Modèle de distraction pour la sélection séquentielle de contenu

Claire Vernade [1], Paul Lagrée [2] & Olivier Cappé

[1] *Université Paris Saclay, Télécom Paristech, 46 rue Barrault 75013 PARIS.*
claire.vernade@telecom-paristech.fr
[2] *paul.lagree@telecom-parsitech.fr*

Résumé. Dans le contexte du marketing sur Internet, il est fréquent que les publicités présentées aux utilisateurs soient hiérarchisées : les mieux placées retiennent l'attention de l'utilisateur et obtiennent plus de clics, indépendamment de leur contenu propre. Pour construire séquentiellement une campagne qui recueille de nombreux clics sans information *a priori* sur la qualité des articles, il faut donc être capable d'apprendre quelle est la meilleure liste ordonnée de L parmi K produits disponibles dans le catalogue. À chaque fois qu'une liste est proposée à l'internaute, celui-ci clique sur certains produits et c'est l'unique information qu'il envoie au système. Dans le cadre de l'apprentissage séquentiel, ce dernier doit alors mettre à jour ses estimateurs afin de proposer une liste potentiellement meilleure au futur visiteur. L'inconvénient des méthodes existantes pour résoudre ce problème réside dans les modèles : ceux-ci négligent la relative inattention de l'utilisateur, ce qui induit une sous-estimation des probabilités de clics pour les produits présentés et de possibles failles dans l'exploration. Nous proposons donc une manière d'inclure cet aspect dans un modèle de *Bandits Manchots* original. Après avoir précautionneusement étudié l'impact de la distraction de l'utilisateur sur les performances asymptotiques des algorithmes, nous exploitons le principe d'*optimisme face à l'incertitude* pour proposer une série d'algorithmes efficaces que nous évaluons expérimentalement.

Mots-clés. Bandits manchots, marketing en ligne, bornes inférieures sur le regret

Abstract. In the context of online marketing, ads are not displayed randomly on Web pages but rather in a carefully designed way because it is assumed that the user should pay more attention to well located ads. Indeed, those one get more clicks than the others, independently of their content. Thus, the system that builds such advertising campaigns sequentially must be able to learn which is the best ordered list of L items among the K available in the catalogue. Each time a list is displayed, the user may click on some items or not, and it shall be all the information she sends to the learner. In a sequential fashion, the latter must update its current estimators in order to be able to propose a potentially better list to a future user. Existing models for such problems do not take into account the user's vanishing attention, inducing a vast enough underestimation of click through rates and possibly exploration flaws. The present work suggests a way of including the user's vanishing attention in an original Multi-Armed Bandit Model. We first analyze

the impact of the new model on the asymptotic performance of the algorithms and then use the *optimism in the face of uncertainty* principle to propose some algorithms that we evaluate empirically.

Keywords. Multi-armed bandits, online advertising, lower bounds on the regret.

1 Introduction : du bandit manchot multiple au modèle de distraction

Les modèles de bandits manchots ont connu un regain d'intérêt récent en raison de leur possible application à la construction de campagnes marketing en ligne [2]. Pour construire un affichage de page Web, on a besoin de tirer plusieurs annonces, disons L, à la fois parmi K; on parle alors de bandit manchot multiple [5]. Dans ce cas précis d'application, on peut considérer que l'annonceur est capable de savoir si l'utilisateur qui vient de passer sur sa page a cliqué ou non sur chacune des publicités choisies. Ce type de retour, dit "semi-bandit", a un aspect quelque peu irréaliste : il est peu probable que l'utilisateur ait minutieusement examiné chaque annonce avant de décider de cliquer ou non pour chacune d'entre elles. Il est au contraire plus vraisemblable qu'il n'ait vu que les emplacements les mieux exposés et qu'il n'ait pas prêté attention aux encarts publicitaires en pied de page ou sur les bords. Dans ce cas, il semble intéressant d'inclure la distraction de l'utilisateur dans le modèle d'apprentissage actif afin de ne pas sous-estimer automatiquement les récompenses associées au publicités les moins bien situées et d'assurer ainsi une meilleure exploration.

