Adversarial Combinatorial Bandits with General Non-linear Reward Functions

Xi Chen [*1] **Yanjun Han** [*2] **Yining Wang** [*3]

Abstract

In this paper we study the adversarial combinatorial bandit with a known non-linear reward function, extending existing work on adversarial linear combinatorial bandit. The adversarial combinatorial bandit with general non-linear reward is an important open problem in bandit literature, and it is still unclear whether there is a significant gap from the case of linear reward, stochastic bandit, or semi-bandit feedback. We show that, with N arms and subsets of K arms being chosen at each of T time periods, the minimax optimal regret is $\widetilde{\Theta}_d(\sqrt{N^d T})$ if the reward function is a d-degree polynomial with $d < K$, and $\Theta_K(\sqrt{N^K T})$ if the reward function is not a low-degree polynomial. Both bounds are significantly different from the bound $O(\sqrt{\mathrm{poly}(N,K)T})$ for the linear case, which suggests that there is a fundamental gap between the linear and non-linear reward structures. Our result also finds applications to adversarial assortment optimization problem in online recommendation. We show that in the worst-case of adversarial assortment problem, the optimal algorithm must treat each individual $\binom{N}{K}$ assortment as independent.

1. Introduction

In this paper we study the *combinatorial bandit* problem, which is a natural extension to the multi-armed bandit problem (Auer et al., 1995) and has applications to online advertising, online shortest paths and many other practical problems (Cesa-Bianchi & Lugosi, 2012; Chen et al., 2013; 2016a;b; Wang & Chen, 2018). In the adversarial combinatorial bandit setting, there are T time periods and N arms. At the beginning of each time period t, an adaptive adversary chooses a reward vector $v_t = (v_{t1}, \cdots, v_{tN}) \in [0,1]^N$ not revealed to the algorithm. The algorithm chooses a subset $S_t \subseteq [N]$ consisting of exactly $K \le N$ distinct arms (i.e., $|S_t| = K$). The algorithm then receives a *bandit* feedback $r_t \in [0,1]$ satisfying

$$\mathbb{E}[r_t | S_t, v_t] = g\left(\sum_{i \in S_t} v_{ti}\right), \tag{1}$$

where $g : \mathbb{R}^+ \to [0,1]$ is a known link function. The objective is to minimize the *regret* of the algorithm compared against a stationary benchmark, defined as

$$\max_{|S^\star|=K} \mathbb{E}\left[\sum_{t=1}^{T} R(S^\star, v_t) - R(S_t, v_t)\right], \tag{2}$$

where $R(S, v_t) := g(\sum_{i \in S} v_i)$ and $\{S_t\}_{t=1}^T$ are the subsets outputted by a regret minimization algorithm.

As far as we know, all existing works on *adversarial* combinatorial bandit studied only the case when the link function is linear (i.e., $g(x) = cx$) (Cesa-Bianchi & Lugosi, 2012; Bubeck et al., 2012; Audibert et al., 2014). While there have been research on combinatorial bandits with general link functions, such results are established exclusively for the *stochastic* setting, in which the reward vectors $\{v_t\}_{t=1}^T$ do *not* change over time (Rusmevichientong et al., 2010; Agarwal & Aggarwal, 2018; Agrawal et al., 2019), and most of them further assume a *semi-bandit* feedback where noisy observations of all v_{ti} with $i \in S_t$ are available (Combes et al., 2015; Kveton et al., 2015; Chen et al., 2016a; 2018a;b). This motivates the following natural question:

Question 1. *For adversarial combinatorial bandit with a non-linear link function $g(\cdot)$ and bandit feedback, is it possible to achieve $\widetilde{O}(\sqrt{\mathrm{poly}(N,K)T})$ regret?*

Note that in adversarial combinatorial bandit with linear link function, *or* stochastic combinatorial (semi-)bandit with general link functions, the $\widetilde{O}(\sqrt{\mathrm{poly}(N,K)T})$ regret targeted in Question 1 can be attained (Bubeck et al., 2012; Combes et al., 2015; Kveton et al., 2015; Chen et al., 2016a; Agrawal et al., 2019). The question also has important practical motivations beyond theoretical/mathematical reasoning, because many interesting applications of combinatorial bandit involve non-linear link functions, such as online assortment

[*]Equal contribution [1]Stern School of Business, New York University, New York, NY 10012, USA [2]Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA [3]Warrington College of Business, University of Florida, Gainesville, FL 32611, USA. Correspondence to: Yining Wang <yining.wang@warrington.ufl.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, PMLR 139, 2021.

optimization with a multinomial logit (MNL) model, which corresponds to a link function of $g(x) = x/(1+x)$. Please see more discussions Sec. 1.2.

1.1. Our results

Below is an informal statement of our main result, as a summary of Theorem 1 later in the paper.

Corollary 1 (Informal)**.** *Fix an arbitrary, known reward function $g : \mathbb{R}^+ \to [0,1]$. If g is a d-degree polynomial for some $d < K$, then the optimal regret is $\widetilde{\Theta}_{g,d,K}\big(\sqrt{N^d T}\big)$.*

Otherwise, if g is either not a polynomial or a polynomial of degree at least K, the optimal regret is $\Theta_{g,K}\big(\sqrt{N^K T}\big)$.

The results in Corollary 1 easily cover the linear link function case $g(x) = x$, with $d = 1$ and the optimal regret being $\widetilde{\Theta}(\sqrt{NT})$ (Bubeck et al., 2012). On the other hand, Corollary 1 shows that when g is not a polynomial, no algorithm can achieve a regret better than $O(\sqrt{N^K T})$. This shows that when g is a general non-linear reward function, the $\binom{N}{K}$ subsets of the N arms can only be treated as "independent" and it is information-theoretically impossible for any algorithm to exploit correlation between subsets of arms to achieve a significantly smaller regret.

1.2. Dynamic assortment optimization

Dynamic assortment optimization is a key problem in revenue management and online recommendation, which naturally serves as a motivation for the generalized combinatorial bandit problem studied in this paper. In the standard setup of dynamic assortment optimization (Agrawal et al., 2019), there are N substitutable products, each associated with a known profit parameter $p_i \in [0,1]$ and an unknown mean utility parameter $v_i \in [0,1]$. At each time, the seller offers an *assortment* (i.e., the recommended set of products) $S_t \subseteq [N]$ of size K, e.g., there are K display spots of recommended products on an Amazon webpage. Then the customer either purchases one of the products being offered (i.e., $i_t \in S_t$) or leaves without making any purchase (i.e., $i_t = 0$). The choice behavior of the customer is governed by the well-known MultiNomial-Logit (MNL) model from economics (Train, 2009):

$$\mathbb{P}[i_t = i | S_t, v] = \frac{v_i}{v_0 + \sum_{j \in S_t} v_j}, \qquad \forall i \in S_t \cup \{0\}, \quad (3)$$

with the definition that $v_0 := 1$, where v_0 denote the utility of no-purchase. The objective for the seller or retailer is to maximize the expected profit/revenue

$$R(S, v) = \sum_{i \in S} p_i \mathbb{P}[i|S, v] = \frac{\sum_{i \in S} p_i v_i}{v_0 + \sum_{i \in S} v_i}.$$

Note also that, in the adversarial setting, the mean utility vector $v_t = \{v_{ti}\}_{i=1}^N$ will be different for each time period $t = 1, 2, \cdots, T$, and will be selected by an adaptive adversary. The regret is then defined as (2).

