Bandits-Manchots Contextuels : Précision Globale Versus Individuelle

Nicolas Gutowski, Tassadit Amghar, Olivier Camp, Fabien Chhel

Bandits-Manchots Contextuels : Précision Globale Versus Individuelle

Nicolas Gutowski[1,2] Tassadit Amghar[2] Olivier Camp[1] Fabien Chhel[1]

[1] ESEO-TECH
10 boulevard Jean Jeanneteau, 49100 Angers, France

[2] LERIA, Université d'Angers (UBL)
2 boulevard Lavoisier, 49000 Angers, France

nicolas.gutowski@eseo.fr

Résumé

Dans la littérature, la plupart des travaux sur les bandits manchots sont évalués à l'aide d'une mesure de la précision globale. Concernant les bandit-manchots contextuels, les approches existantes ont pour objectif d'atteindre une personnalisation individuelle. Ainsi, leur précision globale devrait refléter la précision individuelle pour chacun des utilisateurs. Afin de mesurer le niveau de personnalisation atteint par ces approches, nous avons défini une nouvelle évaluation comparant les précisions individuelles des recommandations faites à chaque utilisateur avec la précision globale. Sur la base de cette comparaison, démontrant des disparités entre la précision individuelle et la moyenne de précision globale, nous proposons *Sliding Window LinUCB (SW-LinUCB)*. *SW-LinUCB* est une combinaison de *LinUCB* (CMAB) et d'un mécanisme de diversification pénalisant les bras sélectionnés trop fréquemment. Notre approche, inspirée d'applications réelles, comme les systèmes de recommandation, ne nécessite pas uniquement d'atteindre une bonne précision globale mais doit aussi tenir compte de la précision individuelle. Nous expérimentons et discutons nos résultats sur plusieurs jeux de données réelles.

Mots Clefs

Apprentissage par renforcement, LinUCB, Bandits-manchots contextuels, Système de recommandation

Abstract

Most works on Multi-Armed Bandits (MAB) focus the evaluations of their methods on a global accuracy performance metric. In the case of Contextual Multi-Armed Bandit (CMAB), the existing algorithms claim to eventually provide full personalization, which might suggest that their global accuracy metric should reflect each user's individual accuracy. In order to verify this, we consider a novel approach of CMAB assessment focused on the evaluation of individual accuracy and compare it to global accuracy. Based on the results of this comparison highlighting some users far from the average global accuracy, we propose *Sliding Window LinUCB* (*SW-LinUCB*), a combination of the original *LinUCB* (CMAB) and a diversification mechanism penalizing arms which are pulled too frequently. It is motivated by the requirements of different real-world applications such as clinical trials or recommender systems, which must converge to a good global accuracy and should equally distribute it among individuals. We experiment and discuss the benefits and losses of the proposed method on several real-world datasets.

Keywords

Recommendation System, Reinforcement learning, LinUCB, Contextual Multi-armed Bandits

1 Introduction

De nos jours, les bandits-manchots contextuels (Contextual Multi-Armed Bandit : CMAB) sont très largement considérés par de nombreuses applications se heurtant à des problèmes de décision séquentielle e.g., les systèmes de recommandation [1], ou encore les essais cliniques [2]. À chaque itération, les algorithmes d'apprentissage de CMAB ont pour objectif de choisir une action optimale (tirer le bras optimal) parmi un ensemble de possibilités, en tenant compte du contexte donné et des récompenses passées obtenues en regard de ces actions. Dans la majorité des cas, les récompenses retournées sont égales à 1 si l'algorithme réalise une bonne classification du contexte par rapport à l'action choisie, et à 0 sinon [1]. Ainsi, la plupart des études évaluent la performance de leurs méthodes, jusqu'à un horizon final, à travers une mesure de précision globale e.g., récompense moyenne, cumul des récompenses, ou nombre de regrets total [1, 3, 4]. Néanmoins, de telles métriques semblent inadéquates pour déterminer la précision à associer à chaque contexte [5]. Ainsi, si nous prenons l'exemple des systèmes de recommandation pour lesquels les utilisateurs peuvent être des visiteurs réguliers, ou des abonnés (exemple : applications mobiles), il semble

essentiel de tenir compte de leurs retours individuels en regard des recommandations qui leur sont faites.
De telles considérations nous ont amené à définir une méthode permettant de mesurer la précision individuelle, et à évaluer différents algorithmes existants au regard de notre nouvelle métrique. Le problème se pose alors comme étant double objectif puisqu'il convient de maximiser la précision individuelle tout en maintenant une bonne précision globale.
Ainsi, nous introduisons une mesure complémentaire de la performance qui est fondée sur l'observation de la fonction de distribution cumulative (Cumulative Distribution Function : CDF) de la précision individuelle. De plus, nous proposons l'adaptation d'une méthode de CMAB que nous nommons *SW-LinUCB*. Cette méthode est bâtie à partir de l'algorithme classique *LinUCB* combiné avec une fenêtre glissante. De ce fait, *SW-LinUCB* a pour objectif d'améliorer la précision individuelle en incorporant un mécanisme de diversification, tout en conservant une précision globale satisfaisante.
En nous appuyant sur des jeux de données d'applications réelles, nous étudions les performances de différentes méthodes existantes (MAB et CMAB) à travers notre mesure de CDF sur l'ensemble des contextes observés. Nous montrons que pour les deux méthodes — pour les CMAB, spécifiquement selon le niveau de parcimonie du vecteur de contexte — un fossé se creuse entre les utilisateurs pour lesquels nous obtenons un haut niveau de précision et ceux défavorisés par un très faible niveau de précision, malgré une précision globale satisfaisante. Nos expériences montrent que *SW-LinUCB* réussit à combler ce fossé tout en maintenant une précision globale qui ne décroit pas plus de 10% par rapport à celle d'origine. En outre, on constate que notre algorithme parvient à atténuer les effets négatifs d'un contexte insuffisamment renseigné.
La contribution de notre article est double : 1) nous proposons une nouvelle mesure de l'évaluation basée sur la précision individuelle ; 2) nous présentons *SW-LinUCB* qui est une adaptation de *LinUCB* dont l'objectif est de maximiser la précision individuelle tout en conservant une précision globale satisfaisante.
Cet article est organisé comme suit : la section 2 présente un état de l'art sur les problèmes de MAB et les différentes métriques utilisées. La section 3 introduit les travaux connexes sur les problématiques de CMAB, l'algorithme *LinUCB* et l'usage de fenêtres glissantes. La section 4 dresse notre problématique et la méthode que nous mettons en place. Dans la section 5 nous exprimons et discutons les résultats de nos expérimentations. Enfin, nous concluons et présentons les perspectives de notre travail dans la section 6.

