT peut également être employé pour des distributions sous-Bernoulli, notamment celles à support borné dans $[0,1]$.

3 Propriétés du B-GLRT

Les propriétés asymptotiques du GLR pour la détection de rupture ont été étudiées par [5] pour les distributions de Bernoulli, et plus généralement pour les familles exponentielles à un paramètre, pour lesquelles le test GLR est défini comme en Définition 1, mais avec $\mathrm{kl}(x,y)$ remplacé par $d(x,y)$, la divergence de Kullback-Leibler entre deux éléments de cette famille exponentielle ayant pour moyennes x et y. Par exemple, le GLR gaussien étudié récemment par [6] correspond à $d(x,y) = 2(x-y)^2$, lorsque la variance est fixée à $\sigma^2 = 1/4$, et les propriétés non-asymptotiques de ce test sont données pour toute distribution $(1/4)$-sous gaussienne.

Nous fournissons de nouveaux résultats non asymptotiques pour le test B-GLRT en supposant que les échantillons (X_i) proviennent d'une distribution sous-Bernoulli, ce qui vaut notamment pour toute distribution bornée sur $[0,1]$.

Probabilité de fausse alarme. Dans la proposition 2 ci-dessous, nous proposons un choix de la fonction de seuil $\beta(n,\delta)$ sous laquelle la probabilité qu'il existe une *fausse alarme* pour des données *i.i.d.* est faible. Pour définir β, nous devons introduire une fonction de seuil $\mathcal{T}(x)$, dont l'expression explicite est donnée dans [2, §3.2] et qui vérifie $\mathcal{T}(x) \simeq x + \ln(x)$ pour x assez grand. L'utilisation de $\mathcal{T}$ pour la construction d'inégalités de concentration uniformes en temps est détaillée dans [8], et en adaptant les techniques utilisées, on peut proposer ce résultat.

Proposition 2 (Fausse alarme)**.** *Soit* $\mathbb{P}_{\mu_0}$ *un modèle probabiliste sous lequel* $X_t \in [0,1]$ *et* $\mathbb{E}[X_t] = \mu_0$ *pour tout* t. *Le test B-GLRT satisfait* $\mathbb{P}_{\mu_0}(\widehat{\tau}_\delta < \infty) \leq \delta$ *avec la fonction seuil* $\beta(n,\delta) = 2\,\mathcal{T}(\frac{\ln(3n\sqrt{n}/\delta)}{2}) + 6\ln(1+\ln(n))$.

Délai de détection. Une autre caractéristique d'un détecteur de rupture est son *délai de détection*, sous un modèle dans lequel un changement de μ_0 à μ_1 survient au moment τ. L'inégalité de Pinsker, donnant $\mathrm{kl}(x,y) \geq 2(x-y)^2$, justifie que le B-GLRT s'arrête plus tôt que le GLR gaussien basé sur la divergence quadratique $d(x,y) = 2(x-y)^2$. En s'inspirant des travaux de [6] pour le contrôle du délai du GLR gaussien, on peut alors prouver le résultat suivant, qui donne une borne supérieure en forte probabilité sur le délai de détection du B-GLRT.

Proposition 3. *Soit* $\mathbb{P}_{\mu_0,\mu_1,\tau}$ *un modèle probabiliste sous lequel* $X_t \in [0,1]$ *et* X_t *a pour moyenne* μ_0 *pour tout* $t \leq \tau$, *et* μ_1 *pour tout* $t > \tau$, *avec* $\mu_0 \neq \mu_1$, *et soit* $\Delta = |\mu_0 - \mu_1|$. *Le test B-GLRT donné en Définition 1 satisfait* $\mathbb{P}_{\mu_0,\mu_1,\tau}(\widehat{\tau}_\delta \geq \tau + u) \leq \exp(-\frac{2\tau u}{\tau+u}\left(\max[0, \Delta - \sqrt{\frac{\tau+u}{2\tau u}\beta(\tau+u,\delta)}]\right)^2)$ *pour tout* u.

Pour un seuil β choisi comme dans la proposition 2, on peut montrer qu'une conséquence de la proposition 3 est que si la rupture τ a lieu après $\Delta^{-2}\ln(1/\delta)$ échantillons, le délai de détection peut être de la même magnitude. Autrement dit, si le test B-GLRT observe assez d'échantillons avant la rupture, son délai $\widehat{\tau}_\delta$ est de l'ordre de $O(\Delta^{-2}\ln(1/\delta))$, en forte probabilité.

Les éléments d'analyse non-asymptotique présentés dans cette section sont cruciaux pour l'analyse du B-GLRT dans le cadre d'un problème de bandit, que nous présentons maintenant.