Let us first consider a special case, where all the products have the profit one (i.e., $p_i \equiv 1$) and only binary purchase/no-purchase actions are observable. That is, one only observes a binary reward at time t, $r_t = \mathbf{1}\{i_t \in S_t\}$, which indicates whether there is a purchase. Then the dynamic assortment optimization question reduces to the generalized (adversarial) combinatorial bandit problem formulated in Eqs. (1,2) with the link function $g(x) = x/(1+x)$. Since $g(x) = x/(1+x)$ is clearly not a polynomial, Corollary 1 shows that $\Theta(\sqrt{N^K T})$ should be the optimal regret. The following corollary extends this to the general case of dynamic assortment optimization, where different products can have different profit parameters.

Corollary 2. *Consider the dynamic assortment optimization question with known profit parameters $\{p_i\}_{i=1}^N \subseteq [0,1]$ and unknown mean utility parameters $\{v_{ti}\}_{t,i=1}^{T,N} \subseteq [0,1]$ chosen by an adaptive adversary. Then there exists an algorithm with regret upper bounded $O_K(\sqrt{N^K T})$.*

The next corollary, on the other hand, shows that the $O(\sqrt{N^K T})$ regret in not improvable, even with a richer non-binary feedback and if all products have the same profits parameter $p_i \equiv 1$.

Corollary 3. *Suppose $p_i \equiv 1$. There exists an adaptive adversary that chooses $\{v_{ti}\}_{t,i=1}^{T,N} \subseteq [0,1]$, such that for any algorithm, the regret is lower bounded by $\Omega_K(\sqrt{N^K T})$.*

Both corollaries are consequences of Proposition 1 and Lemma 3 later in the paper.

1.3. Proof techniques

As we shall see later in this paper, the upper bounds of $\widetilde{O}_{g,K,d}(\sqrt{N^d T})$ or $O_{g,K}(\sqrt{N^K T})$ in Corollary 1 are relatively easier to establish, via reduction to known adversarial multi-armed or linear bandit algorithms. The key challenge is to establish corresponding $\Omega(\sqrt{N^d T})$ and $\Omega(\sqrt{N^K T})$ *lower bounds* in Corollary 1.

In this section we give a high-level sketch of the key ideas in our lower bound proof. For simplicity we consider only the case when $g(\cdot)$ is not a polynomial function. The key insight is to prove the existence of a distribution μ on $v \in [0,1]^n$, such that for any $S \subseteq [n]$, $|S| = K$, the following holds on the choice distribution $\mathbb{P}(\cdot|S, v)$ with $\mathbb{P}_0 \neq \mathbb{P}_1$:

$$\mathbb{E}_{v \sim \mu}[\mathbb{P}(\cdot|S, v)] \equiv \begin{cases} \mathbb{P}_0, & \text{if } S = S^\star, \\ \mathbb{P}_1, & \text{if } S \neq S^\star. \end{cases} \quad (4)$$

Intuitively, Eq. (4) shows that, no information is gained unless an algorithm *exactly* guesses the optimal subset $S^\star$, even if the subset S_t produced by the algorithm only differ by a single element from $S^\star$. Since there are $\binom{N}{K}$ different

subsets, the question of guessing the optimal subset $S^\star$ exactly correct is similar to locating the best arm of a multi-armed bandit question with $\binom{N}{K} = \Theta_K(N^K)$ arms, which incurs a regret of $\Omega_K(\sqrt{N^K T})$.

To gain deeper insights into the construction of $v \sim \mu$ that satisfies Eq. (4), it is instructive to consider some simpler bandit settings in which a small regret can be achieved and understand why the construction of μ does not apply there.

- The first setting is dynamic assortment optimization under the stationary setting, in which the $\{v_t\}_{t=1}^T$ vectors remain the same for all T periods. The results of (Agrawal et al., 2019) achieve $\widetilde{O}(\sqrt{NT})$ regret in this setting. In this setting, the vector v is deterministic and fixed, and therefore the laws $\mathbb{P}(\cdot|S, v)$ and $\mathbb{P}(\cdot|S', v)$ must be correlated as long as $S \cap S' \neq \emptyset$. This means that Eq. (4) cannot be possibly satisfied, with every subset $S \neq S^\star$ revealing no information about $S^\star$.

- The second setting is the adversarial combinatorial bandit with a linear link function $g(x) = cx$, for which an $\widetilde{O}_K(\sqrt{NT})$ regret is attainable (Bubeck et al., 2012; Combes et al., 2015). When g is linear, the expectation of the mixture distribution $\mathbb{E}_{v\sim\mu}[\mathbb{P}(\cdot|S, v)]$ is

$$\mathbb{E}_{v\sim\mu}[R(S, v)] = g(\langle \mathbf{1}_S, \mathbb{E}_\mu[v]\rangle),$$

  where $\mathbf{1}_S \in \{0, 1\}^n$ is the indicator vector of the subset $S \subseteq [N]$. Clearly, this is impossible to achieve (4) as there is no vector $w \in \mathbb{R}^N$ satisfying that $\langle \mathbf{1}_S, w\rangle$ is constant for all $S \neq S^\star$ and $\langle \mathbf{1}_S, w\rangle \neq \langle \mathbf{1}_{S^\star}, w\rangle$.

- The third setting is a special stochastic combinatorial bandit, where $v \sim \mu$ is random but there exists a total ordering of the stochastic dominance relation among the components $(\mu_1, \cdots, \mu_N)$ of $\mu = \prod_{i=1}^N \mu_i$, and an increasing g. In this setting, it was shown in (Agarwal & Aggarwal, 2018) that a regret of $\widetilde{O}_K(N^{1/3}T^{2/3})$ can be achieved, and the stochastic dominance requirement implies that once an element of $[N]\backslash S^\star$ is replaced by an element of $S^\star$, the expected reward must increase. Therefore, (4) cannot hold in this scenario either.

1.4. Other related works

Combinatorial bandit is a classical question in machine learning and has been extensively studied under the settings of stochastic semi-bandits (Chen et al., 2013; Combes et al., 2015; Kveton et al., 2015; Chen et al., 2016a;b; Wang & Chen, 2018; Merlis & Mannor, 2019; 2020), stochastic bandits (Agarwal & Aggarwal, 2018; Rejwan & Mansour, 2020; Kuroki et al., 2020), and adversarial linear bandits (Cesa-Bianchi & Lugosi, 2012; Bubeck et al., 2012; Audibert et al., 2014; Combes et al., 2015). In the above-mentioned works, either the reward link function $g(\cdot)$ is linear, or the model is stochastic (stationary).

There is another line of research on dynamic assortment optimization with the multinomial logit model, which is a form of combinatorial bandit with general reward functions (Rusmevichientong et al., 2010; Agrawal et al., 2017; Chen et al., 2018a;b; Chen & Wang, 2018; Chen et al., 2019; Agrawal et al., 2019). All of these works are carried out under the stochastic setting, with the exception of (Chen et al., 2019) which studied an ε-contamination model and obtained a regret upper bound $\widetilde{O}(\sqrt{NT} + \varepsilon T)$. Clearly, with the adversarial setting in this paper ($\varepsilon = 1$) the regret bound in (Chen et al., 2019) becomes linear in T and thus meaningless.

