Dynamic pricing and assortment under a contextual MNL demand

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider dynamic multi-product pricing and assortment problems under an unknown demand over T periods, where in each period, the seller decides on the price for each product or the assortment of products to offer to a customer who chooses according to an unknown Multinomial Logit Model (MNL). Such problems arise in many applications, including online retail and advertising. We propose a randomized dynamic pricing policy based on a variant of the Online Newton Step algorithm (ONS) that achieves a $O(d\sqrt{T}\log(T))$ regret guarantee under an adversarial arrival model. We also present a new optimistic algorithm for the adversarial MNL contextual bandits problem, which achieves a better dependency than the state-of-the-art algorithms in a problem-dependent constant κ (potentially exponentially small). Our regret upper bound scales as $\tilde{O}(d\sqrt{\kappa T} + \log(T)/\kappa)$, which gives a stronger bound than the existing $\tilde{O}(d\sqrt{T}/\kappa)$ guarantees.

1 Introduction

In this paper, we consider the contextual dynamic pricing and assortment optimization problems faced by a seller who sequentially observes a contextual demand under bandit feedback. The goal is to learn the underlying model parameters in order to minimize the seller's cumulative regret. These problems arise in numerous applications, including pricing and product recommendation in online retail, as well as click-through rate predictions for web search results. Sequential learning is especially important in settings involving short selling seasons or where no historical data is available.

In both the dynamic pricing and assortment problems, we consider a seller who sells a set of N products over a time horizon of T periods. A new customer arrives in each period t and is offered a set of products or prices. In the pricing problem, the customer arrives with a consideration set and the seller has to decide on the prices of the different products in the consideration set. In the assortment problem, the seller selects the subset of products to offer to the customer. In both cases we assume that the customer's purchase decisions are made according to a multinomial logit model (MNL)([33], [29]) which is widely used in modeling customer preferences.

Non-contextual dynamic pricing, where the seller sells identical products to the customers over time, was initially investigated by [23] under various assumptions on the demand curve, and has since been extensively studied (see for instance [8], [15], [7] and [16]). We refer the reader to [14] for an in-depth survey of the area. On the other hand, dynamic assortment optimization was first studied by [9] under the assumption of independent demand for the different products. Dynamic assortment under a MNL choice model has been recently considered (see for instance [37], [38] and [40]). In particular, UCB and Thompson sampling-based policies are proposed in [3] and [4]. These policies achieve optimal regrets of $O(\sqrt{NT})$ in the non-feature based setting. Another related stream of literature is that of combinatorial bandits (see for instance [11], [24] and [35]), and in particular, top k combinatorial bandits ([36]). In this framework, the agent can pull a subset of arms of cardinality less

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

than k in each round and the total reward obtained is a function of the individual rewards for the arms played. However, the reward obtained in our setting for each individual arm depends on the whole set of arms played in period t, whereas the rewards are supposed independent in the aforementioned works.

Feature-based dynamic pricing has recently received a lot of attention and been well characterized when only a single product is involved at each step (see for instance [20], [21], [5] [6], [13], [28], [27], [25] and [34]). The proposed policies achieve the optimal regrets of $\tilde{O}(\log(T))$ or $\tilde{O}(\sqrt{T})$, depending on whether a well-separability assumption is satisfied or not. We would like to highlight [6], which is the first work to incorporate contextual information about the customers under the form of feature-dependent price sensitivities.

The MNL contextual dynamic assortment problem is an extension of the generalized linear bandits problem, which has been widely studied (see for instance [18], [26] and [2]). UCB and TS based policies have recently been proposed for the MNL contextual dynamic assortment problem (see for instance [31], [32] and [12]).

Problem formulation. In this paper, we consider a setting where multiple products are offered to the customer in each period. We study a feature-based model, where the product utilities are a function of both products and customers features. In particular, we assume that in each period t, the customer's utility for each product $j \in [N]$ can be written under the form $x_{t,j}^\top \theta^*$, where θ^* is an unknown model parameter and $x_{t,j} \in \mathbb{R}^d$ is a feature vector, which can be adversarially chosen. Furthermore, we allow feature-dependent price sensitivities in our pricing setting and suppose that the price sensitivities can be expressed as $x_{t,j}^\top \alpha^*$ for some unknown parameter α^*. Typically, the number of products is significantly larger than the number of features ($d << N$). The objective is to learn across the products and obtain a regret upper bound scaling with d instead of N.

