Online Bayesian Recommendation with No Regret

YIDING FENG, Microsoft Research New England, USA
WEI TANG, Washington University in St. Louis, USA
HAIFENG XU, University of Chicago, USA

CCS Concepts: • **Theory of computation** → **Online learning algorithms**; • **Computing methodologies** → **Online learning settings**.

Additional Key Words and Phrases: Bayesian persuasion, regret minimization, online learning, dynamic pricing

ACM Reference Format:
Yiding Feng, Wei Tang, and Haifeng Xu. 2022. Online Bayesian Recommendation with No Regret. In *Proceedings of the 23rd ACM Conference on Economics and Computation (EC '22), July 11–15, 2022, Boulder, CO, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3490486.3538327

Motivated by the *video recommendation* in short-video platforms such as TikTok, Instagram Reels and YouTube Shorts, we introduce and study the *online Bayesian recommendation* problem. Here we describe the problem in the language of video recommendation. Consider a sequential interaction between a video platform and a population of users with the same private preference and belief. At each time, there is a video displayed by the platform to an incoming user. To capture the uncertain characteristics of the video, we study a *Bayesian* model, in which the payoff-relevant characteristics of the video is captured by a (random) *state* of the video. The platform and user each have their own preferences over the video states, which are captured by their utility functions respectively. We assume a natural *information asymmetry* between the platform and users — only the platform can privately observe the realized state of each video, whereas all users only have a prior belief about the video state. Notably, the platform also has its own prior belief over the video state, which is allowed to be different from the users' belief. The platform designs and commits to a recommendation strategy which makes different levels of recommendation (e.g., "standard", "recommended", "highly recommended") based on his private information about the video, i.e., its realized state. After observing the recommendation level, together with her initial belief, the user forms a posterior belief about the video and decides either to watch this video or skip it.

In the idealized situation when the platform knew both the user's preferences and prior beliefs, this sequential Bayesian recommendation problem turns out to be a standard Bayesian persuasion problem and thus can be solved by a linear program [1–3]. This paper, however, addresses the more realistic yet challenging situation in which the platform does not know user's preferences neither user's prior beliefs. Therefore, the platform has to adaptively update his recommendation strategy based on user's past behaviors, so as to maximize its own accumulated utility. The goal of this paper is to design online learning policies with no *Stackelberg regret* for the platform.

Main results. Our first result is an online algorithm (Algorithm 1) that achieves $O(2^m \cdot \log\log T)$ regret, where T is the number of rounds and m is the number of video states. In the optimum

EC '22, July 11–15, 2022, Boulder, CO, USA

ACM ISBN 978-1-4503-9150-4/22/07.
https://doi.org/10.1145/3490486.3538327

policy in hindsight that has the complete knowledge of users' preference and belief, the signaling schemes in all rounds are identical and can be solved separately as the classic Bayesian persuasion problem. By the revelation principle, this optimum signaling scheme in hindsight is a *direct* signaling scheme which has binary recommendation level. In particular, it specifies an order (based on users' preference and belief) over all states and recommends every state above a threshold state in this order. When the platform has no knowledge of users' preference or belief, the order as well as the threshold state specified in the optimum signaling scheme in hindsight remains unknown. Unfortunately, designing an online policy to pin down this order with logarithm regret seems implausible. On the other hand, suppose this order is given, a specific binary search over the threshold state can be accomplished with the double logarithm regret dependence on T, under a careful treatment due to the feedback feature. We formalize this idea and design Algorithm 1. To overcome the uncertainty of the aforementioned order, Algorithm 1 enumerates over all possible orders over all states. As a consequence, Algorithm 1 achieves the double logarithm dependence on the number of rounds T, but exponential dependence on the number of states m.

To also shed lights for problem instances with large m, we introduce Algorithm 2 that achieves $O(\text{poly}(m \cdot \log T))$ regret. Algorithm 2 is designed by phrasing the problem as optimizing a linear program with membership oracle access. In particular, the optimum signaling scheme in hindsight can be formulated as the optimal solution of a linear program as follows. Every feasible solution corresponds to a signaling scheme. The objective is the platform's utility. The constraints are the feasibility constraint and the persuasiveness constraint. Here the feasibility constraint ensures that every feasible solution of the linear program is indeed a signaling scheme, and the persuasiveness constraint ensures that the user prefers to follow the recommendation. When the platform does not know users' preference or belief, the persuasiveness constraint remains unknown. Nonetheless, the platform may check the persuasiveness of a given signaling scheme by deploying it to users. In this sense, the platform obtains a membership oracle for the aforementioned linear program.

Similar to the optimum policy in hindsight, both Algorithm 1 and Algorithm 2 only use direct signaling schemes with binary recommendation level. Such direct signaling schemes are also prevalent in real-world applications such as *For You* in TikTok. However, one may wonder whether restricting to direct signaling schemes with binary recommendation level is still without loss of generality (i.e., whether the revelation principle still holds) in our situations with unknown user preferences. We prove that introducing more recommendation levels cannot improve the regret's order-wise dependence on T. Namely, no online policy can achieve a regret better than $\Omega(\log \log T)$ even for problem instances with binary state. To show this negative result, we first reduce the single-item dynamic pricing problem [4] to a special case of our online Bayesian recommendation problem with binary state, and the signaling schemes are restricted to have binary signal space. Then, we argue that in our problem with $m = 2$, every online policy can be converted into an online policy which only uses direct signaling scheme with the same regret. **The full version of this paper can be found at https://arxiv.org/abs/2202.06135.**

Acknowledgment. Haifeng Xu is supported by a NSF grant CCF-2132506 and a Google Faculty Research Award.

REFERENCES

[1] Ricardo Alonso and Odilon Camara. Bayesian persuasion with heterogeneous priors. *Journal of Economic Theory*, 165: 672–706, 2016.

[2] Shaddin Dughmi and Haifeng Xu. Algorithmic bayesian persuasion. *SIAM Journal on Computing*, 50(3):STOC16–68, 2019.

[3] Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

[4] Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *FOCS*., pages 594–605, 2003.