4 Application aux bandits multi-bras

Le modèle de bandit stochastique à plusieurs bras est très utilisé pour décrire des situations d'allocation séquentielle de ressources dans des environnement aléatoires. Ce modèle est donné par un ensemble de K *bras*, auxquels sont associés des lois de probabilité, pouvant par exemple modéliser l'efficacité d'un traitement dans un essai clinique, l'occupation d'un canal radio dans un système de radio intelligente, ou encore la qualité d'une recommendation dans un système de recommandation.

Un flux (aléatoire) de récompenses $(X_{i,t})_{t\in\mathbb{N}^*}$ est associé à chaque bras $i \in \{1,\dots,K\}$. Ces récompenses sont supposées bornées dans $[0,1]$. À chaque instant t, un agent sélectionne un bras $I_t \in \{1,\dots,K\}$, en se basant sur les observations passées, et reçoit la récompense correspondante $r(t) = X_{I_t,t}$. L'objectif de l'agent est de maximiser la somme de ses récompenses.

Dans le modèle standard, les récompenses du bras i sont supposées être i.i.d. mais cette hypothèse est restrictive pour de nombreuses applications, et nous considérons ici des distributions qui peuvent évoluer au cours du temps. Ainsi on note $\mu_i(t) = \mathbb{E}[X_{i,t}]$ la récompense moyenne du bras i à l'instant t et i_t^* *un* bras de moyenne maximale, *i.e.*, $\mu_{i_t^*}(t) = \max_i \mu_i(t)$, appelé *un* bras optimal. Une stratégie π choisit le prochain bras à jouer en fonction de l'historique des décisions passées et des récompenses obtenues. La performance de π est mesurée par son *regret* : la différence entre la récompense attendue obtenue par une stratégie oracle jouant un bras optimal i_t^* au temps t, et celle de la stratégie π, $R_T^\pi = \mathbb{E}[\sum_{t=1}^T \big(\mu_{i_t^*}(t) - \mu_{I_t}(t)\big)]$.

Dans le modèle stationnaire par morceaux, nous supposons en outre qu'il y a un nombre relativement petit de *ruptures*, noté $\Upsilon_T = \sum_{t=1}^{T-1} \mathbb{1}(\exists i \in \{1,\dots,K\} : \mu_i(t) \neq \mu_i(t+1))$. Nous définissons la k-ième rupture par $\tau^k = \inf\{t > \tau^{(k-1)} : \exists i : \mu_i(t) \neq \mu_i(t+1)$. Les récompenses $(X_{i,t})$ associées à chaque bras sont donc *i.i.d.* sur chaque segment $[\tau^k+1, \tau^{k+1}]$.

Notons qu'une rupture signifie qu'il y a (au moins) un bras dont la moyenne a changé, et différents scénarios sont possibles selon l'application : des ruptures peuvent arriver sur tous les bras simultanément, ou seulement sur un ou quelques bras (par exemple un seul des K canaux radio d'un réseau sans fil devient subitement utilisé par un appareil communicant à haut débit).

Les premières approches proposées pour les bandits stationnaires par morceaux combinent un algorithme de bandit et un mécanisme d'*oubli des récompenses* : l'utilisation d'un facteur d'oubli (discount, D-UCB [9]) ou d'une fenêtre glissante (SW-UCB [10]). Si le facteur d'oubli ou la fenêtre sont choisis en connaissant T et le nombre de ruptures Υ_T, on peut obtenir un regret de $\mathcal{O}(\sqrt{\Upsilon_T T \ln(T)})$ [10]. Ces approches dites passivement adaptatives sont empiriquement moins efficaces que des approches *activement adaptatives* proposées plus récemment. Ces dernières combinent un algorithme de bandit et un algorithme de détection séquentielle de rupture observant les récompenses de chaque bras, pour réinitialiser la mémoire des observations d'un ou de tous les bras dès qu'une rupture est détectée. Deux algorithmes récents, CUSUM-UCB [11] et Monitored UCB (M-UCB, [12]) ont un regret en $\mathcal{O}(\sqrt{\Upsilon_T T \ln(T)})$, et une performance pratique bien supérieure à celle de SW-UCB.

Nous proposons un nouvel algorithme de ce type, appelé GLR-klUCB, qui combine l'algorithme klUCB, connu pour être optimal pour les bandits à distributions de Bernoulli [7], avec le détecteur de rupture B-GLRT. Lorsqu'une rupture est détectée sur un bras, on peut envisager de réinitialiser la mémoire de ce bras uniquement (comme CUSUM-UCB) ou celle de tous les bras (comme M-UCB). Nous proposons dans [2] la première analyse conjointe des deux approches mais par soucis de simplicité, nous présentons dans cette section uniquement les résultats obtenu pour GLR-klUCB basé sur des réinitialisations globales, défini formellement comme l'algorithme 1 ci-contre.