2 Contexte

Le problème du Bandit-Manchot (Multi-Armed Bandit : MAB) est un sujet qui a suscité de nombreuses recherches depuis sa première formalisation en 1952 [6]. De nombreuses formulations ont pu être proposées : stochastiques [7, 8, 4], ou encore Bayésiennes [9]. Plus précisément, le défi pour tout problème de MAB consiste à construire une stratégie visant à tirer le bras optimal sans connaissance préalable de la rentabilité de chacun des bras disponibles. La résolution de ce problème consiste à trouver un compromis entre l'exploration de l'ensemble des bras pour en déduire leurs rentabilités et l'exploitation de ce qui a été inféré pour favoriser la sélection des bras optimaux. Une version étendue de ce problème prend en compte le contexte. Il s'agit du problème de Bandit-Manchot Contextuel (Contextual Multi-Armed Bandit : CMAB) [10, 3]. Ainsi, dans une approche CMAB, le défi visant à déterminer le bras optimal reste le même que pour un problème de MAB mais doit tenir compte du contexte des utilisateurs.

Dans la littérature, le critère le plus fréquemment observé pour mesurer la performance d'un algorithme de bandit reste la précision globale — i.e. le nombre de fois qu'une récompense positive a été obtenue en tirant les différents bras [7, 8, 1, 3, 4]. Néanmoins, en fonction du domaine dans lequel les bandits sont appliqués, l'évaluation de leur performance peut nécessiter de s'ouvrir à d'autres critères. En effet, comme c'est tout particulièrement le cas pour les systèmes de recommandations, il a été observé dans certaines études que les mesures de précision ne sont pas suffisamment adaptées et pourraient être préjudiciables et nuire à la satisfaction des utilisateurs [5]. De ce fait, même si des algorithmes de CMAB tels que *LinUCB* [1] ou encore *Contextual Thompson Sampling* [3] permettent à terme une personnalisation complète auprès de chaque utilisateur, une autre étude soutient en revanche que ceux-ci nécessitent un si grand nombre d'itérations pour atteindre cette personnalisation qu'ils risquent de causer la frustration des utilisateurs avant d'y parvenir [11].

De telles constats ont conduit des recherches sur les CMAB et les systèmes de recommandations vers deux directions : 1) Tenter de réduire le nombre d'itérations nécessaires pour atteindre la personnalisation pour chaque utilisateur [11, 12], 2) Prendre en considération d'autres critères d'évaluation de la performance comme : la qualité [13], la diversité et la nouveauté [14], la couverture et la sérendipité [15], ou encore la satisfaction utilisateur [16].

La majorité des travaux tendent à montrer que la diversification serait l'un des points-clé pour améliorer la satisfaction utilisateur. Par exemple cela permettrait de mieux répondre aux besoins éphémères des utilisateurs [17], de les aider à découvrir de nouveaux éléments [13], ou d'éviter les recommandations redondantes [18].

À notre connaissance, aucune approche n'aborde spécifiquement le problème de la recherche d'un compromis entre précision individuelle et précision globale pour les CMAB à travers l'usage de techniques de diversification. Ceci constitue l'objectif principal de notre travail.

3 Travaux Antérieurs

Cette section présente les concepts clés sous-jacents à notre approche : le problème de CMAB, l'algorithme *LinUCB* qui le résout, et un principe de diversification s'appuyant sur une fenêtre glissante.

3.1 Bandits-Manchots Contextuels

Les approches contextuelles du problème de bandits-manchots (CMAB) [10] ont été très largement étudiées via des méthodes telles que *LinUCB* [1], Contextual Thompson Sampling *CTS* [3] ou encore Neural Bandit [19]. Ces méthodes résolvent le problème de CMAB en supposant une dépendance linéaire entre la récompense attendue d'une action et son contexte. Selon les travaux de Langford [10], le problème de CMAB peut être défini comme suit : Soit $A = \{a_1, ..., a_k\}$ un ensemble donné de k bras indépendants. Soit $X \subseteq \mathbb{R}^d$ l'ensemble de vecteurs de contexte de dimension d caractérisant un utilisateur et son environnement e.g., $x \in X$ est un vecteur binaire codant les caractéristiques telles que : l'âge, le sexe, le métier, les préférences, les spécialités, la localisation ou encore les caractéristiques des bras eux-mêmes. Soit l'horizon $T \in \mathbb{N}^*$, à chaque itération $t \in [1, T]$, le contexte x_t incluant l'utilisateur, est pris en considération afin de permettre la sélection du bras optimal compte tenu des récompenses obtenues lors des itérations précédentes. Pour chaque itération t, soit $r_t = (r_{t,a_1}, ..., r_{t,a_k})$ le vecteur de récompense où r_{t,a_i} correspond à la récompense obtenue après avoir sélectionné le bras a_i et $r_{t,a_i} \in \{0, 1\}$ dans notre cas où les récompenses sont tirées depuis des distributions de Bernoulli. Soit $\mathcal{D}_{x,r}$ la distribution conjointe entre les contextes x et les récompenses r, et soit $\theta_{t,a}$ le vecteur de coefficients inconnu (restant à déterminer) associé au bras a à l'itération t. Nous supposons que les récompenses attendues d'un bras a à l'itération t est une fonction linéaire du vecteur de contexte x_t de dimension d tel que $\mathbb{E}[r_{t,a}|x_t] = \widehat{\theta}_a^\top x_t$ où $\widehat{\theta}_a$ représente le vecteur de coefficients estimé associé au bras a. Ainsi, soit $\Pi : X \to A$ l'ensemble des politiques possibles où la politique optimale devant être déterminée est $\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{r,x}[r_{t,\pi(x)}]$. Alors, soit $\pi_t \in \Pi$ la politique empruntée par un algorithme de CMAB $\mathcal{A}$ à l'itération t. Par conséquent, dans le cadre d'un environnement stationnaire où $\mathcal{D}_{x,r}$ ne varie pas, le pseudo-regret instantané à l'itération t peut alors être défini tel que $\rho_t(\mathcal{A}) = \mathbb{E}_{r,x}[r_{t,\pi^*(x_t)} - r_{t,\pi(x_t)}]$ et le pseudo-regret cumulé tel que $\rho(\mathcal{A}) = \sum_{t=1}^{T} \rho_t(\mathcal{A})$.
Les algorithmes tels que *LinUCB* [1] ou *Contextual Thompson Sampling* (*CTS*) [3] ont été modélisés et largement étudiés afin de résoudre ce problème de CMAB. Aussi, à la section suivante nous rappelons l'un des plus populaires d'entres eux : *LinUCB*.

