Near-Optimal MNL Bandits Under Risk Criteria

Guangyu Xi, [1] **Chao Tao,** [2] **Yuan Zhou** [3]

[1] University of Maryland, College Park
[2] Indiana University Bloomington
[3] University of Illinois at Urbana-Champaign
gxi@umd.edu, taochao@iu.edu, yuanz@illinois.edu

Abstract

We study MNL bandits, which is a variant of the traditional multi-armed bandit problem, under risk criteria. Unlike the ordinary expected revenue, risk criteria are more general goals widely used in industries and business. We design algorithms for a broad class of risk criteria, including but not limited to the well-known conditional value-at-risk, Sharpe ratio, and entropy risk, and prove that they suffer a near-optimal regret. As a complement, we also conduct experiments with both synthetic and real data to show the empirical performance of our proposed algorithms.

Introduction

Dynamic assortment optimization is one of the fundamental problems in online learning. It has wide applications in industries, for example retailing and advertisement. To motivate the study of the problem, let us consider e-commerce companies like Amazon and Wish who want to sell products to online users when they visit the websites and search for some type of products, for example, headphones. Such companies usually have a variety of products with that type in a warehouse to sell. Due to the space constraints of a website, it is not possible to exhibit all of the available products. Hence, each time when an online user visits the website, only a limited number of products can be displayed. When an online user buys a product, the company will get some profit. So one natural goal for the company is to display on the website an assortment consisting of several products such that the expected revenue is maximized. However, in practice, a company may have more complex strategies other than simply maximizing its revenue, and general *risk criteria* may be better choices to serve such goals. For example, in risk management, a very common risk criterion called *expected shortfall* or *conditional value at risk* (CVaR) is defined as the expected revenue under a certain percentile. If we only consider the expected revenue, we may lead to focus on recommending some products producing high revenue but purchased only by a small portion of users. If the company wishes to maintain a higher level of active and diversified users, then CVaR is more appropriate. Whether it is still possible for the sales manager of the company to design

a near-optimal sales strategy when the goal is changed, for example to a kind of risk criteria, is a very practical problem and to the best of our knowledge, has not been studied before.

Suppose a company has N products of a certain category to sell during a sales season, which can be represented by $[N]$, where $[N] \stackrel{\text{def}}{=} \{1, 2, \ldots, N\}$ and each product corresponds to an element in $[N]$. Let T be the total number of times such products are searched during a sales season and S_t be the assortment displayed by the website at tth time of the request. The aforementioned sales activity can be modeled by the following game which runs in T time steps: at time step $1 \leqslant t \leqslant T$, when an assortment $S_t \subset [N]$ is displayed by the website, the online user will make a choice, i.e., whether to buy a product in S_t or purchase nothing. Following the previous motivation example, we add a cardinality constraint, which means the number of products in S_t can not exceed a predefined number $K \leqslant N$. Let c_t denote the choice of the online user at time t. When $c_t = i$, it means that the online user buys product i. For convenience, we use $c_t = 0$ to represent the situation when the online user does not purchase anything. In general, c_t can be viewed as a random variable and there is no doubt that the *multinomial logit model* (MNL) (Agrawal et al. 2019) has become the most popular one to model the behavior of the online user, i.e., c_t, when S_t is provided. Dynamic assortment optimization with the MNL choice model is also called MNL bandits. In this model, each product i is assumed to be related to an *unknown* preference parameter v_i and the probability that a visiting online user chooses product i given assortment S_t is defined by

$$\mathbb{P}(c_t = i) \stackrel{\text{def}}{=} \frac{v_i}{1 + \sum_{j \in S_t} v_j}, \qquad (1)$$

where we set the preference parameter of no-purchase $v_0 = 1$. Note that this assumption does not harm the model too much since one can easily scale v_i's to satisfy this condition. Following the literature, we also assume no-purchase is the most frequent choice i.e., $0 \leqslant v_i \leqslant 1$, which is often a reasonable assumption in sales activities.

During the last decade, MNL bandits has attracted much attention (Rusmevichientong, Shen, and Shmoys 2010; Sauré and Zeevi 2013; Agrawal et al. 2017, 2019; Dong et al. 2020). However, all of the previous works consider maxi-

mizing the expected revenue, which is not always appropriate for practical applications. In this paper, we are interested in designing algorithms for a general class of risk criteria.

Problem Formulation

Suppose for each product $i \in [N]$, selling it successfully can make the company a profit of r_i, which is *known* beforehand. Without loss of generality, we assume $r_i \in (0, 1]$. This can always be achieved by proper scaling. Moreover, the profit for no-purchase is $r_0 = 0$. Then at time step $t \geqslant 1$, when assortment $S_t \subset [N]$ and preference parameter vector $\boldsymbol{v} = (v_1, \dots, v_N)$ are provided, the profit can be represented by a random variable $X(S_t, \boldsymbol{v})$ defined by

$$\mathbb{P}(X(S_t, \boldsymbol{v}) = r_i) = \mathbb{P}(c_t = i) = \frac{v_i}{1 + \sum_{j \in S_t} v_j} \quad (2)$$

for $i = 0$ and all $i \in S_t$. In addition, we denote by $F(S_t, \boldsymbol{v})$ the cumulative distribution function to $X(S_t, \boldsymbol{v})$. Given time horizon T, one natural goal, as explained in the introduction, is to find a policy equipped by the decision maker such that the expected revenue, i.e.,

$$\sum_{t=1}^{T} R(S_t, \boldsymbol{v}) = \sum_{t=1}^{T} \mathbb{E}\left[X(S_t, \boldsymbol{v})\right] \quad (3)$$

is maximized, where $R(S_t, \boldsymbol{v})$ represents the expected profit when S_t is served. This has been investigated previously in (Agrawal et al. 2017, 2019).