1.1 Our contributions.

In this work, we introduce two new algorithms for the dynamic contextual pricing and assortment problems. Our main contribution is the following.

Dynamic pricing. We present a dynamic pricing policy for the multi-product MNL model with adversarial contexts and feature-dependent price sensitivities that achieves a $O(d\log(T)\sqrt{T})$ regret bound. This is near-optimal given the $\Omega(\sqrt{T})$ lower bound from [22]. Based on structural properties of the MNL model, and more specifically, on the self-concordant-like property of the MNL log likelihood function, we propose to use a variant of the Online Newton Step method ([19]) to update our estimators of the model parameters. We combine it with a random price shock policy to force exploration.

Closest to our pricing setting is [22], in which the authors design a multi-product dynamic pricing policy under a MNL choice model and feature-dependent price sensitivities. The proposed algorithm achieves a $\tilde{O}(\sqrt{T})$ regret. However, it relies on a bayesian assumption, namely the feature vectors are drawn i.i.d. from some unknown distribution. Our work considers an adversarial context and uses an ONS method to update the parameters. The connection with the Online Convex Optimization framework has been exploited in the literature (see for example the stochastic online gradient descent in [20] for the single product setting without price sensitivities). The problem we consider is more challenging since the presence of feature-dependent price sensitivities implies the existence of *uninformative prices* (i.e. prices p such that no pricing policy can hope to learn the true model parameters when repeatedly pricing at price p). Note that all prices are informative in [20]. The link between uninformative prices and the difficulty to design low regret policies was first pointed out by [8], which shows that no algorithm can achieve better regret than $\Omega(\sqrt{T})$ in settings involving such prices. We address this challenge by using appropriate randomized price shocks that force exploration. Adding random shocks is a classical exploration strategy in sequential learning problems and has been considered in the context of dynamic pricing (see for instance [30]).

Our results imply a $O(d\log(T))$ regret in the single product setting without price sensitivity considered in [20], under the extra assumption that the noise follows a logistic distribution. This improves over the $O(\sqrt{T})$ regret bound the authors provide. We have become aware that a concurrent

recent paper [41] obtained a logarithmic regret for the contextual single product pricing problem without price sensitivity through a variant of the online newton method. We would like to underscore that our theoretical results were obtained independently from this work. Furthermore, the algorithm proposed in [41] uses the exp-concavity parameter in the descent step and the regret upper bound scales with this potentially exponentially small constant. By leveraging the self-concordance property of the logistic loss, we design an algorithm which does not use such parameter. We present in Appendix D some numerical comparison between the two algorithms, which confirms that our algorithm achieves a significantly better regret than the one in [41] as the exp-concavity parameter decreases.

In addition to our main contribution, we present a new algorithm for MNL contextual bandits.

MNL contextual bandits. We consider a setting with uniform product revenues and propose a new UCB-based algorithm for the MNL assortment problem with adversarial contexts. Our algorithm achieves a regret bound of order $\tilde{O}(d\sqrt{T})$, which is optimal as a function T. One major limitation of the state-of-the-art algorithms for the MNL contextual bandits problem ([31], [32] and [12]) is that the regret bounds scale with a problem dependent constant, which we will refer to as κ (see Section 3 for a formal definition); κ quantifies the level of non-linearity of the model and can be exponentially small even for moderate size instances ([17]). Our regret bound can be expressed as $C_1 d\sqrt{\kappa T} + C_2 d^2 \log(T)/\kappa$. Therefore, we give a significantly stronger bound than [31], whose regret term of order $\sqrt{T}$ scales with $1/\kappa$. Note that our first order term improves for small values of κ (when the model is further from the linear model). Moreover, a prior knowledge of the value of κ is often presupposed in the existing algorithms, which may be a major hindrance to their practical implementation. The quantity κ appears for example in the design of the exploration bonuses of the UCB-MNL algorithm of [31]. Our algorithm does not rely on an a priori knowledge of κ. Our policy is based on optimistic parameter search instead of exploration bonuses, as used in [31]. The analysis relies on a concentration result on the MLE estimators, which uses a generalization of the Bernstein-like tail inequality for self-normalized vectorial martingales in [2] and [17], and leverages the self-concordant property of the MNL log loss.