3.2 LinUCB

Nous avons d'abord décidé de bâtir et d'expérimenter notre approche à partir de *LinUCB* [1] qui reste l'un des algorithmes de CMAB les plus célèbres présentés dans la littérature.

LinUCB [1] est un algorithme contextuel à bornes supérieures de confiance qui renforce rapidement la sélection des bras optimaux en ajoutant un *bonus* (l'écart de la récompense) au gain total calculé. À chaque itération t, *LinUCB* sélectionne le bras $a \in A$ avec le gain calculé $p_{t,a}$ maximum parmi l'ensemble des bras disponibles. $p_{t,a}$ est construit à partir d'une combinaison linéaire du coefficient $\theta_{t,a}$ et du vecteur de caractéristiques x_t auxquels vient s'ajouter l'écart de récompense qui représente la valeur d'action optimiste du gain obtenu. Le vecteur de coefficient $\widehat{\theta_a}$ est construit à partir de la matrice D_a de dimension $n \times d$ (n recommandations en correspondance de d caractéristiques), et $b_a \in \mathbb{R}^d$ représente le vecteur de réponse correspondant, dont les poids pour chaque dimension sont fonction des récompenses obtenues. Plus précisément, $\widehat{\theta_a} = (D_a^\top D_a + I_d)^{-1} b_a$ où I_d représente la matrice identité de dimension $d \times d$. Par conséquent, à chaque itération t, *LinUCB* sélectionne le bras a_t tel que $a_t = \arg\max_{a \in A} p_{t,a}$ où $p_{t,a} = \widehat{\theta_a}^\top x_t + \alpha\sqrt{x_t^\top (D_a^\top D_a + I_d)^{-1} x_t}$. Ainsi, $\widehat{\theta_a}^\top x_t$ représente l'espérance de récompense et $\alpha\sqrt{x_t^\top (D_a^\top D_a + I_d)^{-1} x_t}$ l'écart de récompense où α est un paramètre pouvant être considéré comme un critère de robustesse face au bruit. De plus, selon [20], il y a une probabilité d'au moins $1 - \delta$ que $\mid \widehat{\theta}_a^\top x_t - \mathbb{E}[r_{t,a}|x_t] \mid \leq \alpha\sqrt{x_t^\top (D_a^\top D_a + I_d)^{-1} x_t}$ avec $\alpha = 1 + \sqrt{ln(2/\delta)/2}$. Si un ensemble de bras contient k bras, alors la borne supérieure du regret sera en $\tilde{\mathcal{O}}\left(\sqrt{kdT}\right)$.

Néanmoins, même si le regret total est ici bien identifié et que sa borne supérieure a été démontrée, il est encore nécessaire de surmonter les problématiques de faible précision individuelle possiblement induite par des environnements non stationnaires ou par des contextes trop pauvres en informations.

Dans la section suivante, nous rappelons plusieurs mécanismes de diversification reposant sur l'utilisation de fenêtres glissantes.

3.3 Mécanismes de diversification

Dans les systèmes de recommandation, la diversité est pertinente pour la satisfaction individuelle. La diversification peut aussi trouver son intérêt dans le cadre d'environnements non-stationnaires afin de permettre à l'algorithme de rester à jour et favoriser les observations les plus récentes. À l'aide d'une fenêtre glissante, des algorithmes tels que *SW-UCB* [21], ou encore *Windows Thompson Sampling with Restricted Context (Windows TSRC)* [12] permettent d'atténuer les effets résultant de la non-stationnarité. De plus, pour résoudre ces mêmes problèmes induits par la non-stationnarité et plus particulièrement dans le cadre d'une problématique de bandits de type *restless* [22], il existe une approche utilisant également une fenêtre glissante et dont l'objectif est de pénaliser les bras qui ont été

tirés trop souvent [23]. Cette approche intéressante a inspiré notre proposition.

4 Problématique et Méthodes

Dans cette section, nous posons notre problème, puis nous définissons notre nouvelle approche *SW-LinUCB*. Notre méthode est basée sur la combinaison de l'algorithme original *LinUCB* [1] et l'utilisation d'une fenêtre glissante inspirée de [23].