1.5. Notations

For a multi-index $\alpha \in \mathbb{N}^d$, let $|\alpha| = \sum_{i=1}^d \alpha_i$, and $D^\alpha f = \partial^{|\alpha|} f / \prod_{i=1}^d \partial x_i^{\alpha_i}$ for a d-variate function f. For $m \in \mathbb{N}$ and interval I, let $C^m(I)$ be the set of m-times continuously differentiable functions on I. For two probability distributions P and Q, let $\mathrm{TV}(P, Q) = \frac{1}{2}\int |dP - dQ|$ and $D_{\mathrm{KL}}(P\|Q) = \int dP \log(dP/dQ)$ be the total variation (TV) distance and the Kullback–Leibler (KL) divergence, respectively. We adopt the standard asymptotic notation: for two non-negative sequences $\{a_n\}$ and $\{b_n\}$, we use the notation $a_n = O_c(b_n)$ to denote that $a_n \leq Cb_n$ for all n and constant $C < \infty$ depending only on c, $a_n = \Omega_c(b_n)$ to denote $b_n = O_c(a_n)$, and $a_n = \Theta_c(b_n)$ to denote both $a_n = O_c(b_n)$ and $a_n = \Omega_c(b_n)$. We also use $\widetilde{O}(\cdot), \widetilde{\Omega}(\cdot), \widetilde{\Theta}(\cdot)$ to denote the respective meanings up to a multiplicative poly-logarithmic factor in (N, T).

2. Problem formulation and results

Suppose there are N arms, T time periods and a known reward function $g : \mathbb{R}_+ \to [0, 1]$. At each time period t, the algorithm outputs a subset $S_t \subseteq [N]$, $|S_t| = K$ and receives a binary bandit feedback $r_t \in \{0, 1\}$. Note that the binary feedback structure can be significantly relaxed for the purpose of upper bounds, as discussed in Sec. 2.1. Let $\mathcal{F}_t = \{S_\tau, r_\tau\}_{\tau\leq t}$ be the filtration of observable statistics at time t, and $\mathcal{V}_t = \{v_\tau\}_{\tau\leq t}$ be the filtration of unobservable reward vectors. Let also $\mathcal{A}$ be an unknown adversary and π be an admissible policy. The reward dynamics are modeled as follows:

$$\begin{aligned} v_t &\sim \mathcal{A}(\mathcal{F}_{t-1}, \mathcal{V}_{t-1}); \\ S_t &\sim \pi(\mathcal{F}_{t-1}); \\ r_t &\sim \mathrm{Bernoulli}(g(\textstyle\sum_{i\in S_t} v_{ti})). \end{aligned}$$

For any $g(\cdot), N, K, T$, the *minimax regret* $\mathfrak{R}(g, N, K, T)$ is defined as

$$\mathfrak{R}(g, N, K, T) := \inf_{\pi} \sup_{\mathcal{A}} \max_{|S^\star|=K} \mathbb{E}\left[\sum_{t=1}^{T} R(S^\star, v_t) - R(S_t, v_t)\right], \quad (5)$$

where $R(S, v_t) = g(\sum_{i\in S} v_{ti})$, and the expectation is taken with respect to both the bandit algorithm π and the adaptive adversary $\mathcal{A}$.

The following theorem is a rigorous statement of the main result of this paper.

Theorem 1. *Fix function $g : \mathbb{R}^+ \to [0, 1]$ that is K-times continuously differentiable on $(0, K)$. If g is a polynomial with degree $d \in [1, K)$, then there exist constants $0 < c_{g,d,K} \le C_{g,d,K} < \infty$ depending only on g, d, K such that, for every $N \ge K$ and $T \ge 1$, it holds that*

$$c_{g,d,K} \le \frac{\mathfrak{R}(g, N, K, T)}{\min\{T, \sqrt{N^d T}\}} \le C_{g,d,K}\sqrt{\log N}.$$

Furthermore, if g is a polynomial with degree at least K or not a polynomial, then there exist constants $0 < c_{g,K} \le C_{g,K} < \infty$ depending only on g, K such that, for every $N \ge K$ and $T \ge 1$, it holds that

$$c_{g,K} \le \frac{\mathfrak{R}(g, N, K, T)}{\min\{T, \sqrt{N^K T}\}} \le C_{g,K}.$$

Remark 1. *Based on Propositions 1 and 2 later, the hidden dependence of the constants $C_{g,d,K}, C_{g,K}$ on (g, d, K) is $O(1)$ and $O(\sqrt{K})$, respectively. However, since our lower bound relies on an existential result (cf. Lemma 2), the hidden dependence of constants $c_{g,d,K}$ and $c_{g,K}$ is unknown. It is an outstanding open question to characterize an explicit dependence on (g, d, K) in the lower bound.*

The results in Theorem 1 cover the linear reward case of $g(x) = cx$ via $d = 1$ and a regret of $\widetilde{\Theta}_K(\min\{T, \sqrt{NT}\})$, which matches the existing results on adversarial linear combinatorial bandits. On the other hand, the results for general non-polynomial reward functions $g(\cdot)$ are quite negative, with a $\Theta_{g,K}(\min\{T, \sqrt{N^K T}\})$ regret showing that all the $\binom{N}{K}$ subsets are essentially independent and the bandit algorithm cannot hope to exploit correlation between overlapping subsets like in the linear case. Finally, the reward function $g(\cdot)$ being a low-degree polynomial interpolates between the linear case and the general case, with a regret of $\widetilde{\Theta}_{g,d,K}(\min\{T, \sqrt{N^d T}\})$ for $d \in (1, K)$ between $\widetilde{\Theta}_K(\min\{T, \sqrt{NT}\})$ and $\Theta_{g,K}(\min\{T, \sqrt{N^K T}\})$.

In the rest of this section we sketch the proofs of Theorem 1 by studying the upper bounds and lower bounds separately. We will also adapt the proof of Theorem 1 to cover the more general dynamic assortment optimization model described in Sec. 1.2.

2.1. Upper bounds

We first prove the $O_K(\min\{T, \sqrt{N^K T}\})$ regret upper bound for general link functions. In fact, we state the following result that is much more general than Theorem 1.

Proposition 1. *Suppose at each time t the adversary $\mathcal{A}$ could choose an arbitrary combinatorial reward model $\{\mathbb{P}_t(\cdot|S)\}_{S\subseteq[N],|S|=K}$, and an arbitrary bandit feedback $\{b_t\}_{t=1}^T$. Let also $r_t(b_t) \in [0, 1]$ denote the reward as a function of b_t. There exists a bandit algorithm π and a universal constant $C < \infty$ such that for any $N \ge K$, $T \ge 1$,*

$$\sup_{\mathcal{A}} \max_{|S^\star|=K} \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}(r_t(b_t)|S^\star, \mathbb{P}_t) - \mathbb{E}(r_t(b_t)|S_t, \mathbb{P}_t)\right] \le C\min\{T, \sqrt{N^K T}\}.$$

Proof. For each subset $S \subseteq [N]$ of size K, let j_S be a constructed arm and $r_t(b_{t,j_S})$ be the bandit reward feedback at time t if the arm j_S is pulled (i.e., subset S is selected). This reduces the problem to an adversarial multi-armed bandit problem with $\binom{N}{K}$ independent arms. Applying the Implicitly Normalized Forecaster (INF) algorithm from (Audibert & Bubeck, 2009), we have the regret upper bound $O(\min\{T, \sqrt{\binom{N}{K}T}\}) = O_K(\min\{T, \sqrt{N^K T}\})$. □

We remark that Proposition 1 is more general and contains the $C_{g,K}\min\{T, \sqrt{N^K T}\}$ upper bound in Theorem 1 as a special case. By considering the feedback model $b_t \in S_t \cup \{0\}$ and $r_t(b_t) = \sum_{i\in S_t} p_i \mathbf{1}\{b_t = i\}$, Proposition 1 also covers the dynamic assortment optimization model described in Sec. 1.2 and Corollary 2.