We would like to mention that a similar result is achieved by [17] and [2] in the logistic bandits setting. Our results cannot, however, be derived from these two works as the MNL contextual bandits problem cannot be formulated as a generalized bandits problem [12]. In this work, we show that the techniques from [17] and [2] can be efficiently extended to the MNL bandit setting. This involves overcoming a few technical challenges that are specific to the MNL problem.

Notations. For any vector $x \in \mathbb{R}^d$ and any positive definite matrix $M \in \mathbb{R}^{d\times d}$, let $||x||_M = \sqrt{x^T M x}$. Also let $B_p(0, W)$ be the d-dimensional ball of radius W under the norm ℓ^p. In the special case of the norm ℓ^2, we will drop the index and refer to the d-dimensional ball as $B(0, W)$. For two symmetric matrices $A, B \in \mathbb{R}^{d\times d}$, $A \succeq B$ means that $A - B$ is semi-definite positive. For $n \in \mathbb{N}$, we use the notation $[n]$ to denote the set $\{1, ..., n\}$. When it is not clear from the context, we will use a bold font to denote vectors.

2 Multi-product dynamic pricing

2.1 Model setting and preliminaries

We consider a dynamic pricing problem for a seller with N products represented by feature vectors $x_1, x_2, ..., x_N \in \mathbb{R}^d$. In each period t, a customer arrives with context $z_t \in \mathbb{R}^d$ and a consideration set $S_t \subseteq [N]$. The sequence of customer features and consideration sets can be chosen adversarially. The seller offers prices $p_{t,j}$ for all $j \in S_t$ and the customer makes a selection based on a MNL model where the attraction parameter depends on product and customer features. For ease of notation, we consider a more general setting where customer features z_t and consideration set S_t are represented by $|S_t|$ adversarially chosen feature vectors $x_{t,1}, ..., x_{t,|S_t|} \in \mathbb{R}^d$.

Notations. For all $t \in [T]$, let $k_t = |S_t|$ and let $K \leq N$ be an upper bound on k_t for all t. We also define X_t as the matrix whose rows are $x_{t,j}$ for $j \in [k_t]$ and we refer to it as the context at time t.

After seeing the prices $\{p_{t,j}\}_{j\in[k_t]}$, the customer purchases product $j \in [k_t] \cup \{0\}$ with probability:

$$q_{t,j}((\theta^*, \alpha^*), \vec{p_t}) = \begin{cases} \frac{e^{x_{t,j}^\top \theta^* - x_{t,j}^\top \alpha^* p_{t,j}}}{1+\sum_{i=1}^{k_t} e^{x_{t,i}^\top \theta^* - x_{t,i}^\top \alpha^* p_{t,i}}} & \text{if } j \geq 1 \\ \frac{1}{1+\sum_{i=1}^{k_t} e^{x_{t,i}^\top \theta^* - x_{t,i}^\top \alpha^* p_{t,i}}} & \text{if } j = 0 \end{cases},$$

where $\theta^*, \alpha^* \in \mathbb{R}^d$ are two model parameters. For all $j \in [k_t]$, $x_{t,j}^\top \alpha^*$ represents the price sensitivity of customer t for product j. Note that the customer has always the possibility to leave without making any purchase (by selecting product $j = 0$). The binary variable $y_{t,j}$ indicates whether the customer has purchased product j at time t.

For brevity, let $\gamma^* = (\theta^*, \alpha^*)$ and $\tilde{x}_{t,j} = [x_{t,j}, -p_{t,j}x_{t,j}]$. We can more simply write the utility $u_{t,j} = x_{t,j}^\top \theta^* - x_{t,j}^\top \alpha^* p_{t,j}$ associated with each product as $u_{t,j} = \tilde{x}_{t,j}^\top \gamma^*$. We make the following assumptions, which are standard in the dynamic pricing literature:

Assumption 2.1. For all t, $j \in [k_t]$, $||x_{t,j}||_2 \leq 1$.

Assumption 2.2. $||(\theta^*, \alpha^*)||_2 \leq W$ for some known constant W.

Although the contexts $\{x_{t,j}\}_{t\in[T],j\in[k_t]}$ can be adversarially chosen, we need to slightly restrict the set of feasible contexts to guarantee the positiveness of the price sensitivity for all products. Following [22], we make the following assumption.