In this paper, instead of expectation, we consider a general class of risk criteria. Some examples of such risk criteria can be found in (Cassel, Mannor, and Zeevi 2018). Suppose $\mathcal{D}$ is the convex set of cumulative distribution functions. In general, we consider the risk criterion U which is a function from $\mathcal{D}$ to $\mathbb{R}$. In the case of expectation, $U(F) = \int x \mathrm{d}F(x)$. In particular, since we assumed that $r_i \in (0, 1]$, we will only need $F \in \mathcal{D}[0, 1]$, where we denote by $\mathcal{D}[0, 1]$ the subspace of $\mathcal{D}$ consisting of F that is the cumulative distribution function of random variable X taking values on $[0, 1]$. The goal of this paper is to find a policy such that the following quantity

$$\mathbb{E}\left[\sum_{t=1}^{T} U(F(S_t, \boldsymbol{v}))\right] \quad (4)$$

is maximized. Let S^* be the smallest assortment such that

$$U(F(S^*, \boldsymbol{v})) = \max_{S \subset [N], |S| \leqslant K} U(F(S, \boldsymbol{v})).$$

The regret of the game after T time steps, which is a quantity measuring the difference between the optimal policy and policy π used by the decision maker, is defined as

$$\mathfrak{R}_T^\pi([N], \boldsymbol{v}, \boldsymbol{r})$$
$$\stackrel{\text{def}}{=} TU(F(S^*, \boldsymbol{v})) - \mathbb{E}\left[\sum_{t=1}^{T} U(F(S_t, \boldsymbol{v}))\right]. \quad (5)$$

where $\boldsymbol{r} = (r_1, \cdots, r_N)$ and $\boldsymbol{v} = (v_1, \cdots, v_N)$.

When it is clear from the context, we usually omit the policy π and parameters $([N], \boldsymbol{v}, \boldsymbol{r})$. Without much effort, we can see that maximizing (4) is equivalent to minimizing the regret (5).

Related Work and Our Contribution

To the best of our knowledge, we are the first to study MNL bandits under general risk criteria.

In the past decade, there have been many works on the MNL bandit problem considering maximizing the expected revenue (3). In (Rusmevichientong, Shen, and Shmoys 2010; Sauré and Zeevi 2013), the authors assumed the gap between the best and the second-to-the-best assortments is known and proposed "Explore-then-Commit" algorithms. Later in (Agrawal et al. 2019), the authors proposed the state-of-the-art UCB-type algorithm with a regret upper bound $\mathcal{O}(\sqrt{NT \ln T})$. Authors in (Agrawal et al. 2017) utilized the Bayesian method i.e., Thompson Sampling to design an algorithm which performs well in practice. For the expected revenue, it is showed in (Chen and Wang 2018) that the lower bound for the regret is $\Omega(\sqrt{NT})$.

There are a lot of previous works studying different risk criteria in multi-armed bandits (Sani, Lazaric, and Munos 2012; Maillard 2013; Galichet, Sebag, and Teytaud 2014; Zimin, Ibsen-Jensen, and Chatterjee 2014; Vakili and Zhao 2016). In (Cassel, Mannor, and Zeevi 2018), the authors established a thorough theory to deal with general risk criteria.

Our Contribution Note that directly applying the algorithm proposed in (Cassel, Mannor, and Zeevi 2018) will lead to a regret of $\Omega\left(\sqrt{\binom{N}{K}T}\right)$ since each assortment corresponds to an arm, which is far from being optimal. We can not simply take each product as an arm since the optimal assortment may consist of multiple products due to that there is an involved relationship between the risk criterion of the assortment and the underlying preference parameters, which is characterized by the following two complex structures: a general (and non-specific) risk function and the multinomial logit choice model. In this paper, we are able to gain a clearer understanding of the challenge raised by these two complex structures. To be more specific, we recognize three mild conditions that are easy to verify and show that the class identified by the aforementioned conditions encompasses most of the risk criteria that are of interest in literature (see Table 1). We also design and analyze the algorithmic framework, proving that our algorithm achieves $\widetilde{\mathcal{O}}(\sqrt{NT})$ regret for any general risk criterion that belongs to the class.

Assumptions

In this section, we first present the aforementioned three assumptions the risk criterion U should satisfy.

Assumption 1 (Quasiconvexity). *U is quasiconvex on $\mathcal{D}[0, 1]$, i.e., for any $\lambda \in [0, 1]$ and $F_1, F_2 \in \mathcal{D}[0, 1]$, it satisfies*

$$U(\lambda F_1 + (1 - \lambda)F_2) \leqslant \max\{U(F_1), U(F_2)\}. \quad (6)$$

In addition to quasiconvexity, we also make the following two assumptions on U.

Assumption 2 (Boundedness). *For any $F \in \mathcal{D}[0, 1]$ it holds that $|U(F)| \leqslant \gamma_1$.*

Assumption 3 (One-sided Lipschitz Condition). *For any* $\boldsymbol{v}' \geqslant \boldsymbol{v}$, *i.e.,* $v_i' \geqslant v_i$ *for all* $i \in [N]$, *and* $S \subset [N]$, *it holds that*

$$U(F(S, \boldsymbol{v}')) - U(F(S, \boldsymbol{v}))$$

$$\leqslant \frac{\gamma_2}{1 + \sum_{i \in S} v_i} \left[\sum_{i \in S} (v_i' - v_i) \right].$$

Note that here γ_1 and γ_2 are universal constants related to the risk criterion U.