We next establish the $O(\min\{T, \sqrt{N^d T \log N}\})$ upper bound for polynomial reward functions.

Proposition 2. *Fix a known d-degree polynomial $g(x) = a_d x^d + a_{d-1}x^{d-1} + \cdots + a_1 x + a_0$. Suppose at each time t, conditioned on the selected subset $S_t \subseteq [N]$ of size K, the bandit feedback r_t is supported on an arbitrary bounded set not necessarily $\{0, 1\}$, such that $\mathbb{E}[r_t|S_t, v_t] = g(\sum_{i\in S_t} v_t)$. Then there exists a bandit algorithm π such that, for any adaptive adversary $\mathcal{A}$,*

$$\max_{|S^\star|=K} \mathbb{E}\left[\sum_{t=1}^{T} R(S^\star, v_t) - R(S_t, v_t)\right] \le C_{g,d,K}\min\{T, \sqrt{N^d T \log N}\},$$

where $R(S, v_t) = g(\sum_{i\in S} v_{ti})$, and $C_{g,d,K} = O(\sqrt{K})$.

Proof. For any n-dimensional vector $x \in \mathbb{R}^n$, let $x^{\otimes d} = (x_{i_1} x_{i_2} \cdots x_{i_d})_{i_1,\cdots,i_d=1}^n \in \mathbb{R}^{n^d}$ be the order-d tensorization of x. It is easy to verify that, for any $0 \le k \le d$ and $S_t \subseteq [N]$, $(\sum_{i\in S_t} v_{ti})^k = \langle v_t, \mathbf{1}_{S_t}\rangle^k = \langle v_t^{\otimes k}, \mathbf{1}_{S_t}^{\otimes k}\rangle$, where $\mathbf{1}_{S_t} \in \{0,1\}^n$ is the indicator vector of S_t. Hence,

$$R(S_t, v_t) = \sum_{k=0}^{d} a_k \langle v_t, \mathbf{1}_{S_t}\rangle^k = \sum_{k=0}^{d} a_k \langle v_t^{\otimes k}, \mathbf{1}_{S_t}^{\otimes k}\rangle.$$

Define $\widetilde{v}_t \in \mathbb{R}^{n^d}$ as

$$\widetilde{v}_t := \sum_{k=0}^{d} a_k \underbrace{v_t \otimes \cdots \otimes v_t}_{k \text{ times}} \otimes \underbrace{\frac{\mathbf{1}}{K} \otimes \cdots \otimes \frac{\mathbf{1}}{K}}_{d-k \text{ times}}.$$

As $\langle \mathbf{1}, \mathbf{1}_{S_t}\rangle = K$, it is easy to verify that, for every $S_t \subseteq [N]$, $R(S_t, v_t) = \mathbb{E}[r_t|S_t, v_t] = \langle \widetilde{v}_t, \mathbf{1}_{S_t}^{\otimes d}\rangle$.

With this transformation, the problem reduces to adversarial linear bandit with dimension $D = n^d$ and fixed action space $\mathcal{A} = \{\mathbf{1}_S^{\otimes d}\}_{S\subseteq[N],|S|=K}$ with $|\mathcal{A}| = \binom{N}{K}$. Applying the EXP2 algorithm with John's exploration and the analysis in (Audibert & Bubeck, 2009), the regret is upper bounded by $O(\sqrt{DT \log |\mathcal{A}|}) = O(\sqrt{N^d T K \log N})$, which proves Proposition 2. □

Propositions 1 and 2 complete the proof of minimax upper bounds in Theorem 1.

2.2. Lower bounds

We first prove the following result corresponding to the minimax lower bounds in Theorem 1.

Lemma 1. *Suppose $r_t \sim$ Bernoulli$(g(\sum_{i\in S_t} v_{ti}))$ for some fixed, known function $g : \mathbb{R}_+^n \to [0,1]$ that is K-times continuously differentiable on $(0, K)$. If g is a degree-d polynomial with $d < K$, then there exists a constant $c_{g,d,K} > 0$ such that for all $N \ge K$, $T \ge 1$,*

$$\mathfrak{R}(g, N, K, T) \ge c_{g,d,K} \min\{T, \sqrt{N^d T}\}.$$

Otherwise, there exists a constant $c_{g,K} > 0$ such that for all $N \ge K$, $T \ge 1$,

$$\mathfrak{R}(g, N, K, T) \ge c_{g,K} \min\{T, \sqrt{N^K T}\}.$$

Our proof is based on the following technical lemma:

Lemma 2. *Let $g \in C^m([0,b])$ be a real-valued and m-times continuously differentiable function on $[0,b]$, with $b \ge m$. Then the following two statements are equivalent:*

1. *g is not a polynomial of degree at most $m-1$;*
2. *there exists a random vector $(X_1, \cdots, X_m)$ supported on $[0,1]^m$, which follows an exchangeable joint distribution μ, and a scalar $x_0 \in [0,1]$, such that*

$$\mathbb{E}_\mu[g(X_1 + \cdots + X_{\ell-1} + (b-\ell+1)x_0)] = \mathbb{E}_\mu[g(X_1 + \cdots + X_\ell + (b-\ell)x_0)] \quad (6)$$

for all $\ell = 1, 2, \cdots, m-1$, and

$$\mathbb{E}_\mu[g(X_1 + \cdots + X_{m-1} + (b-m+1)x_0)] < \mathbb{E}_\mu[g(X_1 + \cdots + X_m + (b-m)x_0)]. \quad (7)$$

The proof of Lemma 2 is deferred to the Sec. 3. The construction of the distributions in Lemma 2 is non-constructive and uses duality existential arguments. Its proof also applies several technical tools from real analysis and functional analysis (Rudin, 1991; Donoghue, 1969; Dudley, 2018).

We are now ready to prove Lemma 1.

Proof. We first prove the case when $g(\cdot)$ is not a polynomial of degree at most $K-1$. We use the construction $(X_1, \cdots, X_K)$ and x_0 in Lemma 2 with $(m, b) = (K, K)$, and construct i.i.d. copies $(X_{t,1}, \cdots, X_{t,K})$ for each $t \in [T]$. Consider the following random strategy of the adversary $\mathcal{A}$ when the optimal subset is $S^\star$: at each time $t \in [T]$, nature assigns $(X_{t,1}, \cdots, X_{t,K})$ to the restriction of v_t to $S^\star$ with probability $\delta \in (0,1]$, and assigns $(x_0, \cdots, x_0)$ otherwise, with the parameter δ to be specified later; nature also assigns $v_{ti} = x_0$ for all $i \notin S^\star$. Suppose an algorithm selects subset $S_t \subseteq [N]$, $|S_t| = K$ at time t, such that $|S_t \cap S^\star| = \ell$. Then

$$\begin{aligned}&\mathbb{E}_{v_t}[\mathbb{E}[r_t|S_t, v_t]] \\ &= \delta \mathbb{E}_\mu[g(X_1 + \cdots + X_\ell + (K-\ell)x_0)] + (1-\delta) g(Kx_0) \\ &= g(Kx_0) + \delta\gamma \times \mathbf{1}\{\ell = K\}, \end{aligned} \quad (8)$$

where $\gamma := \mathbb{E}_\mu[g(X_1 + \cdots + X_K)] - g(Kx_0) > 0$. Essentially, Eq. (8) shows that the marginal distribution of r_t conditioned on S_t is Bernoulli$(g(Kx_0))$ if $S_t \ne S^\star$, but Bernoulli$(g(Kx_0) + \delta\gamma)$ if $S_t = S^\star$.