4.1 Énoncé du problème

Soit $\mathcal{U} = \{u_1, ..., u_n\}$ l'ensemble des n agents disponibles dans un problème de bandits, et pouvant par exemple correspondre dans le cadre d'applications réelles, à des utilisateurs ou encore des patients. Inspiré par [24], nous supposons pour chaque bras $a \in A$ et étant donné $x \in X \subseteq \mathbb{R}^d$, que $\mathcal{U}$ peut être partitionné en un nombre $m_a(x)$ de clusters $\mathcal{U}_{1,a}(x), \mathcal{U}_{2,a}(x), ..., \mathcal{U}_{m_a(x),a}(x)$ d'utilisateurs partageant les mêmes comportements vis à vis des récompenses qu'ils octroient à chaque bras a. Faisant maintenant l'hypothèse de l'existence d'un vecteur de contexte optimal $x^* \in X^*$ qui posséderait toutes les caractéristiques pertinentes associées, avec une confiance de 100%, au bras optimal correspondant et cela pour chaque contexte disponible. Comme *LinUCB* suppose une dépendance linéaire entre la récompense attendue d'une action et son contexte tel que $\mathbb{E}[r_{t,a}|x_t] = \widehat{\theta}_a^\top x_t$, alors lorsque x^* est fourni, *LinUCB* converge vers une précision de 100% et offre une personnalisation pour chaque individu. Cela signifie que toute précision individuelle convergera également vers 100%. Cependant, dans les situations réelles, x peut manquer d'informations et rester incomplet pour différentes raisons telles que : un manque d'information sur les caractéristiques des bras, une mauvaise modélisation du contexte c'est-à-dire un contexte spécifié de manière incomplète, des restrictions dues à des problématiques de confidentialité et de protection de la vie privée, un profil mal renseigné, des informations manquantes sur l'environnement de l'utilisateur (par exemple, une localisation temporairement indisponible). Dans les cas où $x \neq x^*$, les algorithmes de CMAB doivent faire face à des contraintes de parcimonies dans les données ou d'incomplétude sur les caractéristiques disponibles puisque les caractéristiques de x^* manquantes dans x ne peuvent pas être prises en compte. En effet, avec un vecteur de contexte insuffisamment décrit, les clusters associés à x^* ne seront pas pris en compte par *LinUCB*, qui peut finalement être incapable de tirer le bras optimal pour différentes situations. Les utilisateurs affectés par cette parcimonie vectorielle pourraient donc se retrouver insatisfaits de la sélection des bras qui leur est proposée par *LinUCB*. Cela entraîne une diminution de la précision globale mais également de la précision individuelle ciblant ces utilisateurs. Ces problématiques nous ont conduits à construire une nouvelle approche visant, à la fois, à garder une bonne précision globale et à atténuer la diminution de la précision individuelle. La sous-section suivante présente notre méthode qui utilise un mécanisme de diversification afin de contrer le manque d'information contextuelle et favoriser la sérendipité.

4.2 Sliding Window LinUCB : SW-LinUCB

Notre Fenêtre Glissante : Notre nouvelle approche combine *LinUCB* et l'utilisation d'une fenêtre glissante permettant de pénaliser la sélection des bras optimaux (tirés plus fréquemment), afin de favoriser l'exploration des bras moins optimaux que nous appellerons ici l'*ensemble des bras sous-optimaux*. Les méthodes utilisant des fenêtres glissantes appliquent généralement un coefficient dit de *discount* pondérant les récompenses obtenues par leurs bras afin de favoriser les observations les plus récentes. Ainsi, il est possible de définir un coefficient de *discount* qui pondère les récompenses cumulées obtenues pour chaque bras tel que $\sum_{t=1}^{T} \gamma_t r_{t,a}$ [23]. Avec $\gamma_t = 1 - \frac{Occ_w(a,t)}{w}$ où w correspond à la taille de la fenêtre glissante et $Occ_w(a,t)$ représente le nombre de fois qu'un bras a a été sélectionné durant les t dernières itérations. $Occ_w(a,t) = \#_1(E_{t,a})$ où $E_{t,a} = \{0..(2^{(w+1)} - 1)\}$ représente les w dernières sélections d'un bras a donné e.g., pour une taille de fenêtre $w = 6$, $E_{t,a} = 101001$ signifie que a a été sélectionné aux itérations $t-6$, $t-4$ et $t-1$. Néanmoins, même si il pourrait être intéressant de combiner une telle méthode à *LinUCB*, celle-ci reste un processus mettant en œuvre une mémoire à court-terme qui, dans notre cas, ne permettra pas de diversifier suffisamment. Dans notre cas, nous devons conserver un processus d'élimination des mauvaises solutions sur le long terme tout en diversifiant suffisamment parmi l'ensemble des bras sous-optimaux. Ainsi, nous proposons une nouvelle méthode de calcul du gain basée sur le $p_{t,a}$ originel tel que

$$p_{t,a}^w = \gamma_t \widehat{\theta_a}^\top x_t + \alpha \sqrt{x_t^\top M_a^{-1} x_t} \quad (1)$$

où $M_a = (D_a^\top D_a + I_d)$. Ce calcul permet à la fois de garder la confiance (élimination à long terme) grâce au *bonus*, et de diversifier suffisamment parmi l'ensemble des bras sous-optimaux en pénalisant temporairement l'espérance calculée des récompenses pour les bras sélectionnés trop fréquemment.

L'algorithme SW-LinUCB : l'objectif de *SW-LinUCB* est de déterminer la politique π qui maximise les récompenses cumulées à l'horizon T tandis qu'une fenêtre glissante force la diversification parmi l'ensemble des bras sous-optimaux. Notre hypothèse est la suivante : en fonction du niveau de parcimonie du vecteur de contexte ($x \neq x^*$), diversifier parmi l'ensemble des bras sous-optimaux atténuera la perte de précision individuelle pour les utilisateurs pour lesquels la méthode d'origine obtient une très faible précision. Notre méthode est décrite dans l'algorithme 1.