We note that quasiconvexity is a natural assumption as most practical risk criteria considered in literature are quasiconvex. In Table 1, we give a list of the risk criteria considered, which are all quasiconvex as shown in (Cassel, Mannor, and Zeevi 2018). To complement, we also show in Table 1 that whether a risk criterion satisfies Assumption 2 and Assumption 3, and give concrete values of γ_1 and γ_2. It turns out that all the risk criteria listed in Table 1 satisfy all three assumptions except VaR, which does not meet Assumption 3 since it is discontinuous in $\boldsymbol{v}$,

Algorithms

Due to space constraints, we only show `RiskAwareUCB`, which is a variant of the UCB-type algorithm proposed in (Agrawal et al. 2019), and its guarantee. In a similar way, we also propose `RiskAwareTS`, a variant of the Thompson Sampling algorithm proposed in (Agrawal et al. 2017). Please refer to Appendix B for its precise description and near-optimal guarantee.

The high level idea of the proposed algorithm `RiskAwareUCB` is as follows. We divide all the time steps i.e., $[T]$ into small episodes. During each episode ℓ, the same assortment S_ℓ is repeatedly provided to the online user until a no-purchase outcome is observed. Specifically, in each episode ℓ, we are providing the assortment

$$\text{argmax}_{S \subset [N], |S| \leqslant K} U(F(S, \widetilde{\boldsymbol{v}}^\ell)),$$

where $\widetilde{\boldsymbol{v}}^\ell$ is an optimistic estimate of the real unknown preference parameters before the start of episode ℓ. The details of `RiskAwareUCB` is described in Algorithm 1 where $t_{i,\ell}$ is the number of times the online users buy product i in the ℓth episode and $\mathcal{T}_i(\ell)$ denotes the collection of episodes for which product i is served until episode ℓ (exclusive).

Then we have the following theoretical upper bound for `RiskAwareUCB`.

Theorem 4. *Suppose the risk criterion* U *satisfies Assumption 1, 2 and 3. The regret (5) incurred by the decision maker using* `RiskAwareUCB` *is upper bounded by* $\widetilde{\mathcal{O}}(\sqrt{NT})$ *after* T *time steps, where* $\widetilde{\mathcal{O}}$ *hides poly-logarithmic factors in* N *and* T.

Before proceeding, we first prove the following key lemma, which says that the risk gain of the optimal assortment calculated by an optimistic estimate of the preference parameters is never worse than that of S^*.

Lemma 5 (Monotone Maximum). *For any* $\boldsymbol{v}' \geqslant \boldsymbol{v}$, *it holds that*

$$\max_{S \subset [N], |S| \leqslant K} U(F(S, \boldsymbol{v}')) \geqslant U(F(S^*, \boldsymbol{v})).$$

Algorithm 1: `RiskAwareUCB`$(N, K, \boldsymbol{r}, U)$

---

1 Initialize $t \leftarrow 1$, $\ell \leftarrow 1$, $\widetilde{v}_i^\ell \leftarrow 1$ for $i \in [N]$ and $\mathcal{T}_i(\ell) \leftarrow \emptyset$ for $i \in [N]$

2 while $t \leqslant T$ **do**

3 $S_\ell \leftarrow \text{argmax}_{S \subset [N], |S| \leqslant K} U(F(S, \widetilde{\boldsymbol{v}}^\ell))$

4   Initialize $t_{i,\ell} \leftarrow 0$ for $i \in [N]$

5   **repeat**

6   Serve S_ℓ and observe customer choice c_t

7   **if** $c_t \neq 0$ **then** $t_{c_t,\ell} \leftarrow t_{c_t,\ell} + 1$

8 $t \leftarrow t + 1$

9   **until** $t > T$ *or* $c_{t-1} = 0$

10   **for** $i \in [N]$ **do**

11   **if** $i \in S_\ell$ **then** $\mathcal{T}_i(\ell + 1) \leftarrow \mathcal{T}_i(\ell) \cup \{\ell\}$

12   **else** $\mathcal{T}_i(\ell + 1) \leftarrow \mathcal{T}_i(\ell)$

13 $\ell \leftarrow \ell + 1$, $T_i(\ell) \leftarrow |\mathcal{T}_i(\ell)|$ for $i \in [N]$

14 $\bar{v}_i^\ell \leftarrow \frac{\sum_{\ell' \in \mathcal{T}_i(\ell)} t_{i,\ell'}}{T_i(\ell)}$ for $i \in [N]$

15 $\widetilde{v}_i^\ell \leftarrow \min \left\{ \bar{v}_i^\ell + \sqrt{\bar{v}_i^\ell \cdot \frac{48 \ln(\sqrt{N}\ell + 1)}{T_i(\ell)}} + \frac{48 \ln(\sqrt{N}\ell + 1)}{T_i(\ell)}, 1 \right\}$ for $i \in [N]$

Proof. Fix S, we first prove that $U(F(S, \boldsymbol{u}))$ is a quasiconvex function with respect to vector $\boldsymbol{u}$. This statement can be easily verified by noticing that for any $\lambda \in [0, 1]$ and $\boldsymbol{u}'$, we have

$$U(F(S, \lambda \boldsymbol{u} + (1 - \lambda)\boldsymbol{u}'))$$

$$= U \left(\frac{\lambda(1 + \sum_{i \in S} u_i)}{\lambda(1 + \sum_{i \in S} u_i) + (1 - \lambda)(1 + \sum_{i \in S} u_i')} F(S, \boldsymbol{u}) \right.$$

$$+ \left. \frac{(1 - \lambda)(1 + \sum_{i \in S} u_i')}{\lambda(1 + \sum_{i \in S} u_i) + (1 - \lambda)(1 + \sum_{i \in S} u_i')} F(S, \boldsymbol{u}') \right)$$

$$\leqslant \max\{U(F(S, \boldsymbol{u})), U(F(S, \boldsymbol{u}'))\}$$

where the last inequality is due to quasiconvexity of U on $\mathcal{D}[0, 1]$.

Next, we show the following lemma.