Let $\mathbb{P}_S$ denote the distribution of $\{r_t\}_{t=1}^T$ when the adversary chooses $S^\star = S$ as the optimal subset. Let $\mathbb{P}_0$ also denote the distribution of $\{r_t\}_{t=1}^T$ with $v_{ti} \equiv x_0$. As $\{r_t\}$ are binary, Eq. (8) implies that, for every $S \subseteq [N], |S| = K$,

$$D_{\mathrm{KL}}(\mathbb{P}_0 \| \mathbb{P}_S) \le \mathbb{E}_0[T_S] \times \Gamma_{g,K}^2 \delta^2, \quad (9)$$

where $\mathbb{E}_0$ is the expectation under $\mathbb{P}_0$, $T_S := \sum_{t=1}^T \mathbf{1}\{S_t = S\}$ is the number of times S is selected and constant $\Gamma_{g,K} < \infty$ depends only on $g(Kx_0)$ and $\mathbb{E}_\mu[g(X_1, \cdots, X_K)]$ and thus only on g, K. By Pinsker's inequality,

$$\begin{aligned}\big|\mathbb{E}_0[T_S] - \mathbb{E}_S[T_S]\big| &\le T \cdot \mathrm{TV}(\mathbb{P}_0, \mathbb{P}_S) \\ &\le \Gamma_{g,K} \delta T \sqrt{\mathbb{E}_0[T_S]}. \end{aligned} \quad (10)$$

Subsequently,

$$\begin{aligned}
\mathfrak{R}(g,N,K,T) &= \max_{|S^\star|=K} \mathbb{E}\left[\sum_{t=1}^{T} R(S^\star,v_t) - R(S_t,v_t)\right] \\
&\geq \frac{1}{\binom{N}{K}} \sum_{|S^\star|=K} \mathbb{E}\left[\sum_{t=1}^{T} R(S^\star,v_t) - R(S_t,v_t)\right] \\
&= \frac{\delta\gamma}{\binom{N}{K}} \sum_{|S^\star|=K} (T - \mathbb{E}_{S^\star}[T_{S^\star}]) \\
&\geq \frac{\delta\gamma}{\binom{N}{K}} \sum_{|S^\star|=K} (T - \mathbb{E}_0[T_{S^\star}] - \left|\mathbb{E}_{S^\star}[T_{S^\star}] - \mathbb{E}_0[T_{S^\star}]\right|) \\
&= \frac{\delta\gamma}{\binom{N}{K}} \left[\binom{N}{K}T - T - \Gamma_{g,K}\delta T \sum_{|S|=K} \sqrt{\mathbb{E}_0[T_S]}\right] \quad (11) \\
&\geq \frac{\delta\gamma}{\binom{N}{K}} \left[\binom{N}{K}T - T - \Gamma_{g,K}\delta T \sqrt{\binom{N}{K}T}\right]. \quad (12)
\end{aligned}$$

Here, Eq. (11) holds because $\sum_{|S|=K} \mathbb{E}_0[T_S] = T$, and Eq. (12) is due to the Cauchy-Schwarz inequality. Setting $\delta = \min\{1, \sqrt{\binom{N}{K}}/(2\Gamma_{g,K}\sqrt{T})\}$, Eq. (12) is lower bounded by the minimum between $\Omega(T)$ and

$$\frac{\delta\gamma}{\binom{N}{K}} \left[\binom{N}{K}T - T - \binom{N}{K}\frac{T}{2}\right] = \Omega_{g,K}\left(\sqrt{\binom{N}{K}T}\right),$$

which is to be demonstrated.

The scenario when $g(\cdot)$ is a polynomial of degree $d < K$ can be proved in an entire similar way. Applying Lemma 2 with $(m,b) = (d,K)$, we could obtain a random vector $(X_1,\cdots,X_d)$ and some $x_0 \in [0,1]$ such that the conditions (6) and (7) hold. The nature uses the same strategy, with the only difference being that the size of the random subset $S^\star$ is d instead of K. In this case, any size-K set S with $S^\star \subseteq S$ gives the optimal reward, and the learner observes the non-informative outcome at time t if and only if $S^\star \nsubseteq S_t$. Consequently, both Eqs. (8, 9) still hold, with $\mathbf{1}\{\ell = K\}$ replaced by $\mathbf{1}\{\ell \geq d\}$ in Eq. (8) and the definition of T_S changed to $T_S = \sum_{t=1}^{T} \mathbf{1}\{S \subseteq S_t\}$ in Eq. (9). Using

$$\sum_{|S|=d} T_S = \sum_{t=1}^{T} \sum_{|S|=d} \mathbf{1}\{S \subseteq S_t\} = \binom{K}{d}T,$$

(12) with $|S| = d$ yields a lower bound of $\mathfrak{R}(g,N,K,T)$:

$$\frac{\delta\gamma}{\binom{N}{d}} \left[\binom{N}{d}T - \binom{K}{d}T - \Gamma_{g,d,K}\delta T \sqrt{\binom{N}{d}\binom{K}{d}T}\right].$$

Setting $\delta = \min\{1, \sqrt{\binom{N}{d}}/(2\Gamma_{g,d,K}\sqrt{\binom{K}{d}T})\}$ and noting that $\binom{K}{d}$ does not depend on (N,T), the above lower bound is simplified to

$$\Omega_{g,d,K}\left(\sqrt{\binom{N}{d}T}\right),$$

which is to be demonstrated. □

Next, we establish an $\Omega(\sqrt{N^K T})$ lower bound for the dynamic assortment optimization model described in Sec. 1.2. It implies Corollary 3 in the introduction section.

Lemma 3. *Consider the dynamic assortment optimization question with mutlinomial logit choice models described in Sec. 1.2. Adopt the unit price model, with $p_i \equiv 1$ for all $i \in [N]$. Then there exists an adversary choosing $\{v_{ti}\}_{t,i=1}^{T,N}$ and a constant $c_K > 0$ depending only on K, such that for any bandit algorithm and $N \geq K$, $T \geq 1$, it holds that*

$$\max_{|S^\star|=K} \mathbb{E}\left[\sum_{t=1}^{T} R(S^\star,v_t) - R(S_t,v_t)\right] \geq c_K \min\{1, \sqrt{N^K T}\},$$

where $R(S,v_t) = \frac{\sum_{i\in S} p_i v_{ti}}{1+\sum_{i\in S} v_{ti}}$.

Proof. It is easy to verify that $R(S,v_t) = g(\sum_{i\in S} v_{ti})$ with $g(x) = x/(1+x)$. Since the link function g here is clearly not a polynomial of any degree, the construction in Lemma 1 should immediately imply an $\Omega(\sqrt{N^K T})$ lower bound. However, in the multinomial logit choice model for dynamic assortment optimization, the bandit feedback is *not* binary. This means that expectation calculations in Eq. (8) are not sufficient, and we have to calculate the KL-divergence between discrete observations directly.