On suppose donc qu'un catalogue de K actions est disponible, chacune d'entre elles étant caractérisée par sa probabilité de clic : $X_k \sim \mathcal{B}(\theta_k)$. Afin d'alléger les notations par la suite, nous supposerons que ces actions sont distinctes et indexées de telle sorte que $\theta_1 > \theta_2 > ... > \theta_K$. La divergence de Kullback-Leibler entre deux distributions de Bernoulli $\mathcal{B}(p)$ et $\mathcal{B}(q)$ sera notée $d(p,q) := p\log(p/q) + (1-p)\log((1-p)/1-q))$. Le but de l'annonceur est donc de trouver la liste optimale de L éléments parmi K, $(\theta_1, ..., \theta_L)$, en choisissant séquentiellement des listes $A_t := (A_t(1), ...A_t(L))$ de L actions dans l'ensemble $\mathcal{A}$ contenant $K!/(K-L)!$ éléments.

Afin de modéliser la décroissance de l'attention de l'utilisateur, nous supposons qu'à chaque position l de la liste est associé un paramètre $\kappa_l \in]0,1]$ qui contrôle la probabilité que celle-ci soit effectivement examinée par l'utilisateur. Il existe donc une variable aléatoire $Y_l \sim \mathcal{B}(\kappa_l)$ censurée contrôlant cet événement. Ceci modifie donc la probabilité de clic de l'item k lorsqu'il est placé dans cette position : l'annonceur observe $Y_l X_k \sim \mathcal{B}(\kappa_l \theta_k)$.

Lorsqu'elle est tirée, une liste $a \in \mathcal{A}$ rapporte à l'annonceur la récompense moyenne $\mu(a)$ correspondant à l'espérance de la somme des clics obtenus, soit d'après ce qui précède $\mu(a) = \sum_{l=1}^{L} \kappa_l \theta_{a(l)}$. Le regret moyen au bout d'un horizon T est donc la somme des différences entre la récompense à chaque instant t, $\mu(A_t)$, et la meilleure récompense

possible $\mu(a^*)$:

$$R(T) = \sum_{t=1}^{T} \mu(a^*) - \mu(A_t) := \sum_{t=1}^{T} \Delta_{A_t}.$$

En notant $\mathbb{E}[N_a(T)]$ l'espérance du nombre de tirages de chaque liste $a \in \mathcal{A}$, on obtient une réécriture du regret qui s'avérera plus pratique par la suite :

$$R(T) = \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[N_a(T)].$$

Commentaires. L'idée de modifier la structure simple de la réponse semi-bandit a été introduite récemment par [9, 10] sous le nom de *modèle en cascade*, en anglais *Cascade Model*. Le modèle considéré par ces travaux est assez différent du nôtre puisqu'il suppose que l'utilisateur scanne la liste de haut en bas et s'arrête dès qu'il rencontre un item intéressant sur lequel il clique. La récompense est donc binaire : elle vaut 1 si un item de la liste est cliqué, 0 sinon. Il est assez aisé d'en déduire que pour apprendre plus rapidement, l'agent a intérêt à placer les items les plus intéressants à la fin de la liste, ce qui paraît inadapté à la recommandation de contenu.

Contributions. La première contribution de cet article réside dans la proposition d'un modèle de distraction censurée détaillé plus haut. Celle-ci introduit une modification du problème de bandits manchots multiple et donc de la borne inférieure sur le regret qui avait été démontrée par [1]. La Section 2 est donc consacrée à l'exposé de la nouvelle borne inférieure adaptée au nouveau type de réponse considéré. Forts de ce résultat, nous proposons dans la Section 3 un ensemble d'algorithmes fondés sur le principe d'optimisme face à l'incertitude dont les performances sont évaluées empiriquement.