5 Expérimentations et Résultats

Jeux de données : L'évaluation de notre proposition se base sur quatre jeux de données. Tout d'abord un jeu de

Algorithme 1 *Sliding Window LinUCB (SW-LinUCB)*

Require: L'ensemble des k bras $a \in A$ disponibles, $\alpha \in \mathbb{R}^+$, l'horizon T, et l'ensemble des n contextes fixes disponibles X

1: $w \leftarrow k$
2: **for** $t = 1$ to T **do**
3: Considérer $x_t \in X$: un utilisateur et son contexte
4: **for all** $a \in A$ **do**
5: **if** a n'a pas encore été sélectionné **then**
6: $Occ_w(a,t) \leftarrow 0$; $M_a \leftarrow I_d$; $b_a \leftarrow 0_{d\times 1}$
7: **end if**
8: $\widehat{\theta}_a \leftarrow M_a^{-1} b_a$
9: **if** $t > w$ **then**
10: Calculer $Occ_w(a,t) = \#_1(E_{t,a})$
11: **end if**
12: $p^w_{t,a} \leftarrow \left(1 - \frac{Occ_w(a,t)}{w}\right) \widehat{\theta}_a^{\top} x_t + \alpha\sqrt{x_t^{\top} M_a^{-1} x_t}$
13: **end for**
14: Sélectionner le bras $a_t = \arg\max_{a_t \in A}(p_{t,a})$ et observer la récompense r_t retournée par l'utilisateur
15: $M_{a_t} \leftarrow M_{a_t} + x_t x_t^{\top}$; $b_{a_t} \leftarrow b_{a_t} + r_t x_t$
16: $\forall a \neq a_t$, mettre à jour toutes les sous-séquences $E_{t,a}$ en ajoutant un bit 0
17: Mettre à jour la sous-séquence E_{t,a_t} en ajoutant un bit 1
18: **if** $t > w$ **then**
19: Réaliser un décalage logique vers la gauche (Left Shift) de $E_{t,a}$ et E_{t,a_t}
20: **end if**
21: **end for**

données a été artificiellement généré afin d'obtenir un x^* garantissant une équiprobabilité entre chacun des bras. Il servira de jeu de contrôle dans nos expérimentations. Enfin, nous avons utilisé trois autres jeux de données d'applications réelles : Recommendation System for Angers Smart City (RS-ASM)[1], Covertype et Poker Hand[2]. Chacun des jeux de données considéré est constitué d'un nombre d'instances, s'appuie sur un contexte d'une dimension donnée et propose un nombre défini de bras (voir Tableau 1).

Jeu	Instances	Dim	Bras	Source
Contrôle	1000	4	4	Generated
RS-ASM	2152	56	18	Kaggle
Covertype	581 012	95	7	UCI
Poker Hand	1 025 010	11	9	UCI

Tableau 1 – Jeux de données

Mesure de Précision Globale : La précision globale est un critère de performance basé sur le total des récompenses positives cumulées à l'horizon T. De ce fait, pour obtenir la précision globale, nous calculons le gain c'est à dire le nombre total de récompenses positives $g(T)$ puis nous calculons enfin la précision (Accuracy : Acc, voir Tableau 2) tel que : $Acc(T) = \frac{g(T)}{T}$ où $g(T) = \sum_{t=1}^{T} r_t$ et $r_t = \{0, 1\}$.

Mesure de Diversité : La diversité de sélection parmi un ensemble fini et fixe de k bras peut être définie comme un critère de dispersion découlant du coefficient de variation (C_v) des bras sélectionnés. Il est par conséquent possible de calculer la diversité (Div, voir Tableau 2) comme suit : $Div(N) = 1 - \frac{c_v(N)}{\sqrt{k}}$ où $N = \{n_{a_1}, ..., n_{a_k}\}$, et n_{a_i} correspond au nombre de fois qu'un bras $a_i \in A$ a été sélectionné. Ainsi, la dispersion de la sélection tend à son maximum quand $c_v(N) \to 0$ alors qu'elle tend vers son minimum quand $c_v(N) \to \sqrt{k}$ [25].

Mesure de Précision Individuelle : La précision individuelle par utilisateur $Acc_u(T)$ peut être définie comme étant $\forall u \in \mathcal{U}, Acc_u(T) = \frac{\sum_{t=1}^{T} r_{t,u}}{T_u}$ où T_u représente le nombre de fois qu'un utilisateur avec son contexte a été sélectionné à l'horizon T, et $r_{t,u}$ correspond à la récompense retournée par u à l'itération t. Les mesures de précision individuelle de chaque utilisateur peuvent être représentées par une fonction de distribution cumulative (CDF). La CDF nous permet ainsi d'observer la distribution de la précision individuelle sur l'ensemble des utilisateurs $\mathcal{U}$ avec leur contexte à l'horizon T (voir Figure 1).

Comparaison des Algorithmes[3] : Nous comparons notre algorithme *SW-LinUCB* avec les méthodes suivantes : *UCB* standard (MAB) [8], et *LinUCB* classique (CMAB) [1]. Notons que *LinUCB* et *SW-LinUCB* auront la même valeur du paramètre α calculé avec $\delta = 0.1$.

Protocole Expérimental : Pour chaque algorithme et pour chaque jeu de données, nous simulons 2 cas différents : 1) Avec le vecteur de contexte complet (vc), 2) Avec une partie tronquée à 25% du vecteur de contexte d'origine (vt). Ici, le terme tronqué représente la proportion (en pourcentage) des caractéristiques, sélectionnées aléatoirement, que nous décidons de perdre au début de l'expérience. Ainsi, pour chacun des différents cas et pour chaque algorithme, nous simulons 20 expériences de $10,000,000$ d'itérations pour Poker Hand et Covertype, et $100,000$ concernant RS-ASM et le jeu de données de contrôle. Comme le nombre d'instances de chaque jeu de données est plus ou moins important, nous devons mettre à l'échelle l'horizon T pour chacun d'entre aux afin d'obtenir une mesure suffisante de la précision individuelle.

