Proceedings of Machine Learning Research vol 65:1–3, 2017

Thompson Sampling for the MNL-Bandit

Shipra Agrawal SA3305@COLUMBIA.EDU
Industrial Engineering and Operations Research, Columbia University

Vashist Avadhanula VAVADHANULA18@GSB.COLUMBIA.EDU
Decision Risk and Operations, Columbia Business School

Vineet Goyal VGOYAL@IEOR.COLUMBIA.EDU
Industrial Engineering and Operations Research, Columbia University

Assaf Zeevi ASSAF@GSB.COLUMBIA.EDU
Decision Risk and Operations, Columbia Business School

We consider a sequential subset selection problem under parameter uncertainty, a decision maker is faced with the problem of determining which subset (of at most cardinality K) of N items to present to users that arrive sequentially and user preferences for said items are unknown. Each user either selects one of the items s/he is offered or selects none. Every item presents some reward which is item-specific. Based on the observations of items users have selected, the decision maker needs to ascertain the composition of the "best bundle," which involves balancing an exploration over bundles to learn the users' preferences, while simultaneously exploiting the bundles that exhibit good reward. This problem arises in many real-world instances, perhaps most notably in display-based online advertising. Here the publisher has to select a set of advertisements to display to users. Due to competing ads, the click rates for an individual ad depends on the overall subset of ads to be displayed; this is referred to as a *substitution* effect. To capture these substitution effects, choice models are often used to specify user preferences in the form of a probability distribution over items in a subset. In this work, we model user preferences over N substitutable items and an outside option (possibility of the offer set having no preferred item), using the widely used multinomial logit (MNL) model with unknown parameters $v_0, v_1, \ldots, v_N$. Under this model the probability that a user chooses item i when set $S \subset \{1, \cdots, N\}$ is offered is given by,

$$p_i(S, \mathbf{v}) = \begin{cases} \dfrac{v_i}{v_0 + \sum_{j \in S} v_j}, & \text{if } i \in S \cup \{0\} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

and therefore the expected revenue when subset S is offered is given by

$$R(S, \mathbf{v}) = \sum_{i \in S} r_i p_i(S, \mathbf{v}) = \sum_{i \in S} \frac{r_i v_i}{1 + \sum_{j \in S} v_j}. \tag{2}$$

We have a specified time horizon T and can offer at most K products at each time step. If S^* is the optimal assortment, our objective is to offer subsets $S_1, \cdots, S_T$ so as to maximize the expected

. Extended abstract. Full version of the paper is available at https://arxiv.org/abs/1706.00977

cumulative rewards over a finite horizon T, or alternatively, minimize the *regret*, defined as

$$Reg(T, \mathbf{v}) = \sum_{t=1}^{T} R(S^*, \mathbf{v}) - E[R(S_t, \mathbf{v})], \tag{3}$$

where $R(S, \mathbf{v})$ is the expected revenue when assortment S is offered as defined in (2). We refer to this problem as the MNL-Bandit problem.

The problem can be formulated as a classical multi-armed (MAB) problem by considering every assortment as an arm, however that would lead to exponentially many arms. Popular extensions of MAB for "large scale" problems include the linear bandit (e.g., Auer (2003)). However, these approaches do not apply directly to our problem, since the revenue corresponding to each offered set is not linear in problem parameters. Other works (see Gopalan et al. (2014)) consider a variant of MAB where one can play a subset of arms in each round and the expected reward is a function of rewards of the arms played. This setting is similar to the MNL-Bandit, though the regret bounds they develop are dependent on the instance parameters as well as the number of possible actions which can be large in our combinatorial problem setting. Rusmevichientong et al. (2010) and Sauré and Zeevi (2013) considered the MNL-Bandit problem in the context of online retail. Both papers develop an "explore first and exploit later" approach, where a fixed set of assortments are explored to learn the MNL parameters to a desired (known) accuracy, and then this information is exploited for the rest of the selling period. Their approaches require prior knowledge of a "separability parameter" that is typically not available in practice. In a more recent paper, Agrawal et al. (2016) show how to exploit specific characteristics of the MNL model to develop a policy based on the principle of "optimism under uncertainty" (UCB-like algorithm, see Auer et al. (2002)) which does not rely on the a priori knowledge of this gap or separation information and achieves a worst-case regret bound of $O(\sqrt{NT \log T})$. It is widely recognized that UCB-type algorithms that optimize the worst case regret typically tend to spend too much time in the exploration phase, resulting in poor performance in practice. To that end, several studies (Oliver and Li (2011), May et al. (2012)) have demonstrated that TS significantly outperforms the state of the art methods in practice.

Motivated by the attractive empirical properties of TS, in this work we focus on a Thompson Sampling (TS) approach to the MNL-Bandit problem. Our main contribution is a TS based approach that is computationally efficient and yet achieves parameter independent (optimal in order) regret bounds. Specifically, we present a computationally efficient TS algorithm for the MNL-Bandit which uses a prior distribution on the parameters of the MNL model such that the posterior update under the MNL-bandit feedback is tractable. A key ingredient in our approach is a two moment approximation of the posterior and the ability to judicially correlate samples, which is done by embedding the two-moment approximation in a normal family. It is shown that our algorithm achieves a worst-case (prior-free) regret bound of $O(\sqrt{NT} \log TK)$ under a mild assumption that $v_0 \geq v_i$ for all i (more on the practicality of this assumption in the full version of the paper). This regret bound is independent of the parameters of the MNL choice model and hence holds uniformly over all problem instances. The regret is comparable to the existing upper bound of $O(\sqrt{NT \log T})$ and the lower bound of $\Omega(\sqrt{NT/K})$ provided by Agrawal et al. (2016) under the same assumptions, yet the numerical results demonstrate that our Thompson Sampling based approach significantly outperforms the UCB-based approach of Agrawal et al. (2016). The methods developed in this paper highlight some of the key challenges involved in adapting the TS approach to the MNL-Bandit, and present a blueprint to address these issues that we hope will be more broadly applicable, and form the basis for further work in the intersection of combinatorial bandits.

References

S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi. 2016. A Near-Optimal Exploration-Exploitation Approach for Assortment Selection. *Proceedings of the 2016 ACM Conference on Economics and Computation (EC)*, 599–600.

P. Auer. 2003. Using Confidence Bounds for Exploitation-exploration Trade-offs. *Journal of Machine Learning Research* 3 , 397–422.

P. Auer, N. Cesa-Bianchi, and P. Fischer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning* 47 , 235–256.

A. Gopalan, S. Mannor, and Y. Mansour. 2014. Thompson Sampling for Complex Online Problems.. In *Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML)*, Vol. 32. 100–108.

B. C. May, N. Korda, A. Lee, and D. S. Leslie. 2012. Optimistic Bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research* 13, 2069–2106.

C. Oliver and L. Li. 2011. An Empirical Evaluation of Thompson Sampling. *In Advances in Neural Information Processing Systems (NIPS)* 24 , 2249-2257.

P. Rusmevichientong, Z. M. Shen, and D.B. Shmoys. 2010. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations Research* 58 (6), 1666–1680.

D. Sauré and A. Zeevi. 2013. Optimal Dynamic Assortment Planning with Demand Learning. *Manufacturing & Service Operations Management* 15 (3), 387–404.