Lemma 6. *Given a quasiconvex function* $V(\boldsymbol{u})$ *defined on* $[0, 1]^n$, *suppose there is a point* $\bar{\boldsymbol{u}} = (\bar{u}_1, \cdots, \bar{u}_n) \in [0, 1]^n$ *satisfying that* $V(\bar{\boldsymbol{u}}) \geqslant V(\boldsymbol{u})$ *for any point* $\boldsymbol{u} \neq \bar{\boldsymbol{u}}$ *such that* $u_i = \bar{u}_i$ *or* 0 *for each* i *from* 1 *to* n. *Then we have that* $V(\boldsymbol{u}') \geqslant V(\bar{\boldsymbol{u}})$ *for any* $\boldsymbol{u}' \geqslant \bar{\boldsymbol{u}}$.

Proof. For the sequence of points $\boldsymbol{u}^{(i)} = (u_1^{(i)}, \ldots, u_n^{(i)})$ with $i = 1, 2, \ldots, n$ such that

$$u_j^{(i)} = \begin{cases} \bar{u}_j & j \neq i \\ 0 & j = i, \end{cases}$$

we have that $V(\boldsymbol{u}^{(i)}) < V(\bar{\boldsymbol{u}})$. For any $\boldsymbol{u}' \geqslant \bar{\boldsymbol{u}}$ and $\boldsymbol{u}' \neq \bar{\boldsymbol{u}}$, we define

$$\lambda_i = \left(\frac{u_i' - \bar{u}_i}{\bar{u}_i} \right) \left(\sum_{i=1}^n \frac{u_i' - \bar{u}_i}{\bar{u}_i} \right)^{-1}.$$

Risk Criterion (Parameter)	Property	γ_1	γ_2
VaR_α	Quasiconvex	1	Not Exist
CVaR_α	Convex	1	$3/\alpha$
nth-moment	Linear	1	1
Entropy risk (θ)	Convex	1	$2e^\theta/\theta$
Below target semi-variance (r)	Linear	r^2	$2r^2$
Negative variance	Convex	$\frac{1}{4}$	6
Mean-variance (ρ)	Convex	$1 + \frac{\rho}{4}$	$2 + 6\rho$
Sharpe ratio (r, ϵ)	Quasiconvex	$\frac{1}{\sqrt{\epsilon}}$	$2\epsilon^{-1/2} + 3\epsilon^{-3/2}$
Sortino ratio (r, ϵ)	Quasiconvex	$\frac{1}{\sqrt{\epsilon}}$	$2\epsilon^{-1/2} + \epsilon^{-3/2}$

Table 1: Widely Used Risk Criteria

Here $\lambda_i \in [0, 1]$ for all $i = 1, 2, \cdots, n$ and $\sum_{i=1}^n \lambda_i = 1$. Then the convex combination of $\boldsymbol{u}^{(i)}$

$$\widetilde{\boldsymbol{u}} = \sum_{i=1}^n \lambda_i \boldsymbol{u}^{(i)}$$

is on the same line as $\boldsymbol{u}'$ and $\bar{\boldsymbol{u}}$. By the quasiconvexity of V, we have that $V(\widetilde{\boldsymbol{u}}) \leqslant \max_{i=1}^n V(\boldsymbol{u}^{(i)}) < V(\bar{\boldsymbol{u}})$. If we define $\lambda = \frac{1}{1 + \sum_{i=1}^n \lambda_i}$, then

$$V(\bar{\boldsymbol{u}}) = V(\lambda\widetilde{\boldsymbol{u}} + (1 - \lambda)\boldsymbol{u}') \leqslant \max\{V(\widetilde{\boldsymbol{u}}), V(\boldsymbol{u}')\},$$

which means we must have $V(\bar{\boldsymbol{u}}) \leqslant V(\bar{\boldsymbol{u}}')$. $\square$

Let $S^\dagger$ be the smallest assortment such that

$$U(F(S^\dagger, \boldsymbol{v}')) = \max_{S \subset [N], |S| \leqslant K} U(F(S, \boldsymbol{v}')). \quad (7)$$

Together with Lemma 6, we obtain that $U(F(S^\dagger, \boldsymbol{v}')) \geqslant U(F(S^*, \boldsymbol{v}')) \geqslant U(F(S^*, \boldsymbol{v}))$, which concludes the proof of this lemma. $\square$

Lemma 7. *Given any $\ell > 0$ and $C_1, C_2 > 0$, we define event*

$$\mathcal{E}_\ell = \left\{ \forall i \in [N], v_i \leqslant \widetilde{v}_i^\ell \leqslant v_i + C_1 \sqrt{\frac{v_i \ln(\sqrt{N}\ell + 1)}{T_i(\ell) \vee 1}} \right.$$
$$\left. + C_2 \frac{\ln(\sqrt{N}\ell + 1)}{T_i(\ell) \vee 1} \right\}.$$

There exist real numbers $C_1, C_2 > 0$ such that

$$\mathbb{P}(\mathcal{E}_\ell) \geqslant 1 - \frac{1}{\ell}$$

for any ℓ.

Lemma 7 can be easily derived from Lemma 4.1 of (Agrawal et al. 2019). So we omit its proof here.

Proof of Theorem 4. Before proceeding, we introduce several notations. Let L be the total number of episodes when `RiskAwareUCB` stops after T steps. Denote l_ℓ by the length of the ℓth episode. Moreover, set $n_i = T_i(L)$, which is the total number of episodes product i is served before the Lth episode.