GLR-klUCB repose sur le calcul d'un *indice* (de type klUCB [7]) pour chaque bras, et choisit en général le bras d'indice le plus élevé. Soit $\tau_i(t)$ l'instant de la dernière réinitialisation pour le bras i avant le temps t, $n_i(t) = \sum_{s=\tau_i(t)+1}^{t} \mathbb{1}(A_s = i)$ le nombre de sélections et $\widehat{\mu}_i(t)$ la moyenne empirique des récompenses obtenues entre la dernière réinitialisation et l'instant t. L'indice du bras i à l'instant t est défini par $\mathrm{UCB}_i(t) = \max\{q \in [0,1] : n_i(t) \times \mathrm{kl}(\widehat{\mu}_i(t), q) \leq f(t - \tau_i(t))\}$ (l. 6) avec la fonction d'exploration $f(t) = \ln(t) + 3\ln(\ln(t))$. La détection d'une rupture sur le bras qui vient d'être joué produit une réinitialisation (l. 12), si $\mathrm{BGLRT}_\delta(Z_1, \dots, Z_n) = \text{Vrai}$ ssi $\sup_{1<s<n} \big[s \times \mathrm{kl}\big(\frac{1}{s}\sum_{i=1}^{s} Z_i, \frac{1}{n}\sum_{i=1}^{n} Z_i\big) + (n-s) \times \big(\frac{1}{n-s}\sum_{i=s+1}^{t} Z_i, \frac{1}{n}\sum_{i=1}^{n} Z_i\big)\big] \geq \beta(n,\delta)$, avec $\beta(n,\delta)$ défini dans la proposition 2, ou $\beta(n,\delta) = \ln(3n^{3/2}/\delta)$, comme nous le recommandons en pratique. Comme les approches proposées précédemment, GLR-klUCB utilise de l'*exploration forcée* paramétrée par $\alpha \in (0,1)$ (lignes 5-6), qui assure que chaque bras est suffisamment échantillonné, afin que les ruptures puissent aussi être détectées sur les bras actuellement sous-échantillonnés par l'algorithme de bandit.

Garanties théoriques. Soit τ^k la position de la k-ième rupture et soit μ_i^k la moyenne du bras i sur le segment $[\tau^k, \tau^{k+1}]$. Nous introduisons aussi $k^* \in \arg\max_i \mu_i^k$ et le plus grand écart à la rupture k comme $\Delta^k = \max_{i=1,\dots,K} |\mu_i^k - \mu_i^{k-1}| > 0$.

1 **Données :** *Paramètres du problème* : $T \in \mathbb{N}^*$, $K \in \mathbb{N}^*$;
2 **Données :** *Paramètres de l'algorithme* : $\alpha \in (0,1)$, $\delta > 0$;
3 **Initialisation :** $\forall i \in \{1,\dots,K\}$, $\tau_i \leftarrow 0$ et $n_i \leftarrow 0$
4 **pour** $t = 1, 2, \dots, T$ **faire**
5 **si** $t \mod \left\lfloor \frac{K}{\alpha} \right\rfloor \in \{1,\dots,K\}$ **alors**
6 $A_t \leftarrow t \mod \left\lfloor \frac{K}{\alpha} \right\rfloor$ `// exploration forcée`
7 **sinon**
8 $A_t \leftarrow \arg\max_{i \in \{1,\dots,K\}} \mathrm{UCB}_i(t)$
9 **fin**
10 Jouer le bras A_t, $n_{A_t} \leftarrow n_{A_t} + 1$
11 Observer une récompense $X_{A_t,t} : Z_{A_t,n_{A_t}} \leftarrow X_{A_t,t}$
12 **si** $\mathrm{BGLRT}_\delta(Z_{A_t,1}, \dots, Z_{A_t,n_{A_t}}) =$ *Vrai* **alors**
13 $\forall i, \tau_i \leftarrow t$ and $n_i \leftarrow 0$. `// réinitialisation`
14 **fin**

Algorithme 1 : GLR-klUCB.

Hypothèse 4. *Soit* $d^k = d^k(\alpha,\delta) = \lceil \frac{4K}{\alpha(\Delta^k)^2}\beta(T,\delta) + \frac{K}{\alpha} \rceil$. *Nous supposons que toutes les séquences sont suffisamment longues :* $\forall k \in \{1,\dots,\Upsilon_T\}$, $\tau^k - \tau^{k-1} \geq 2\max(d^k, d^{k-1})$.