De plus, pour simuler un flux de données d'utilisateurs se présentant pour recevoir une recommandation (voir ligne 3 de l'algorithme 1), nous sélectionnons séquentiellement et aléatoirement les contextes disponibles dans l'ensemble du jeu de données. Ensuite, nous déterminons les moyennes et écart-types de précision globale et de diversité de sélection des bras sur l'ensemble des 20 simulations. De plus, nous calculons la précision individuelle et déduisons sa CDF dont les données et la représentation sont représentées Figure 1 et Tableau 2. Enfin, nous réalisons un test

[1] `https://www.kaggle.com/`
[2] `http://archive.ics.uci.edu/ml/`
[3] Voir notre étude préliminaire `https://git.io/vxCcv`

		Mesures globales		Distribution de la Précision Individuelle				
		Acc	Div	10%	Q_1	Med	Q_3	90%
UCB	Contrôle	$0.25_{\pm\varepsilon}$	${10^{-3}}_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.25_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$
	RS-ASM	$0.52_{\pm 0.08}$	${10^{-3}}_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.80_{\pm 0.40}$	$1.00_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$
	Poker Hand	$0.47_{\pm 0.04}$	${10^{-3}}_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.60_{\pm 0.49}$	$1.00_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$
	Covertype	$0.41_{\pm 0.05}$	${10^{-3}}_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.90_{\pm 0.30}$	$0.90_{\pm 0.30}$
LinUCB (*vc*)	Contrôle	$1.0_{\pm\varepsilon}$	$0.997_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$
	RS-ASM	$0.78_{\pm\varepsilon}$	$0.86_{\pm\varepsilon}$	$0.04_{\pm 0.02}$	$0.77_{\pm 0.02}$	$0.95_{\pm\varepsilon}$	$0.99_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$
	Poker Hand	$0.53_{\pm\varepsilon}$	$0.06_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.90_{\pm 0.01}$	$1.00_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$
	Covertype	$0.72_{\pm\varepsilon}$	$0.44_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$
SW-LinUCB (*vc*)	Contrôle	$0.991_{\pm\varepsilon}$	$0.997_{\pm\varepsilon}$	$0.98_{\pm\varepsilon}$	$0.99_{\pm\varepsilon}$	$0.99_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$
	RS-ASM	$0.76_{\pm\varepsilon}$	$0.88_{\pm\varepsilon}$	$0.06_{\pm 0.02}$	$0.68_{\pm 0.01}$	$0.92_{\pm 0.01}$	$0.98_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$
	Poker Hand	$0.48_{\pm\varepsilon}$	$0.34_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.22_{\pm 0.02}$	$0.50_{\pm\varepsilon}$	$0.72_{\pm 0.02}$	$0.87_{\pm 0.02}$
	Covertype	$0.69_{\pm\varepsilon}$	$0.47_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.40_{\pm\varepsilon}$	$0.89_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$
LinUCB (*vt*)	Contrôle	$0.749_{\pm\varepsilon}$	$0.88_{\pm 0.07}$	$0.35_{\pm 0.09}$	$0.52_{\pm 0.04}$	$0.88_{\pm 0.04}$	$1.00_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$
	RS-ASM	$0.629_{\pm\varepsilon}$	$0.33_{\pm 0.01}$	$0.01_{\pm\varepsilon}$	$0.06_{\pm 0.01}$	$0.92_{\pm 0.01}$	$0.97_{\pm\varepsilon}$	$0.99_{\pm\varepsilon}$
	Poker Hand	$0.50_{\pm\varepsilon}$	$0.01_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.84_{\pm 0.04}$	$1.00_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$
	Covertype	$0.60_{\pm\varepsilon}$	$0.35_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$
SW-LinUCB (*vt*)	Contrôle	$0.746_{\pm\varepsilon}$	$0.96_{\pm 0.02}$	$0.43_{\pm 0.02}$	$0.50_{\pm\varepsilon}$	$0.82_{\pm 0.01}$	$0.99_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$
	RS-ASM	$0.567_{\pm\varepsilon}$	$0.69_{\pm\varepsilon}$	$0.08_{\pm 0.01}$	$0.33_{\pm 0.01}$	$0.61_{\pm 0.01}$	$0.85_{\pm 0.01}$	$0.95_{\pm 0.01}$
	Poker Hand	$0.48_{\pm\varepsilon}$	$0.33_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.26_{\pm 0.01}$	$0.50_{\pm\varepsilon}$	$0.67_{\pm\varepsilon}$	$0.85_{\pm\varepsilon}$
	Covertype	$0.56_{\pm\varepsilon}$	$0.41_{\pm\varepsilon}$	$0.00_{\pm\varepsilon}$	$0.25_{\pm\varepsilon}$	$0.62_{\pm\varepsilon}$	$0.88_{\pm\varepsilon}$	$1.00_{\pm\varepsilon}$

Tableau 2 – Résultats sur plusieurs jeux de données avec vecteur complet (*vc*) et vecteur tronqué (*vt*) ($\varepsilon = 0.0009$)

de *Kruskal-Wallis* pour vérifier l'inégalité des moyennes obtenues sur les critères observés sur l'ensemble des algorithmes, puis nous complétons ces tests par des comparaison deux à deux en réalisant des tests de rang de *Wilcoxon* pour mettre en évidence la significativité statistique de ces inégalités.

5.1 Analyse Globale

Les analyses ci-dessous s'appuient sur les résultats présentés dans la Tableau 2 dont les CDFs sont illustrées Figure 1.

Tests Statistiques : Un test de Kruskal-Wallis pour chaque expérience nous indique qu'il y a une différence significative entre les mesures de précision de chacun des 3 algorithmes ($p < 0.01$). De plus, le test des rangs signés de Wilcoxon met en évidence une différence significative entre chaque paire d'algorithmes ($p < 0.01$).

Diversité : En ce qui concerne les expériences sur le jeu de données de contrôle nous observons comme attendu que lorsque nous fournissons un vecteur optimal x^* les deux algorithmes de CMAB diversifient à 100%. Néanmoins, pour chaque jeu de données, lorsque nous perdons 25% de l'information du vecteur d'origine, alors la diversité décroit pour les deux algorithmes de CMAB. En effet, ils ne réussissent pas à trouver la bonne politique en regard de la règle de correspondance cachée entre récompenses et dimensions du contexte puisqu'une partie pertinente de ce contexte a été tronquée. En revanche, même si *LinUCB* ne diversifie pas autant que dans le cas où nous lui fournissons un vecteur plus complet, *SW-LinUCB* quant à lui offre dans ces mêmes conditions une meilleure diversification que l'algorithme original. Enfin, comme prévu, *UCB* agit comme un algorithme glouton à bornes supérieures de confiance : il trouve le bras optimal et continue de le tirer tout au long des itérations ce qui résulte en une valeur de diversité proche de 0.