Using the law of total expectation, we rewrite the regret as

$$\mathfrak{R}_T = \mathbb{E}\left[\sum_{\ell=1}^L l_\ell \left(U(F(S^*, \boldsymbol{v})) - U(F(S_\ell, \boldsymbol{v})) \right) \right]$$
$$= \mathbb{E}\left[\sum_{\ell=1}^L \mathbb{E}[l_\ell(U(F(S^*, \boldsymbol{v})) - U(F(S_\ell, \boldsymbol{v}))) \mid \mathcal{H}_\ell] \right],$$

where $\mathcal{H}_\ell$ is the history before episode ℓ. Since S_ℓ is determined by $\mathcal{H}_\ell$, there is

$$\mathfrak{R}_T = \mathbb{E}\left[\sum_{\ell=1}^L \mathbb{E}[l_\ell \mid \mathcal{H}_\ell](U(F(S^*, \boldsymbol{v})) - U(F(S_\ell, \boldsymbol{v}))) \right].$$

Given S_ℓ, we know that l_ℓ follows a geometric distribution with parameter $1/(1 + \sum_{i \in S_\ell} v_i)$. Hence we have $\mathbb{E}[l_\ell \mid \mathcal{H}_\ell] \leqslant 1 + \sum_{i \in S_\ell} v_i$. We put inequality here since the last episode may end due to time limit. Using aforementioned inequality, we further derive

$$\mathfrak{R}_T \leqslant \mathbb{E}\left[\sum_{\ell=1}^L \mathbb{E}\left[\left(1 + \sum_{i \in S_\ell} v_i \right) \right.\right.$$
$$\left.\left. \times (U(F(S^*, \boldsymbol{v})) - U(F(S_\ell, \boldsymbol{v}))) \right] \right]$$
$$= \mathbb{E}\left[\sum_{\ell=1}^L \mathbb{E}\delta_\ell \right], \quad (8)$$

where we have defined $\delta_\ell \stackrel{\text{def}}{=} (1 + \sum_{i \in S_\ell} v_i) \times (U(F(S^*, \boldsymbol{v})) - U(F(S_\ell, \boldsymbol{v})))$.

We now focus on bounding $\mathbb{E}\delta_\ell$. By a simple calculation, we get

$$\mathbb{E}\delta_\ell = \mathbb{E}[\delta_\ell \mathbb{1}_{\mathcal{E}_\ell^c}] + \mathbb{E}[\delta_\ell \mathbb{1}_{\mathcal{E}_\ell}]$$
$$\leqslant 2\gamma_1(N + 1)\mathbb{P}(\mathcal{E}_\ell^c) + \mathbb{E}[\delta_\ell \mid \mathcal{E}_\ell]\mathbb{P}(\mathcal{E}_\ell)$$
$$\leqslant \frac{2\gamma_1(N + 1)}{\ell} + \mathbb{E}[\delta_\ell \mid \mathcal{E}_\ell]\mathbb{P}(\mathcal{E}_\ell), \quad (9)$$

where in the second last inequality, we upper bound δ_ℓ using $v_i \leqslant 1$ and Assumption 2, and the last inequality is due to

10400

Lemma 7. By Lemma 5 and Assumption 3, we get

$$\mathbb{E}[\delta_\ell \mid \mathcal{E}_\ell]$$

$$\leqslant \mathbb{E}\left[(1 + \sum_{i \in S_\ell} v_i)(U(F(S_\ell, \widetilde{\boldsymbol{v}}^\ell) - U(F(S_\ell, \boldsymbol{v})) \mid \mathcal{E}_\ell\right]$$

$$\leqslant \mathbb{E}\left[\gamma_2 \sum_{i \in S_\ell} \left(C_1 \sqrt{\frac{v_i \ln(\sqrt{N}\ell + 1)}{T_i(\ell) \vee 1}}\right.\right.$$

$$\left.\left.+ C_2 \frac{\ln(\sqrt{N}\ell + 1)}{T_i(\ell) \vee 1}\right)\right], \tag{10}$$

where in the last equality we have used the definition of event $\mathcal{E}_\ell$. By (9) and (10), we have

$$\mathbb{E}\delta_\ell \leqslant \frac{2\gamma_1(N+1)}{\ell} + \mathbb{E}\left[\gamma_2 \sum_{i \in S_\ell} \left(C_1 \sqrt{\frac{v_i \ln(\sqrt{N}\ell + 1)}{T_i(\ell) \vee 1}}\right.\right.$$

$$\left.\left.+ C_2 \frac{\ln(\sqrt{N}\ell + 1)}{T_i(\ell) \vee 1}\right)\right]. \tag{11}$$

Putting (11) back into (8), we derive

$$\mathfrak{R}_T$$

$$\leqslant 2\gamma_1(N+1)\underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L} \frac{1}{\ell}\right]}_{(*)}$$

$$+ \gamma_2 C_1 \sqrt{\ln(\sqrt{N}T + 1)}\underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L} \sum_{i \in S_\ell} \sqrt{\frac{v_i}{T_i(\ell) \vee 1}}\right]}_{(**)}$$

$$+ \gamma_2 C_2 \ln(\sqrt{N}T + 1)\underbrace{\mathbb{E}\left[\sum_{\ell=1}^{L} \sum_{i \in S_\ell} \frac{1}{T_i(\ell) \vee 1}\right]}_{(***)}. \tag{12}$$

Note that $(*) \leqslant \sum_{\ell=1}^{T} \ell^{-1} \leqslant \ln T + \gamma$, where γ is Euler's constant. Next we bound $(**)$. Let $\mathcal{K}$ be the set $\{(i, \ell) : T_i(\ell) = 0\}$. It is easy to see $|\mathcal{K}| \leqslant N$. Together with observation $\sum_{i=1}^{j} \frac{1}{\sqrt{i}} \leqslant 2\sqrt{j}$ and Jensen's inequality, we derive

$$(**) \leqslant N + \mathbb{E}\left[\sum_{i \in [N]} \left(\sqrt{v_i} \sum_{j=1}^{n_i} \frac{1}{\sqrt{j}}\right)\right]$$

$$\leqslant N + 2\mathbb{E}\left[\sum_{i \in [N]} \sqrt{v_i n_i}\right] \leqslant N + 2\sqrt{N\mathbb{E}\left[\sum_{i \in [N]} v_i n_i\right]},$$