Recall that in the MNL model, the bandit feedback i_t is supported on $S_t \cup \{0\}$, with $\mathbb{P}[i_t = i|S_t,v_t] = v_{ti}/(1+\sum_{j\in S_t} v_{tj})$ for all $i \in S_t$ and $\mathbb{P}[i_t = 0|S_t,v_t] = 1/(1+\sum_{j\in S_t} v_{tj})$. Suppose $|S_t \cap S^\star| = \ell < K$. Then

$$\begin{aligned}
&\mathbb{E}_{v_t}[\mathbb{P}[i_t = 0|S_t,v_t]] \\
&= 1 - \mathbb{E}_\mu[g(X_1 + \cdots + X_\ell + (K-\ell)x_0)] \\
&= 1 - g(Kx_0) = 1/(1+Kx_0). \quad (13)
\end{aligned}$$

For every $i \in S_t \backslash S^\star$, $v_{ti} \equiv x_0$, and therefore

$$\begin{aligned}
&\mathbb{E}_{v_t}[\mathbb{P}[i_t = i|S_t,v_t]] \\
&= x_0(1 - \mathbb{E}_\mu[g(X_1 + \cdots + X_\ell + (K-\ell)x_0)]) \\
&= x_0(1 - g(Kx_0)) = x_0/(1+Kx_0). \quad (14)
\end{aligned}$$

Next consider any $i \in S_t \cap S^\star$. Because $X_1,\cdots,X_\ell$ are exchangeable, the probabilities $\mathbb{E}_{v_t}[\mathbb{P}[i_t = i|S_t,v_t]]$ are the same for all $i \in S_t \cap S^\star$. Define $\beta := \mathbb{E}_{v_t}[\mathbb{P}[i_t = i|S_t,v_t]]$

for some $i \in S_t \cap S^\star$. By the law of total probability and Eqs. (13,14) we have that

$$1 = \sum_{i \in S_t \cup \{0\}} \mathbb{E}_{v_t}[\mathbb{P}(i_t = i | S_t, v_t)] = \ell\beta + \frac{1 + (K-\ell)x_0}{1 + Kx_0}.$$

Consequently,

$$\beta = \frac{1}{\ell}\left[1 - \frac{1 + (K-\ell)x_0}{1 + Kx_0}\right] = \frac{x_0}{1 + Kx_0}. \tag{15}$$

Comparing Eqs. (13,14,15), we conclude that for any $|S_t| = K$, $S_t \neq S^\star$, $\mathbb{E}_{v_t}[\mathbb{P}[i_t = 0 | S_t, v_t]] = 1/(1 + Kx_0)$ and $\mathbb{E}_{v_t}[\mathbb{P}[i_t = i | S_t, v_t]] = x_0/(1 + Kx_0)$ for all $i \in S_t$. This shows that all $|S_t| = K$, $S_t \neq S^\star$ are information theoretically indistinguishable. On the other hand, for $S_t = S^\star$, with probability $1 - \delta$ all elements of v_t are assigned v_0, for which $\mathbb{P}[\cdot | S_t = S^\star, v_t] = \mathbb{P}[\cdot | S_t \neq S^\star, v_t]$. Hence, $D_{\mathrm{KL}}(\mathbb{P}_0 \| \mathbb{P}_{S^\star}) \leq O(\delta^2) \times \mathbb{E}_0[T_{S^\star}]$. The rest of the proof is identical to the proof of Lemma 1 when the link function g is not a polynomial. □

3. Proof of Lemma 2

We first prove the easy direction $2 \Rightarrow 1$, whose contrapositive is that if g is a polynomial of degree at most $m-1$, then (6) and (7) cannot hold simultaneously. In fact, defining $s(x) := g(x + bx_0) - g(bx_0)$ and $Y_i := X_i - x_0$ for all $i \in [m]$, condition (6) implies that for all $\ell = 1, \cdots, m-1$, it holds that $\mathbb{E}[s(Y_1 + \cdots + Y_\ell)] = 0$. By exchangeability of $(Y_1, \cdots, Y_m)$, this also shows that $\mathbb{E}[s_\ell(Y_1, \cdots, Y_m)] = 0$, where for all $\ell = 1, 2, \cdots, m-1$,

$$s_\ell(Y_1, \cdots, Y_m) := \frac{1}{\binom{m}{\ell}} \sum_{S \subseteq [m] : |S| = \ell} s\left(\sum_{i \in S} Y_i\right). \tag{16}$$

Since s is a polynomial of degree at most $m-1$ and $s(0) = 0$, the following lemma shows that $\mathbb{E}[s_m(Y_1, \cdots, Y_m)] = 0$, a contradiction to (7).

Lemma 4. *For m reals $y_1, \cdots, y_m$ and any polynomial s of degree at most $m-1$, define $s_0 \equiv s(0)$ and s_ℓ as in (16) for $\ell = 1, \cdots, m$. Then the following identity holds:*

$$\sum_{\ell=0}^{m} (-1)^\ell \binom{m}{\ell} s_\ell(y_1, \cdots, y_m) = 0.$$

Proof. By linearity it suffices to prove Lemma 4 for $s(x) = x^d$, $d \in \{0, 1, \cdots, m-1\}$. In this case, simple algebra verifies that the coefficient of $\prod_{i=1}^m y_i^{r_i}$ with $r_i \geq 0, \sum_{i=1}^m r_i = d$ in $s_\ell(y_1, \cdots, y_m)$ is

$$\frac{d!}{\prod_{i=1}^m (r_i!)} \cdot \frac{\binom{m-I(r)}{\ell - I(r)}}{\binom{m}{\ell}},$$

where $I(r) := \sum_{i=1}^m \mathbf{1}(r_i > 0)$, and $\binom{a}{b} := 0$ if $b < 0$. Consequently, the coefficient of $\prod_{i=1}^m y_i^{r_i}$ in the LHS of the equation is

$$\begin{aligned} &\frac{d!}{\prod_{i=1}^m (r_i!)} \sum_{\ell=0}^{m} (-1)^\ell \binom{m - I(r)}{\ell - I(r)} \\ &= \frac{d!}{\prod_{i=1}^m (r_i!)} \sum_{\ell=I(r)}^{m} (-1)^\ell \binom{m - I(r)}{\ell - I(r)} = 0, \end{aligned}$$

where the last identity makes use of the inequality $I(r) \leq d < m$. □

Next we prove the hard direction $1 \Rightarrow 2$. The proof makes use of the idea in convex analysis: after fixing x_0, the problem is to find an exchangeable distribution of $(X_1, \cdots, X_m)$ from the convex set of all such distributions; moreover, both constraints (6) and (7) are linear in the joint distribution, and the Dirac point measure on $(x_0, \cdots, x_0)$ satisfies (6) as well as (7) if $>$ is replaced by $\geq$. Therefore, there are two convex sets in total, one being the family of all exchangeable joint distributions and one from the constraints (6) and (7), and their closure has an intersection point, i.e. the Dirac point measure at $(x_0, \cdots, x_0)$. Now our target is to show that these sets have a non-empty intersection *without* taking the closure, and the following lemma characterizes a sufficient condition via duality.

Lemma 5. *For $g \in C([0, b])$ and a fixed $x_0 \in [0, 1]$, there exists an exchangeable Borel distribution on $[0, 1]^m$ such that both (6) and (7) hold, if the following condition holds: for every non-zero vector $\lambda = (\lambda_1, \cdots, \lambda_m)$ with $\lambda_m \geq 0$, and the functions $g_\ell(x_1, \cdots, x_m)$ defined as*

$$g_\ell(x_1, \cdots, x_m) = \frac{1}{\binom{m}{\ell}} \sum_{S \subseteq [m] : |S| = \ell} g\left(\sum_{i \in S} x_i + (b - \ell)x_0\right)$$

for $\ell = 1, 2, \cdots, m$, the point $(x_0, \cdots, x_0)$ is not a global maxima of the function $\sum_{\ell=1}^m \lambda_\ell g_\ell(x)$ over $[0, 1]^m$.