Assumption 2.3. For all $t \in T, j \in [k_t]$, $x_{t,j}^\top \alpha^* \geq L$ for some known constant L.

Pricing policy and Benchmark. We consider non-anticipating pricing policies π, which depend only on the history up to time $t-1$, $\mathcal{H}_{t-1} = (X_1, \vec{p_1}, \vec{y_1}, ..., X_{t-1}, \vec{p_{t-1}}, \vec{y_{t-1}})$, and the current context X_t. The objective of the seller is to minimize its cumulative expected regret:

$$R^\pi(T) = \sum_{t=1}^{T} \left[\sum_{i=1}^{k_t} q_{t,i}(\gamma^*, \vec{p_t^*}) p_{t,i}^* - \mathbb{E}^\pi \left(\sum_{i=1}^{k_t} q_{t,i}(\gamma^*, \vec{p_t}) p_{t,i} \right) \right],$$

where $\vec{p_t^*}$ is the optimal vector of prices in period t. The expectation is taken over the random feedback and any source of randomization in the policy.

2.2 Dynamic pricing policy

Our algorithm combines the two following ingredients: a variant of the ONS method and random price shocks. In particular, we maintain estimators of the parameters which are updated in each iteration by using a variant of ONS on the log likelihood. In each step, our algorithm selects a myopic vector of prices based on the current estimators and adds random price shocks to force exploration and avoid uninformative prices.

We first give the details on the update of the parameters. Let ℓ_t denote the log loss at time t:

$$\ell_t(\gamma) = -\sum_{j=0}^{k_t} y_{t,j} \log(q_{t,j}(\gamma, \vec{p_t})).$$

For a time-dependent sequence of positive regularizers $\{\lambda_t\}_{t\geq 1}$, the estimator obtained before projection after conducting one step of our descent method is:

$$\hat{\gamma_t} = \gamma_{t-1} - \tfrac{1}{\mu} H_{t-1}^{-1} \nabla \ell_{t-1}(\gamma_{t-1}),$$

where $\mu = \frac{1}{2(1+\sqrt{6KW})}$ and $H_t = \sum_{s=1}^{t} \nabla^2 \ell_s(\gamma_s) + \lambda_t I_d$ is the regularized Hessian of the negative log likelihood. Note that for the MNL model, ℓ_t is convex, hence $H_t \succ 0$ for all t.

Let B_t denote the set $B(0, W) \cap \{(\theta, \alpha) \mid x_{t,j}^\top \alpha \geq L \text{ for } j \in [k_t]\}$. B_t represents the set of parameters which satisfy Assumptions 2.2 and 2.3. We obtain the new estimator γ_t by projecting $\hat{\gamma_t}$ on the feasible set of parameters:

$$\gamma_t = \Pi_{B_t}^{H_{t-1}}(\hat{\gamma_t}),$$

where $\Pi_{B_t}^{H_{t-1}}(y) = \arg\min_{x\in B_t} (x-y)^T H_{t-1}(x-y)$ is the projection relatively to the norm induced by H_{t-1}. As a result, during all the course of the algorithm, our estimator γ_t also satisfies the lower and upper boundedness assumptions.

Finally, the seller chooses a perturbation factor $\delta_t = \frac{1}{Wt^{1/4}}$ and computes, independently for each product $j \in [k_t]$, a random price shock $\Delta p_{t,j}$ which takes value δ_t with probability $1/2$ and $-\delta_t$ with probability $1/2$. The seller posts the vector of prices $\vec{p_t}$ which is the sum of the random price shocks and the myopic vector of prices $g(X_t\alpha_t, X_t\theta_t)$ (see Appendix B for a formal definition). The pseudo code of our dynamic pricing policy is presented in Algorithm 1.

Algorithm 1 Online Newton method for multiple product pricing

Input: Upper bound W on $||(\theta^*, \alpha^*)||_2$, lower bound L, sequence of regularizers $\{\lambda_t\}_{t\geq 1}$.