where n_i is the total number of episodes product i is served before the Lth episode. Noting that

$$T \geqslant \mathbb{E}\left[\sum_{\ell=1}^{L-1} \mathbb{E}[l_\ell | \mathcal{H}_\ell]\right]$$

$$= \mathbb{E}\left[\sum_{\ell=1}^{L-1} (1 + \sum_{i \in S_\ell} v_i)\right] \geqslant \mathbb{E}\left[\sum_{i \in [N]} v_i n_i\right],$$

we obtain $(**) \leqslant N + 2\sqrt{NT}$. Finally, we bound $(***)$ using

$$(***) \leqslant N + \mathbb{E}\left[\sum_{i \in [N]} (\ln n_i + \gamma)\right]$$

$$\leqslant N(1 + \gamma) + N \ln \frac{\sum_{i \in [N]} n_i}{N}$$

$$\leqslant N(1 + \gamma) + N \ln T$$

Putting inequalities of $(*)$, $(**)$ and $(***)$ back into (12), we obtain

$$\mathfrak{R}_T \leqslant 2\gamma_1(N+1)(\ln T + \gamma)$$

$$+ \gamma_2 C_1 \sqrt{\ln(\sqrt{N}T + 1)}(2\sqrt{NT} + N)$$

$$+ \gamma_2 C_2 \ln(\sqrt{N}T + 1)(N \ln T + N(1 + \gamma))$$

$$= \widetilde{\mathcal{O}}(\sqrt{NT})$$

and the proof is complete. $\qquad \square$

Examples of Risk Criteria

In this section, we show that conditional value-at-risk, Sharpe Ratio, and entropy risk all satisfy Assumption 2 and Assumption 3. For the proof of the other risk criteria listed in Table 1, we refer to Appendix A.

For proving the one-sided Lipschitz condition, the following lemma is useful. The proof of Lemma 8 is in Appendix A.

Lemma 8. *For any* $\boldsymbol{v}' \geqslant \boldsymbol{v}$, *i.e.,* $v_i' \geqslant v_i$ *for all* $i \in [N]$, *and* $S \subset [N]$, *it holds that*

$$\sum_{i \in S} \left|\frac{v_i'}{1 + \sum_{i \in S} v_i'} - \frac{v_i}{1 + \sum_{i \in S} v_i}\right| \leqslant$$

$$\frac{2}{1 + \sum_{i \in S} v_i}\left[\sum_{i \in S} (v_i' - v_i)\right]. \tag{13}$$

Conditional Value-at-risk

Given $\alpha \in (0, 1]$, the conditional value-at-risk at α percentile for $F \in \mathcal{D}[0, 1]$ is defined as

$$\mathrm{CVaR}_\alpha(F) \overset{\text{def}}{=} \frac{1}{\alpha} \int_0^\alpha \mathrm{VaR}_\beta(F)\mathrm{d}\beta.$$

An equivalent definition is

$$\mathrm{CVaR}_\alpha(F) = \frac{1}{\alpha}\left(\alpha - \int_0^1 (F(x) \wedge \alpha)\mathrm{d}x\right).$$

Proposition 9. CVaR_α *satisfies Assumption 2 and Assumption 3 with* $\gamma_1 = 1$ *and* $\gamma_2 = 3/\alpha$.

Proof. It is easy to see that $|\mathrm{CVaR}_\alpha(F(S, \boldsymbol{v}))| \leqslant 1$, which implies $\gamma_1 = 1$.

We now show the value of γ_2. Without loss of generality, we can assume that the profit of different products are different since for those items with the same revenue, we can

combine them into one product and the corresponding abstraction parameter is the sum of those of the arms. Given assortment S, we denote the products in S by $[|S|]$ in the increasing order of their profit. Then for any $k + 1 \in [|S|]$ and $x \in [r_k, r_{k+1})$

$$F(S, \boldsymbol{v}'; x) - F(S, \boldsymbol{v}; x) = \frac{1 + \sum_{i=1}^{k} v_i'}{1 + \sum_{i=1}^{|S|} v_i'} - \frac{1 + \sum_{i=1}^{k} v_i}{1 + \sum_{i=1}^{|S|} v_i}.$$

Note that for any $x \in [r_k, r_{k+1})$, by Lemma 8

$$
\begin{aligned}
&|F(S, \boldsymbol{v}'; x) - F(S, \boldsymbol{v}; x)| \\
&\leqslant \left| \frac{1}{1 + \sum_{i=1}^{|S|} v_i'} - \frac{1}{1 + \sum_{i=1}^{|S|} v_i} \right| \\
&\quad + \sum_{i=1}^{k} \left| \frac{v_i'}{1 + \sum_{i=1}^{|S|} v_i'} - \frac{v_i}{1 + \sum_{i=1}^{|S|} v_i} \right| \\
&\leqslant \frac{\sum_{i=1}^{|S|} |v_i' - v_i|}{\left(1 + \sum_{i=1}^{|S|} v_i'\right)\left(1 + \sum_{i=1}^{|S|} v_i\right)} \\
&\quad + \frac{2}{1 + \sum_{i=1}^{|S|} v_i} \left[\sum_{i \in S} (v_i' - v_i) \right] \\
&\leqslant \frac{3}{1 + \sum_{i=1}^{|S|} v_i} \left[\sum_{i \in S} (v_i' - v_i) \right].
\end{aligned}
$$

Clearly the difference between $F(S, \boldsymbol{v}'; x) \wedge \alpha$ and $F(S, \boldsymbol{v}; x) \wedge \alpha$ satisfies the same bound. Hence

$$
\begin{aligned}
&|\mathrm{CVaR}_\alpha(F(S, \boldsymbol{v}')) - \mathrm{CVaR}_\alpha(F(S, \boldsymbol{v}))| \\
&\leqslant \frac{1}{\alpha} \left(\int_0^1 |F_X(x) \wedge \alpha - F_Y(x) \wedge \alpha| \mathrm{d}x \right) \\
&\leqslant \frac{3/\alpha}{1 + \sum_{i=1}^{|S|} v_i} \left[\sum_{i=1}^{|S|} (v_i' - v_i) \right].
\end{aligned}
$$

$\square$

Sharpe Ratio

Given a minimum average reward $r \in [0, 1]$ and the regularization factor ϵ, for $F \in \mathcal{D}[0, 1]$ we define

$$Sh_{r,\epsilon}(F) = \frac{U^1(F) - r}{\sqrt{\epsilon + \sigma^2(F)}},$$

where $U^1(F)$ is the mean and $\sigma^2(F)$ is the variance.