Précision Globale VS Individuelle : Comme attendu, *LinUCB* obtient la meilleure performance globale dans tous les cas et pour tout jeu de données. Sans surprise, la précision globale diminue lorsque nous tronquons le vecteur de contexte, mais il est important de noter que même avec le niveau d'éparsité choisi dans notre expérience, les algorithmes de CMABs restent encore meilleurs que l'algorithme de MAB représenté par *UCB*. Cependant, nous observons dans tous les cas (sauf quand $x = x^*$), que *LinUCB* crée un écart de précision individuelle très important entre les utilisateurs. D'autre part, sur l'ensemble des jeux de données et dans tous les cas (sauf quand $x = x^*$), *SW-LinUCB* perd en précision globale par rapport à *LinUCB* mais en revanche trouve, grâce à son mécanisme de diversification, un meilleur compromis en ce qui concerne la distribution de la précision individuelle. Enfin, pour tous les jeux de données, nous observons que plus l'incomplétude du vecteur de contexte est importante, plus un fossé se crée entre les différentes précisions individuelles d'où résultent distinctement une classe de précisions que l'on peut catégoriser de hautes et une classe de précisions dites basses. De la même manière, on remarque que plus x tend vers x^*, plus la distribution de la précision individuelle est uniformément répartie parmi les utilisateurs.

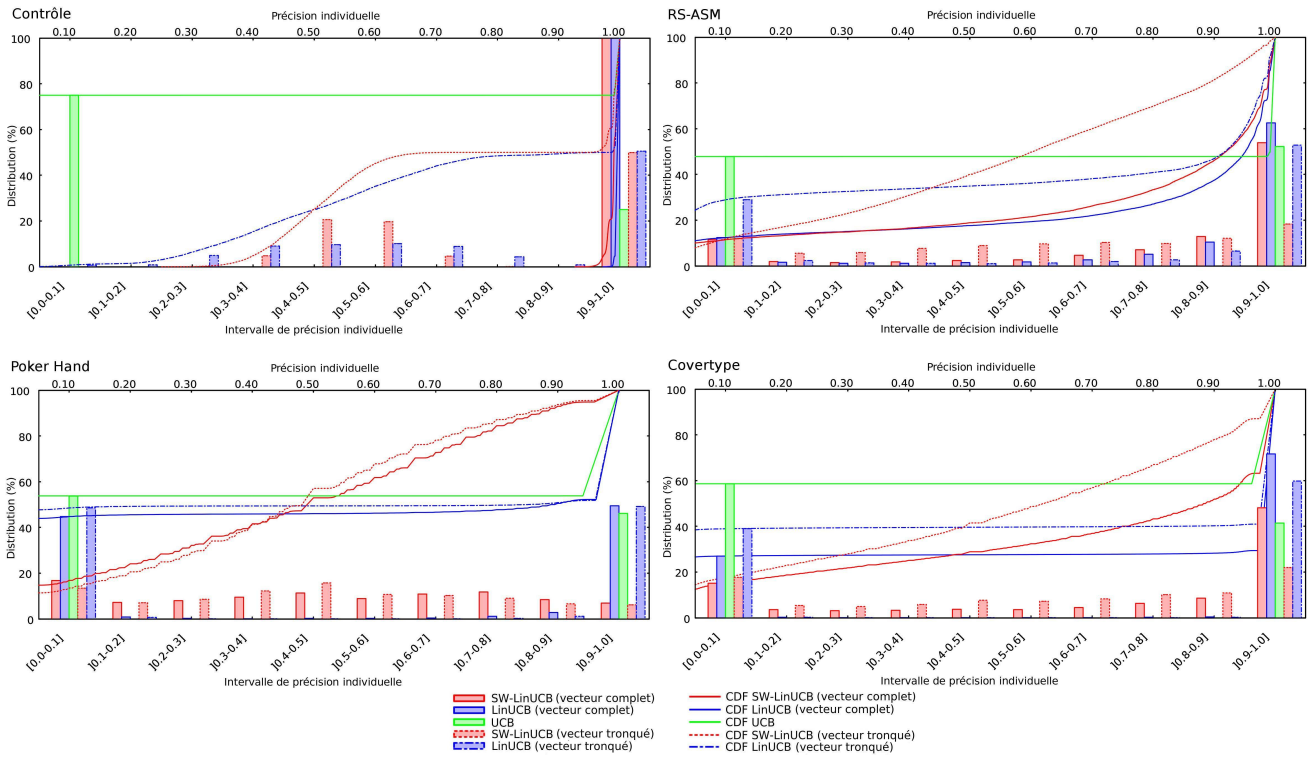

FIGURE 1 – Distribution de la précision individuelle pour chaque algorithme

5.2 Analyse Spécifique sur Covertype

Diversité : On observe que *SW-LinUCB* diversifie plus ($vc : Div = 0.47$, $vt : Div = 0.41$) que *LinUCB* ($vc : Div = 0.44$, $vt : Div = 0.35$) alors que l'algorithme *UCB* continue de tirer le même bras tout au long des itérations ($Div = 10^{-3}$). De plus, nous remarquons que lorsque x tend vers x^*, les caractéristiques fournies en tant que dimension du vecteur de contexte permettent à *LinUCB* et *SW-LinUCB* de diversifier d'avantage.

Précision Globale VS Individuelle : On observe que *LinUCB* conserve une meilleure précision globale ($vc : Acc = 0.72$, $vt : Acc = 0.60$) que *SW-LinUCB* ($vc : Acc = 0.69$, $vt : Acc = 0.56$). De plus, il est important d'observer que le niveau d'incomplétude du vecteur de contexte n'est pas encore assez important pour permettre à notre algorithme de MAB *UCB* d'être plus précis (0.41). En outre, on observe Figure 1 et Tableau 2, que *SW-LinUCB* reste le meilleur en termes de distribution de la précision individuelle ($vc : Q_1 = 0.40$, $vt : Q_1 = 0.25$) que *LinUCB* ($vc : Q_1 = 0.00$, $vt : Q_1 = 0.00$). Ces derniers résultats montrent que notre mécanisme de diversification permet d'augmenter la précision individuelle de la classe dite basse avec la méthode d'origine. Enfin, la comparaison entre les résultats *vc* et *vt* montre que, pour les deux méthodes de CMAB, les précisions globales et individuelles sont toutes deux proportionnelles au niveau d'information et de complétude du vecteur contexte.