2 Bornes inférieures sur le regret

Pour le modèle en cascade, une borne inférieure a été récemment démontrée par [4]. Un tel résultat repose sur des arguments de changement de mesure que l'on retrouve dans un contexte plus général dans [7] puis plus récemment dans [8] sous la forme de la proposition 1 qui soutiendra notre résultat principal.

Proposition 1 *([8]) Soit $B(\theta)$ l'ensemble des changements de mesures qui améliorent la récompense optimale sous θ sans modifier les actions optimales:*

$$B(\theta) = \{\lambda \in \Theta \,|\forall l \leq L, \theta_l = \lambda_l \text{ et } \mu_\theta(a^*) < \mu_\lambda(a^*)\}$$

Alors, pour tout algorithme uniformément efficace, on a

$$\forall \lambda \in B(\theta), \quad \liminf_{T\to\infty} \frac{\sum_{a\in\mathcal{A}} I_a(\theta,\lambda)\mathbb{E}_\theta[N_a(T)]}{\log(T)} \geq 1$$

où $I_a(\theta,\lambda) := \sum_{l=1}^{L} d(\kappa_l\theta_{a(l)}, \kappa_l\lambda_{a(l)})$.

Grâce à cette proposition, nous sommes en mesure de proposer une formulation variationnelle de la borne inférieure recherchée.

Theorem 2 *Le regret moyen de tout algorithme uniformément efficace est borné inférieurement par*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\theta R(T)}{\log T} \geq \ell(\theta)$$

$$\begin{aligned} \textit{où} \quad \ell(\theta) &= \inf_{c\succeq 0} \sum_{a\in\mathcal{A}} \Delta_a(\theta) c_a \,, \\ t.q. \quad & \inf_{\lambda\in B(\theta)} \sum_{a\in\mathcal{A}} I_a(\theta,\lambda) c_a \geq 1 \end{aligned}$$

La preuve est immédiate et consiste simplement à réécrire le regret comme dans (1) puis à imposer la proposition 1 comme contraintes sur les coefficients c_a introduits et correspondant grossièrement aux $\mathbb{E}[N_a(T)]/\log T$.

Il reste donc à proposer une résolution de ce problème d'optimisation. Pour cela, nous suivons [4] qui propose de prouver que le vecteur c recherché est très parcimonieux : ses coefficients positifs ne concernent que des actions du type $v_l^k := (1, 2, ..., l-1, k, l+1, ..., L)$ introduisant un bras sous optimal k en position l dont le regret est noté $\Delta_{k,l}$.

Theorem 3 *Le regret moyen d'un algorithme uniformément efficace est borné inférieurement par*

$$\liminf_{T\to\infty} \frac{\mathbb{E}_\theta R(T)}{\log T} \geq \sum_{k=L+1}^{K} \min_{l\in\{1,...,L\}} \frac{\Delta_{k,l}(\theta)}{d(\kappa_l\theta_k, \kappa_l\theta_L)}$$

La borne inférieure est donc explicite, moyennant la détermination, pour chaque action sous-optimale $k > L$, de la position d'exploration optimale qui assure le meilleur ratio entre regret induit et quantité d'information gagnée.

3 Algorithmes optimistes pour chaque modèle

Le principe d'optimisme face à l'incertitude consiste à construire pour chaque action k un intervalle de confiance pour θ_k et à ensuite choisir les bras en suivant la borne supérieure de cet intervalle de confiance que nous appelons l'indice de l'action. Pour les problèmes à tirages multiples, [3] suggère de sélectionner la liste en suivant l'ordre décroissant des indices des actions. Nous détaillons donc deux manières différentes d'obtenir des intervalles de confiance. L'équivalent bayésien de ces algorithmes, appelé Thompson Sampling, consiste à échantillonner un paramètre pour chaque action selon sa distribution a posteriori. Cela nécessite de faire appel à un algorithme de rejet dans notre cas.