Proposition 10. $Sh_{r,\epsilon}$ *satisfies Assumption 2 and Assumption 3 with* $\gamma_1 = \frac{1}{\sqrt{\epsilon}}$ *and* $\gamma_2 = 2\epsilon^{-1/2} + 3\epsilon^{-3/2}$.

Proof. Since $\sigma^2(F) > 0$, $U^1(F) \in [0, 1]$ and $r \in [0, 1]$, it is easy to see that $\gamma_1 = \frac{1}{\sqrt{\epsilon}}$. For the value of γ_2, by the Lipschitz property of the mean and variance (see Appendix

A), we have

$$
\begin{aligned}
&|Sh_{r,\epsilon}(F(S, \boldsymbol{v}')) - Sh_{r,\epsilon}(F(S, \boldsymbol{v}))| \\
&\leqslant \left| \frac{U^1(F(S, \boldsymbol{v}')) - U^1(F(S, \boldsymbol{v}))}{\sqrt{\epsilon + \sigma^2(F(S, \boldsymbol{v}'))}} \right| + |U^1(F(S, \boldsymbol{v})) - r| \\
&\quad \times \left| \frac{1}{\sqrt{\epsilon + \sigma^2(F(S, \boldsymbol{v}'))}} - \frac{1}{\sqrt{\epsilon + \sigma^2(F(S, \boldsymbol{v}))}} \right| \\
&\leqslant \frac{1}{\sqrt{\epsilon}} |U^1(F(S, \boldsymbol{v}')) - U^1(F(S, \boldsymbol{v}))| \\
&\quad + \frac{|\sigma^2(F(S, \boldsymbol{v}')) - \sigma^2(F(S, \boldsymbol{v}))|}{\sqrt{\epsilon + \sigma^2(F(S, \boldsymbol{v}'))} \sqrt{\epsilon + \sigma^2(F(S, \boldsymbol{v}))}} \\
&\quad \times \frac{1}{\sqrt{\epsilon + \sigma^2(F(S, \boldsymbol{v}'))} + \sqrt{\epsilon + \sigma^2(F(S, \boldsymbol{v}))}} \\
&\leqslant \frac{1}{\sqrt{\epsilon}} |U^1(F(S, \boldsymbol{v}')) - U^1(F(S, \boldsymbol{v}))| \\
&\quad + \frac{1}{2\epsilon^{\frac{3}{2}}} |\sigma^2(F(S, \boldsymbol{v}')) - \sigma^2(F(S, \boldsymbol{v}))| \\
&\leqslant \frac{2\epsilon^{-1/2} + 3\epsilon^{-3/2}}{1 + \sum_{i \in S} v_i} \left[\sum_{i \in S} (v_i' - v_i) \right].
\end{aligned}
$$

$\square$

Entropy Risk

Given the risk aversion parameter $\theta > 0$, the entropy risk measure for $F \in \mathcal{D}[0, 1]$ is defined as

$$U^{ent}(F) = -\frac{1}{\theta} \ln \left(\int_0^1 e^{-\theta x} \mathrm{d}F(x) \right).$$

Proposition 11. U^{ent} *satisfies Assumption 2 and Assumption 3 with* $\gamma_1 = 1$ *and* $\gamma_2 = 2e^\theta / \theta$.

Proof. By Jensen's inequality, we always have $|U^{ent}(F)| \leqslant 1$. Given a fixed $\theta > 0$, we know that

$$\sum_{i \in S} e^{-\theta r_i} \frac{v_i}{1 + \sum_{i \in S} v_i} \in [e^{-\theta}, 1].$$

By the convexity of the log function and Lemma 8, we have that

$$
\begin{aligned}
&|U^{ent}(F(S, \boldsymbol{v}')) - U^{ent}(F(S, \boldsymbol{v}))| \\
&= \frac{1}{\theta} \left| \ln \left(\sum_{i \in S} \frac{e^{-\theta r_i} v_i'}{1 + \sum_{i \in S} v_i'} \right) - \ln \left(\sum_{i \in S} \frac{e^{-\theta r_i} v_i}{1 + \sum_{i \in S} v_i} \right) \right| \\
&\leqslant \frac{e^\theta}{\theta} \left| \sum_{i \in S} \frac{e^{-\theta r_i} v_i'}{1 + \sum_{i \in S} v_i'} - \sum_{i \in S} \frac{e^{-\theta r_i} v_i}{1 + \sum_{i \in S} v_i} \right| \\
&\leqslant \frac{2e^\theta / \theta}{1 + \sum_{i \in S} v_i} \left[\sum_{i \in S} (v_i' - v_i) \right].
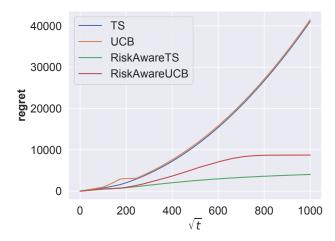\end{aligned}
$$

$\square$

Figure 1: Synthetic Data

Figure 2: Real Data

Experimental Evaluation

We evaluate `RiskAwareUCB` and `RiskAwareTS` against UCB and TS where the last two algorithms are set to maximize the expected revenue in both synthetic and real data. Based on the BanditPyLib library (Holtz, Tao, and Xi 2020), all of the algorithms[1] are implemented in Python3.