Proof. We prove the contrapositive of this result. Let X be the topological vector space of all finite Borel signed measures on $[0, 1]^m$ (which are also Radon measures by Ulam's theorem; cf. (Dudley, 2018, Theorem 7.1.4)) equipped with the weak-* topology (as the dual of $C([0, 1]^m)$), and $A \subseteq X$ be the collection of all exchangeable Borel probability measures (i.e. invariant to permutations). Moreover, for $\varepsilon > 0$, we define $B_\varepsilon \subseteq X$ as the collection of all signed Borel measures μ such that

$$\int_{[0,1]^m} (g_\ell(x) - g(bx_0)) \mu(\mathrm{d}x) = 0$$

for $\ell = 1, \cdots, m-1$ and

$$\int_{[0,1]^m} (g_m(x) - g(bx_0) - \varepsilon) \mu(\mathrm{d}x) = 0.$$

Therefore, the non-existence of such an exchangeable distribution implies that $A \cap B_\varepsilon = \emptyset$ for all $\varepsilon > 0$. Since $[0,1]^m$ is compact, the set A is a closed subset of the unit weak-* ball, and is therefore compact due to Banach–Alaoglu (see, e.g. (Rudin, 1991, Theorem 3.15)). Moreover, as $g_\ell \in C([0,1]^m)$, the set B_ε is closed under the weak-* topology. Finally, as both sets are convex, and (Rudin, 1991, Theorem 3.10) shows that the dual of X under the weak-* topology is $C([0,1]^m)$, (Rudin, 1991, Theorem 3.4) implies that there exist some $f_\varepsilon \in C([0,1]^m)$ and $\gamma_\varepsilon \in \mathbb{R}$ such that

$$\sup_{\mu \in A} \int f_\varepsilon \mathrm{d}\mu < \gamma_\varepsilon \le \inf_{\nu \in B_\varepsilon} \int f_\varepsilon \mathrm{d}\nu. \tag{17}$$

As B_ε is a linear subspace of X, the RHS of (17) must be zero (otherwise it would be $-\infty$). Then by (Rudin, 1991, Lemma 3.9), we must have $f_\varepsilon(x) = \sum_{\ell=1}^{m-1} \lambda_{\varepsilon,\ell}(g_\ell(x) - g(bx_0)) + \lambda_{\varepsilon,m}(g_m(x) - g(bx_0) - \varepsilon)$ for some vector $\lambda_\varepsilon = (\lambda_{\varepsilon,1}, \cdots, \lambda_{\varepsilon,m})$. Plugging this back into the first inequality of (17), and defining $\mu_0 \in A$ as the Dirac point measure on $(x_0, \cdots, x_0)$, we arrive at

$$\sup_{\mu \in A} \int \left(\sum_{\ell=1}^m \lambda_{\varepsilon,\ell} g_\ell(x) \right) (\mu(\mathrm{d}x) - \mu_0(\mathrm{d}x)) < \lambda_{\varepsilon,m}\varepsilon. \tag{18}$$

Since $g_\varepsilon(x) = \sum_{\ell=1}^m \lambda_{\varepsilon,\ell} g_\ell(x)$ is a symmetric function (i.e. invariant to permutations of the input), the LHS of (18) is simply $\max_{x \in [0,1]^m} g_\varepsilon(x) - g_\varepsilon(x_0, \cdots, x_0) \ge 0$. Then $\lambda_{\varepsilon,m} > 0$, and by multiplying a positive constant to all entries of λ_ε, we may assume that $\max_\ell |\lambda_{\varepsilon,\ell}| = 1$ and $\max_{x \in [0,1]^m} g_\varepsilon(x) - g_\varepsilon(x_0, \cdots, x_0) < \varepsilon$. Choosing $\varepsilon_n \to 0$, the compactness of $[-1,1]^m \ni \lambda_{\varepsilon_n}$ implies some subsequence $\lambda_{\varepsilon_{n_k}} \to \lambda$ as $k \to \infty$. Taking the limit along this subsequence, it is clear that λ is a non-zero vector with $\lambda_m \ge 0$, and $(x_0, \cdots, x_0)$ is a global maxima of the function $g(x) = \sum_{\ell=1}^m \lambda_\ell g_\ell(x)$. □

Now it remains to choose a suitable $x_0 \in [0,1]$ such that the condition in Lemma 5 holds. We choose x_0 to be any point in $(0,1)$ such that $g^{(\ell)}(bx_0) \neq 0$ for all $\ell \in [m]$, whose existence is ensured by the following lemma, which itself is a standard exercise on the Baire category theorem.

Lemma 6 (A slight variant of (Donoghue, 1969), Page 53)**.** *If $g \in C^m$ satisfies $g^{(n_x)}(x) = 0$ for some $n_x \in [m]$ and every $x \in (0,1)$, then g is a polynomial of degree at most $m-1$ on $(0,1)$.*

Now we assume by contradiction that the function $f(x) := \sum_{\ell=1}^n \lambda_\ell g_\ell(x)$ attains its maximum at $x = (x_0, \cdots, x_0)$ for some non-zero vector λ. We will prove by induction on $k = 1, 2, \cdots, m$ the following claims:

1. $D^\alpha f(x_0, \cdots, x_0) = 0$ for all multi-indices $\alpha = (\alpha_1, \cdots, \alpha_m) \in \mathbb{N}^m$ with $|\alpha| := \sum_{\ell=1}^k \alpha_\ell = k$;
2. $\sum_{\ell=1}^m \lambda_\ell \cdot \binom{m-k}{\ell-k} / \binom{m}{\ell} = 0$, with $\binom{0}{0} := 1$ and $\binom{a}{b} := 0$ for $b < 0$.

For the base case $k = 1$, the first claim is exactly the first-order condition for the maxima, and the second claim follows from the first claim, the identity $\nabla g_\ell(x_0, \cdots, x_0) = g'(bx_0)\mathbf{1} \cdot \binom{m-1}{\ell-1} / \binom{m}{\ell}$, and our choice of x_0 that $g'(bx_0) \neq 0$. Now suppose that these claims hold for $1, 2, \cdots, k-1$. Then for $|\alpha| = k$ with $I(\alpha) := \sum_{\ell=1}^m \mathbf{1}(\alpha_\ell > 0) \le k-1$, simple algebra gives that

$$D^\alpha f(x_0, \cdots, x_0) = g^{(k)}(bx_0) \cdot \sum_{\ell=1}^m \lambda_\ell \frac{\binom{m-I(\alpha)}{\ell-I(\alpha)}}{\binom{m}{\ell}} = 0,$$

where the last identity is due to the inductive hypothesis for $k' = I(\alpha) < k$. Therefore, for the first claim it remains to consider multi-indices α with $|\alpha| = I(\alpha) = k$. By the inductive hypothesis and the Taylor expansion for multivariate functions, for $\delta > 0$ sufficiently small and every binary vector $(\varepsilon_1, \cdots, \varepsilon_m) \in \{\pm 1\}^m$, it holds that

$$\begin{aligned} &f(x_0 + \delta\varepsilon_1, \cdots, x_0 + \delta\varepsilon_m) = f(x_0, \cdots, x_0) \\ &+ g^{(k)}(bx_0) \sum_{\ell=1}^m \lambda_\ell \frac{\binom{m-k}{\ell-k}}{\binom{m}{\ell}} \delta^k \sum_{S \subseteq [m]: |S| = k} \prod_{i \in S} \varepsilon_i + o(\delta^k). \end{aligned}$$

As $\mathbb{E}[\sum_{S \subseteq [m]: |S| = k} \prod_{i \in S} \varepsilon_i] = 0$ when $\varepsilon_1, \cdots, \varepsilon_m$ are Radamacher random variables, and the term inside the expectation is not identically zero, we conclude that this term could be either positive or negative after carefully choosing $(\varepsilon_1, \cdots, \varepsilon_m)$. As a result, in order for $(x_0, \cdots, x_0)$ to be a maxima of f, the above expansion implies that $\sum_{\ell=1}^m \lambda_\ell \cdot \binom{m-k}{\ell-k} / \binom{m}{\ell} = 0$ (recall that $g^{(k)}(bx_0) \neq 0$), which is exactly the second claim for k. The remainder of the first claim also follows from the second claim, and the induction is complete.