L'hypothèse 4 demande que deux ruptures consécutives soient suffisamment éloignées : leur distance dépend de l'ampleur du changement le plus important qui se produit à ces deux ruptures (*i.e.*, Δ^k et Δ^{k+1}). Sous cette hypothèse, nous fournissons une borne à temps fini du regret, dépendant explicitement des paramètres du problème. La borne utilise les divergences $\mathrm{kl}(\mu_i^k, \mu_{k^*}^k)$ exprimant la difficulté du problème de bandit (stationnaire) entre deux ruptures, les termes Δ^k exprimant la difficulté du problème de détection du rupture, et les paramètres α et δ. Nous simplifions ici le Théorème 5 de [2], en isolant les deux termes dominants et sans expliciter les constantes.

Théorème 5. *Soit* $\alpha = \sqrt{\Upsilon_T \ln(T)/T}$ *et* $\delta = 1/\sqrt{\Upsilon_T T}$. *Pour tout problème satisfaisant l'Hypothèse 4, le regret de* GLR-klUCB *avec les paramètres* α *et* δ *satisfait*

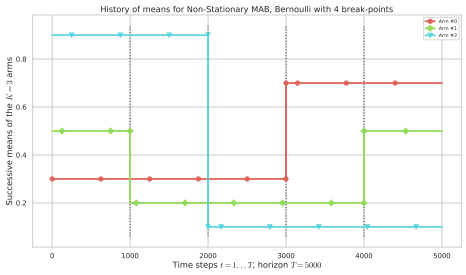
$$R_T = \mathcal{O}\left(\frac{K}{(\Delta^{change})^2} \sqrt{\Upsilon_T T \ln(T)} + \frac{(K-1)}{\Delta^{opt}} \Upsilon_T \ln(T) \right),$$

où Δ^{opt} *est la plus petite valeur de l'écart de sous-optimalité sur chaque segment stationnaire, et* $\Delta^{change} = \min_k \Delta^k$.

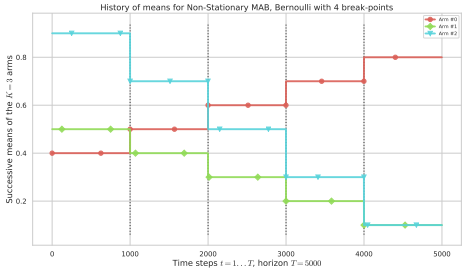
Avantages de notre approche. Les garanties théoriques proposées pour CUSUM et M-UCB exigent des paramètres choisis avec une *certaine connaissance préalable* sur les moyennes des bras. Si les tests utilisés par ces deux algorithmes utilisent un seuil h qui doit être calibré en fonction de T et Υ_T, ce qui est aussi requis par les approches précédentes pour obtenir de bonnes garanties, les paramètres ε pour CUSUM et w pour M-UCB nécessitent la connaissance de Δ^{change}, la plus petite valeur d'un changement. Nous proposons le premier algorithme qui atteint un regret en $\mathcal{O}(\sqrt{\Upsilon_T T \ln(T)})$ *sans exiger cette connaissance*. Enfin, d'un point de vue pratique, même si le test B-GLRT proposé est plus complexe à mettre en œuvre que le test utilisé par M-UCB, nous proposons deux heuristiques pour l'accélérer sans réduire ses performants en terme de regret.

5 Simulations et interprétation

Cette section présente les résultats d'une étude expérimentale comparant GLR-klUCB à d'autres approches pour des exemples (synthétiques) de problèmes de bandits stationnaires par morceaux. La Figure 1 présente les moyennes des bras pour deux problèmes stationnaires par morceaux à $K = 3$ bras et $\Upsilon = 4$ ruptures pour un horizon $T = 5000$. Les récompenses sont générées sous une loi de Bernoulli. Notons que dans le problème 1, un seul bras change à chaque rupture, contre tous les bras dans le problème 2. Nous reportons également des résultats pour un troisième problème, présenté en Section 6 de [2] et inspiré de [12], pour lequel des données synthétiques ont été obtenues à partir de manipulations, depuis une base de données de clics d'utilisateurs sur la page d'accueil de *Yahoo!*. Les simulations ont été effectuées avec SMPyBandits [13].



(a) **Problème 1** : 4 ruptures se produisent sur un seul bras.



(b) **Problème 2** : 4 ruptures se produisent sur tous les bras.

Figure 1 – Moyennes $\mu_i(t)$ des bras pour deux problèmes.