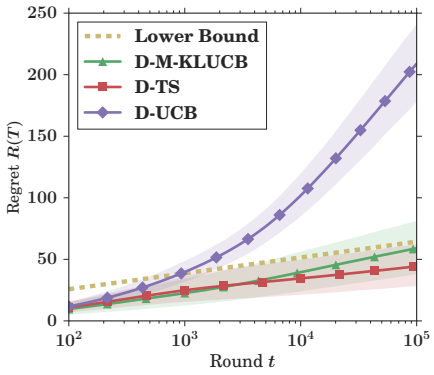

Figure 1: Regret moyen et déciles extrêmes pour les trois algorithmes adaptés au modèle de distraction.

Indice de type UCB. Le premier type d'indice repose sur l'exploitation de l'inégalité d'Azuma-Hoeffding. On introduit les compteurs de tirages : $N_{k,l}(t)$ désigne le nombre de fois où l'action k a été tirée en position l, $N_k(t) = \sum_{l=1}^{L} N_{k,l}(t)$ le nombre total de tirages de k et de manière équivalente $S_k(t)$ le nombre de succès, donc de 1, observés suite à ces tirages. En outre, on peut définir un estimateur du nombre réel d'observations $\tilde{N}_k(t) := \sum_{l=1}^{L} \kappa_l N_{k,l}(t)$.

Proposition 4 *Soit $\epsilon > 0$, et une action k fixée de paramètre θ_k, on définit l'indice suivant*

$$U_k^{UCB}(t, \delta_t) = \frac{S_k}{\tilde{N}_k} + \sqrt{\frac{N_k}{\tilde{N}_k}} \sqrt{\frac{\delta_t}{2\tilde{N}_k}}.$$

Alors, pour tout $\delta > 0$,

$$\mathbb{P}\left(U_k^{UCB}(t, \delta) \leq \theta_k\right) \leq e\delta \log(t) e^{-\delta}$$

En pratique, on choisira $\delta_t = (1 + \epsilon) \log(t)$.

Indice de type KL-UCB. Il est aussi possible de construire un intervalle de confiance tel que [6] à l'aide du résultat de [11]:

$$U_k^L(t, \delta) = \sup_{q \in [\theta_k^{\min}, 1]} \left\{ q \,\middle|\, \sum_{l=1}^{L} N_{k,l} d\left(S_{k,l}/N_{k,l}, \kappa_l q\right) \leq \delta \right\}$$

où $\theta_k^{\min}$ est simplement le minimum de la fonction convexe en argument du sup.

Expériences et validation empirique. Bien que l'analyse statistique en temps fini du regret reste à faire, il est intéressant de vérifier la validité des algorithmes proposés. On réalise donc une simulation avec $K = 5$, $L = 3$, $\kappa = (0.9, 0.6, 0.3)$ et $\theta = (0.45, 0.35, 0.25, 0.15, 0.05)$. Les résultats sont présentés Figure 1. Il semble que, contrairement à UCB, KL-UCB et Thomspon Sampling soient tous deux optimaux.

References

[1] Venkatachalam Anantharam, Pravin Varaiya, and Jean Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays - part I: IID rewards. *Automatic Control, IEEE Transactions on*, 32(11):968–976, 1987.

[2] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

[3] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning*, pages 151–159, 2013.

[4] Richard Combes, Stefan Magureanu, Alexandre Proutiere, and Cyrille Laroche. Learning to rank: Regret lower bounds and efficient algorithms. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 231–244. ACM, 2015.

[5] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation. In *New Frontiers in Dynamic Spectrum, 2010 IEEE Symposium on*, pages 1–9. IEEE, 2010.

[6] Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. *Conference on Learning Theory*, pages 359–376, 2011.

[7] Todd L Graves and Tze Leung Lai. Asymptotically efficient adaptive choice of control laws in controlled markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997.

[8] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 2015.

[9] Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits : Learning to rank in the cascade model. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[10] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 535–543, 2015.

[11] Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bounds and optimal algorithms. In *Proceedings of COLT, Barcelona, Spain*, 2014.