Finally we derive a contradiction from the above result, and thereby show that such a non-zero vector λ cannot exist. In fact, the second claim of the inductive result for $k = 1, 2, \cdots, m$ constitutes a linear system $A\lambda = 0$ for the vector λ, where A is a upper triangular matrix with non-zero diagonal entries. Therefore, we have $\lambda = 0$, which is a contradiction.

4. Conclusion and future directions

In this paper we study the adversarial combinatorial bandit problem with general reward functions g, with a complete characterization of the minimax regret depending on whether g is a low-degree polynomial. For the most general case when g is *not* a polynomial, including dynamic assortment optimization under the multinomial logit choice models, our results imply an $\Omega_K(\sqrt{N^K T})$ regret lower bound, hinting that it is not possible for any bandit algorithm to

exploit inherent correlation among subsets/assortments. We believe it is a promising future research direction to study models that interpolate between the stochastic and the adversarial settings, in order to achieve intermediate regret bounds.

When g is a general non-linear function, the adversarial construction in our lower bound proof can be interpreted as a *latent variable model*: first sample $v_t \sim \mu$, and then sample $r_t \sim \mathbb{P}(\cdot|S_t, v_t)$. To foster identifiability, one common assumption is to have access to multiple observations $\{r_t^1, \cdots, r_t^M\}$ conditioned on the same latent variable v_t, such as "high dimensionality" assumptions in learning Gaussian mixture models (Ge et al., 2015) and learning topic models with multiple words per document (Anandkumar et al., 2015). This motivates the following model: at the beginning of each $\tau \in \{1, 2, \cdots, T/M\}$ epoch, an adaptive adversary chooses $v_\tau \in [0, 1]^N$; afterwards, the bandit algorithm produces subsets $\{S_\tau^1, \cdots, S_\tau^M\} \subseteq [N]$ and observes feedback $r_\tau^t \sim \mathbb{P}(\cdot|S_\tau^t, v_\tau)$ for $t = 1, 2, \cdots, M$. Clearly, with $M = T$ we recover the stochastic (stationary) setting in which $\{v_t\}$ do not change, and with $M = 1$ we recover the adversarial setting in which v_t is different for each time period. An intermediate value of $1 < M < T$ is likely to result in interesting interpolation between the two settings, and we leave this as an interesting future research question.

References

Agarwal, M. and Aggarwal, V. Regret bounds for stochastic combinatorial multi-armed bandits with linear space complexity. *arXiv preprint arXiv:1811.11925*, 2018.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. Thompson sampling for the mnl-bandit. *arXiv preprint arXiv:1706.00977*, 2017.

Agrawal, S., Avadhanula, V., Goyal, V., and Zeevi, A. Mnl-bandit: A dynamic learning approach to assortment selection. *Operations Research*, 67(5):1453–1485, 2019.

Anandkumar, A., Foster, D. P., Hsu, D., Kakade, S. M., and Liu, Y.-K. A spectral algorithm for latent dirichlet allocation. *Algorithmica*, 72(1):193–214, 2015.

Audibert, J.-Y. and Bubeck, S. Minimax policies for adversarial and stochastic bandits. In *COLT*, pp. 217–226, 2009.

Audibert, J.-Y., Bubeck, S., and Lugosi, G. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2014.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE Annual Conference on Foundations of Computer Science (FOCS)*, pp. 322–331. IEEE, 1995.

Bubeck, S., Cesa-Bianchi, N., and Kakade, S. M. Towards minimax policies for online linear optimization with bandit feedback. In *Conference on Learning Theory*, pp. 41–1. JMLR Workshop and Conference Proceedings, 2012.

Cesa-Bianchi, N. and Lugosi, G. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

Chen, W., Wang, Y., and Yuan, Y. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 151–159, 2013.

Chen, W., Hu, W., Li, F., Li, J., Liu, Y., and Lu, P. Combinatorial multi-armed bandit with general reward functions. *Advances in Neural Information Processing Systems (NIPS)*, 29:1659–1667, 2016a.

Chen, W., Wang, Y., Yuan, Y., and Wang, Q. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(1):1746–1778, 2016b.

Chen, X. and Wang, Y. A note on a tight lower bound for capacitated mnl-bandit assortment selection models. *Operations Research Letters*, 46(5):534–537, 2018.

Chen, X., Wang, Y., and Zhou, Y. Dynamic assortment selection under the nested logit models. *Production and Operations Management (to appear)*, 2018a.

Chen, X., Wang, Y., and Zhou, Y. An optimal policy for dynamic assortment planning under uncapacitated multinomial logit models. *Mathematics of Operations Research (to appear)*, 2018b.

Chen, X., Krishnamurthy, A., and Wang, Y. Robust dynamic assortment optimization in the presence of outlier customers. *arXiv preprint arXiv:1910.04183*, 2019.

Combes, R., Talebi Mazraeh Shahi, M. S., Proutiere, A., et al. Combinatorial bandits revisited. *Advances in neural information processing systems*, 28:2116–2124, 2015.

Donoghue, W. F. *Distributions and Fourier transforms*. Academic Press, 1969.

Dudley, R. M. *Real analysis and probability*. CRC Press, 2018.

Ge, R., Huang, Q., and Kakade, S. M. Learning mixtures of gaussians in high dimensions. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing (STOC)*, pp. 761–770, 2015.

Kuroki, Y., Xu, L., Miyauchi, A., Honda, J., and Sugiyama, M. Polynomial-time algorithms for multiple-arm identification with full-bandit feedback. *Neural Computation*, 32(9):1733–1773, 2020.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvari, C. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pp. 535–543, 2015.

Merlis, N. and Mannor, S. Batch-size independent regret bounds for the combinatorial multi-armed bandit problem. *arXiv preprint arXiv:1905.03125*, 2019.

Merlis, N. and Mannor, S. Tight lower bounds for combinatorial multi-armed bandits. *arXiv preprint arXiv:2002.05392*, 2020.

Rejwan, I. and Mansour, Y. Top-k combinatorial bandits with full-bandit feedback. In *Algorithmic Learning Theory*, pp. 752–776. PMLR, 2020.

Rudin, W. *Functional analysis*. Mcgraw-Hill Inc, New York, 1991.

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

Train, K. *Discrete choice methods with simulation*. Cambridge University Press, 2nd edition, 2009.

Wang, S. and Chen, W. Thompson sampling for combinatorial semi-bandits. *arXiv preprint arXiv:1803.04623*, 2018.