Le tableau 1 montre le regret final R_T obtenu pour différents algorithmes, estimé à partir de 1000 répétitions indépendantes. Nous évaluons deux variantes de GLR-klUCB, basées sur des réinitialisations globales (algorithme 1) et locales, un algorithme de bandit classique (klUCB), un oracle effectuant une réinitialisation de tous les bras au moment des ruptures, et différents algorithmes passivement ou activement adaptatifs. On observe que les meilleures stratégies réalistes sont activement adaptatives (M-UCB, CUSUM-UCB et GLR-klUCB) et parmi celles-ci GLR-klUCB est souvent la plus efficace.

Une direction de recherche future sera de comprendre et de justifier l'observation surprenante suivante : la variante à réinitialisations locales semble toujours obtenir de meilleures performances que celle à réinitialisations globales, même sur des problèmes comme le pb. 2 où chaque rupture est globale.

Algorithmes \ Problèmes	Pb 1	Pb 2	Pb 3
Oracle-Restart klUCB	$\mathbf{37 \pm 37}$	$\mathbf{45 \pm 34}$	$\mathbf{257 \pm 86}$
klUCB	270 ± 76	162 ± 59	529 ± 148
Discounted-klUCB	1456 ± 214	1442 ± 440	1376 ± 37
SW-klUCB	177 ± 34	182 ± 34	1794 ± 71
M-klUCB	290 ± 29	534 ± 93	645 ± 141
CUSUM-klUCB	148 ± 32	152 ± 42	$\mathbf{490 \pm 133}$
GLR-klUCB (Local)	$\mathbf{74 \pm 31}$	$\mathbf{113 \pm 34}$	513 ± 97
GLR-klUCB (Global)	97 ± 32	134 ± 33	621 ± 103

Table 1 – Regret moyen $\pm$ 1 écart-type, pour différents algorithmes sur les pb. 1, 2 (avec $T = 5000$) et 3 ($T = 20000$).

6 Conclusion

Nous avons proposé une nouvelle analyse non asymptotique d'un test de détection séquentielle de rupture, le B-GLRT, pour des données issues de distributions à support borné. Notre analyse contrôle la probabilité de fausse alarme et le délai de détection de ce test, et nous permet d'en proposer une combinaison avec l'algorithme de bandit klUCB. Nous montrons que l'algorithme résultant est efficace pour le problème de bandit stationnaire par morceaux. GLR-klUCB a un regret moyen et des performances empiriques compétitives avec l'état de l'art, sans autre connaissance préalable que celle du nombre de ruptures.

References

[1] M. Basseville, I. Nikiforov, *et al.*, *Detection of Abrupt Changes: Theory And Application*. 1993.

[2] L. Besson and E. Kaufmann, "Combining the Generalized Likelihood Ratio Test and kl-UCB for Non-Stationary Bandits." Preprint, hal.archives-ouvertes.fr/hal-02006471, February 2019.

[3] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

[4] D. Siegmund and E. Venkatraman, "Using the Generalized Likelihood Ratio Statistic for Sequential Detection of a Change Point," *The Annals of Statistics*, pp. 255–271, 1995.

[5] T. Lai and H. Xing, "Sequential change-point detection when the pre-and post-change parameters are unknown," *Sequential Analysis*, vol. 29, 2010.

[6] O.-A. Maillard, "Sequential change-point detection: Laplace concentration of scan statistics and non-asymptotic delay bounds," in *ALT*, 2019.

[7] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz, "Kullback-Leibler Upper Confidence Bounds For Optimal Sequential Allocation," *Annals of Statistics*, vol. 41(3), 2013.

[8] E. Kaufmann and W. Koolen, "Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals." arXiv, 2018.

[9] L. Kocsis and C. Szepesvári, "Discounted UCB," in *2nd PASCAL Challenges Workshop*, 2006.

[10] A. Garivier and E. Moulines, "On Upper-Confidence Bound Policies For Switching Bandit Problems," in *ALT*, 2011.

[11] F. Liu, J. Lee, and N. Shroff, "A Change-Detection based Framework for Piecewise-stationary Multi-Armed Bandit Problem," in *AAAI*, 2018.

[12] Y. Cao, W. Zheng, B. Kveton, and Y. Xie, "Nearly Optimal Adaptive Procedure for Piecewise-Stationary Bandit: a Change-Point Detection Approach," in *AISTATS*, 2019.

[13] L. Besson, "SMPyBandits: an Open-Source Research Framework for Single and Multi-Players Multi-Arms Bandits (MAB) Algorithms in Python," 2018. See SMPyBandits.GitHub.io/.