6 Conclusion et Perspectives

Dans cet article, nous proposons une nouvelle mesure pour les algorithmes de décision séquentielle visant à évaluer la distribution de la précision individuelle. Nous soutenons que dans certains cas pratiques, la mesure de précision globale n'est pas suffisante pour évaluer les algorithmes de CMAB et que la mesure de précision individuelle doit également être prise en compte. De plus, nous proposons une nouvelle approche adaptée de l'algorithme original *LinUCB* visant à la fois à améliorer la précision individuelle et à maintenir une bonne précision globale. Nous montrons qu'en privilégiant la diversité, notre algorithme *SW-LinUCB* offre un compromis entre précision globale et individuelle que nous pensons mieux adapté à un certain nombre d'applications du monde réel comme les systèmes de recommandations ou les essais cliniques.

Ainsi en perspectives, il semble pertinent de considérer les deux opportunités suivantes : 1) Mettre en place des techniques permettant la construction d'un vecteur plus précis notamment par l'observation approfondie du contexte de l'application concrète qui en découle afin de déterminer les dimensions manquantes et pertinentes ; 2) Concevoir un algorithme pour résoudre le problème multi-objectifs de la maximisation des trois critères de précision globale, de précision individuelle et de diversité. Nous pensons qu'une approche portfolio c'est à dire tirant parti des avantages de plusieurs algorithmes (notamment *LinUCB* et *SW-LinUCB*) pourrait être envisagée.

Références

[1] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 661–670.

[2] S. S. Villar, J. Bowden, and J. Wason, "Multi-armed bandit models for the optimal design of clinical trials : benefits and challenges," *Statistical science : a review journal of the Institute of Mathematical Statistics*, vol. 30, no. 2, p. 199, 2015.

[3] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *International Conference on Machine Learning*, 2013, pp. 127–135.

[4] D. Bouneffouf and R. Feraud, "Multi-armed bandit problem with known trend," *Neurocomputing*, vol. 205, pp. 16–21, 2016.

[5] S. M. McNee, J. Riedl, and J. A. Konstan, "Being accurate is not enough : how accuracy metrics have hurt recommender systems," in *CHI'06 extended abstracts on Human factors in computing systems*. ACM, 2006, pp. 1097–1101.

[6] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, pp. 527–535, 1952.

[7] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[8] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 397–422, 2002.

[9] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Conference on Learning Theory*, 2012, pp. 39–1.

[10] J. Langford and T. Zhang, "The epoch-greedy algorithm for multi-armed bandits with side information," in *Advances in neural information processing systems*, 2008, pp. 817–824.

[11] L. Zhou and E. Brunskill, "Latent contextual bandits and their application to personalized recommendations for new users," *In International Joint Conferences on Artificial Intelligence (IJCAI)*, 2016.

[12] D. Bouneffouf, I. Rish, G. A. Cecchi, and R. Feraud, "Context attentive bandits : Contextual bandit with restricted context," *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2017.

[13] S. Craw, B. Horsburgh, and S. Massie, "Music recommenders : user evaluation without real users ?" *In International Joint Conferences on Artificial Intelligence (IJCAI)*, 2015.

[14] A. Lacerda, "Contextual bandits for multi-objective recommender systems," in *Intelligent Systems (BRACIS), 2015 Brazilian Conference on*. IEEE, 2015, pp. 68–73.

[15] M. Ge, C. Delgado-Battenfeld, and D. Jannach, "Beyond accuracy : evaluating recommender systems by coverage and serendipity," in *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 2010, pp. 257–260.

[16] X. Wang, Y. Guo, and C. Xu, "Recommendation algorithms for optimizing hit rate, user satisfaction and website revenue." in *In International Joint Conferences on Artificial Intelligence (IJCAI)*, 2015, pp. 1820–1826.

[17] A. Ashkan, B. Kveton, S. Berkovsky, and Z. Wen, "Optimal greedy diversity for recommendation." *In International Joint Conferences on Artificial Intelligence (IJCAI)*, pp. 1742–1748, 2015.

[18] L. Hu, L. Cao, S. Wang, G. Xu, J. Cao, and Z. Gu, "Diversifying personalized recommendation with user-session context," *In International Joint Conferences on Artificial Intelligence (IJCAI)*, 2017.

[19] R. Allesiardo, R. Féraud, and D. Bouneffouf, "A neural networks committee for the contextual bandit problem," in *International Conference on Neural Information Processing*. Springer, 2014, pp. 374–381.

[20] T. J. Walsh, I. Szita, C. Diuk, and M. L. Littman, "Exploring compact reinforcement-learning representations with linear regression," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009, pp. 591–598.

[21] A. Garivier and E. Moulines, "On upper-confidence bound policies for switching bandit problems," in *International Conference on Algorithmic Learning Theory*. Springer, 2011, pp. 174–188.

[22] P. Whittle, "Restless bandits : Activity allocation in a changing world," *Journal of applied probability*, vol. 25, no. A, pp. 287–298, 1988.

[23] A. Goëffon, F. Lardeux, and F. Saubion, "Simulating non-stationary operators in search algorithms," *Applied Soft Computing*, vol. 38, pp. 257–268, 2016.

[24] S. Li, A. Karatzoglou, and C. Gentile, "Collaborative filtering bandits," in *The 39th International ACM SIGIR Conference on Information Retrieval (SIGIR)*, 2016.

[25] J. Katsnelson and S. Kotz, "On the upper limits of some measures of variability," *Theoretical and Applied Climatology*, vol. 8, no. 1, pp. 103–107, 1957.