Synthetic Data In this experiment, we fix the number of products $N = 10$, cardinality limit $K = 4$, horizon $T = 10^6$, and set the goal to be $U = \text{CVaR}_{0.5}$. We generate 10 uniformly distributed random input instances where $v_i \in [0, 1]$ and $r_i \in [0.1, 1]$. For each input instance, we run 20 repetitions and compute their average as the regret. Figure 1 shows how the worst regret among all input instances changes with square root of time.

Real Data In this experiment, we consider the "UCI Car Evaluation Database" dataset from the Machine Learning Repository (Dua and Graff 2017) which contains 6 categorical attributes for $N = 1728$ cars and consumer ratings for each car. We fix cardinality limit $K = 100$, horizon $T = 10^6$, and set the goal to be $U = \text{CVaR}_{0.05}$.

By transforming each attribute to a one-hot vector, we obtain an attribute vector $m_i \in \{0, 1\}^{21}$ for each car. There are four different values for customer ratings i.e., "acceptable", "good", "very good", and "unacceptable". We decode "unacceptable" by 0 and others by 1 to represent whether the customer has the intention to buy the car. We use logistic regression to predict whether the customer is likely to buy the car and the probability that the customer buys car i is modeled by

$$\frac{1}{1 + \exp(-\theta^{\text{T}} m_i)},$$

where $\theta \in \mathbb{R}^{21}$ is an *unknown* parameter. After the model is fit with L_2 regularization, we set the preference parameter v_i
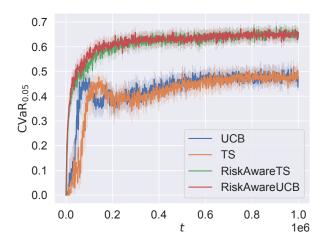
of car i to be the same as the probability predicted by logistic regression. Since there is no profit data available for cars in this dataset, we generate uniformly distributed profit r_i from $[0.1, 1]$ for each car.

We run the experiment for 40 repetitions and compute the average $\text{CVaR}_{0.05}$ for every consecutive 1000 revealed profits. To save time, when computing the assortment with the best $\text{CVaR}_{0.05}$, we do a local search, i.e., try to replace a car, add a car or delete a car, and stops if we can not find a better assortment. Figure 2 shows the results of the experiment.

Discussion From Figure 1, we can see that both `RiskAwareUCB` and `RiskAwareTS` suffer a $\sqrt{t}$-rate regret. Moreover, `RiskAwareTS` performs better than `RiskAwareUCB`, which aligns with literature that Thompson Sampling performs better in practice. From Figure 2, we can see that the obtained $\text{CVaR}_{0.05}$ under UCB and TS are far from optimal. However, `RiskAwareUCB` and `RiskAwareTS` perform roughly the same. For both of these experiments, we can see the proposed algorithms `RiskAwareUCB` and `RiskAwareTS` perform much better than UCB and TS.

Conclusion

In this work, we have shown the near-optimal algorithms for a general class of risk criteria, which only need to satisfy three mild assumptions. Experiments with both synthetic and real data are conducted to validate our results and show that ordinary algorithms suffer a worse performance when the goal is changed.

Acknowledgments

---

[1]Please refer to https://github.com/Alanthink/aaai2021 for the source code.

References

Agrawal, S.; Avadhanula, V.; Goyal, V.; and Zeevi, A. 2017. Thompson Sampling for the MNL-Bandit. In *Conference on Learning Theory (COLT)*.

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

Cassel, A.; Mannor, S.; and Zeevi, A. 2018. A general approach to multi-armed bandits under risk criteria. In *Conference on Learning Theory (COLT)*.

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

Dong, K.; Li, Y.; Zhang, Q.; and Zhou, Y. 2020. Multinomial Logit Bandit with Low Switching Cost. In *International Conference on Machine Learning (ICML)*.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml (Accessed on 2020-09-08).

Galichet, N.; Sebag, M.; and Teytaud, O. 2014. Exploration vs Exploitation vs Safety: Risk-averse Multi-Armed Bandits. In *Asian Conference on Machine Learning (ACML)*.

Holtz, C.; Tao, C.; and Xi, G. 2020. BanditPyLib: a lightweight python library for bandit algorithms. Online at: https://github.com/Alanthink/banditpylib. URL https://github.com/Alanthink/banditpylib. Documentation at https://alanthink.github.io/banditpylib-doc.

Kamath, G. 2015. Bounds on the Expectation of the Maximum of Samples from a Gaussian. http://www.gautamkamath.com/writings/gaussian_max.pdf (Accessed on 2020-09-08).

Maillard, O. 2013. Robust Risk-Averse Stochastic Multi-armed Bandits. In *Algorithmic Learning Theory (ALT)*.

Rusmevichientong, P.; Shen, Z. M.; and Shmoys, D. B. 2010. Dynamic Assortment Optimization with a Multinomial Logit Choice Model and Capacity Constraint. *Operations Research* 58(6): 1666–1680.

Sani, A.; Lazaric, A.; and Munos, R. 2012. Risk-Aversion in Multi-armed Bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 3284–3292.

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

Vakili, S.; and Zhao, Q. 2016. Risk-Averse Multi-Armed Bandit Problems Under Mean-Variance Measure. *IEEE Journal of Selected Topics in Signal Processing* 10(6): 1093–1111.

Zimin, A.; Ibsen-Jensen, R.; and Chatterjee, K. 2014. Generalized Risk-Aversion in Stochastic Multi-Armed Bandits. *arXiv preprint arXiv:1405.0833* .