Initialize $\theta_1, \alpha_1 \in B_1$
for all $t \geq 1$ **do**
For $j \in [k_t]$, let $\Delta p_{t,j} = \begin{cases} \frac{1}{Wt^{1/4}} \text{ w/p } \frac{1}{2} \\ \frac{-1}{Wt^{1/4}} \text{ w/p } \frac{1}{2} \end{cases}$
For $j \in [k_t]$, set $p_{t,j} = g(X_t\alpha_t, X_t\theta_t)_j + \Delta p_{t,j}$
Post prices $\vec{p_t}$
Set $\gamma_{t+1} = (\theta_{t+1}, \alpha_{t+1})$ as follows:

$$\gamma_{t+1} = \Pi_{B_{t+1}}^{H_t}\left(\gamma_t - \frac{1}{\mu}H_t^{-1}\nabla\ell_t(\gamma_t)\right)$$

end for

Using an online descent method for the parameters estimation allows us to obtain a low regret algorithm despite the presence of adversarial contexts. We would like to note that based on our current analysis, a simpler online method such as a stochastic online gradient descent (as proposed in [20] for single-product dynamic pricing without price sensitivity) would not allow us to obtain sublinear regret. We would also like to mention that our method is different from the Online Newton Step presented in [19], which is the classic online analogue of the Newton method. The ONS method moves into the direction of $A_t^{-1}\nabla\ell_t$, where A_t^{-1} is an approximation of the inverse of the Hessian. In our case, we move directly into the direction of the inverse of the Hessian multiplied by $\nabla\ell_t$, and leverage self-concordant-like properties of the negative log likelihood function of the MNL model to show the convergence of the estimators. This allows us to avoid using the exp-concavity parameter $1/\kappa$ (often referred to as β in the literature) in the descent step, as is done by the ONS algorithm.

We are ready to present our regret bound.

Theorem 2.4. *Setting $\lambda_1 = 1$, $\lambda_t = d\log(t)$ for all $t \geq 1$, there is a constant C depending only on W, L, K such that the regret of Algorithm 1 is bounded as:*

$$R^\pi(T) \leq Cd\log(T)\sqrt{T}.$$

The proof of Theorem 2.4 is presented in Appendix B. We would like to point out that following [22], we treat the maximum size of a consideration set K as a constant.

Under the assumptions of [20] (single product dynamic pricing with adversarial contexts and constant price sensitivity), and assuming that the noise has a logistic distribution, we can show that our policy achieves a regret with a dependency on T of order $O(\log(T) + \sum_{t=1}^T t\delta_t)$ where $\{\delta_t\}_{t=1}^T$ represent possible drifts in the ground truth parameter. This contrasts with the $O(\sqrt{T} + \sum_{t=1}^T \sqrt{t}\delta_t)$ upper bound established in [20].

2.3 High level ideas and sketch of the proof.

We provide here the main ideas in the proof of Theorem 2.4. The technical details are presented in Appendix B. We first follow the classical regret analysis for dynamic pricing policies and decompose the regret between a term due to the error in the estimation of the parameters and a term due to the random price shocks. In particular, we have the following lemma.

Lemma 2.5. *There exist constants C_1, C_2 depending only on W, L, K, such that:*

$$R^\pi(T) \leq \mathbb{E}\left[C_1C_2^2\sum_{t=1}^T\sum_{j=1}^{k_t}((\alpha_t - \alpha^*)^T x_{t,j})^2 + ((\theta_t - \theta^*)x_{t,j})^2) + C_1\sum_{t=1}^T ||\Delta\vec{p_t}||^2\right].$$

Since the variances of the random price shocks are $\frac{1}{W^2\sqrt{t}}$, the second term is $O(\sqrt{T})$. Therefore, to exhibit a $\tilde{O}(\sqrt{T})$ regret bound, it suffices to focus on upper bounding the first term. Note that it follows from (1) that the regret upper bound does not require the global convergence of the estimators to the true parameters. We only need to show that they converge sufficiently fast in the directions given by the contexts seen throughout the T periods.

For any sequence of prices $\{\vec{p_t}\}_{t=1}^T$, our online descent method allows us to control the convergence of the estimated utilities $u_{t,j} = x_{t,j}^\top \theta_t - x_{t,j}^\top \alpha_t p_{t,j}$ to the true utilities $u_{t,j}^* = x_{t,j}^\top \theta^* - x_{t,j}^\top \alpha^* p_{t,j}$ for each product $j \in [k_t]$. In particular, we have the following lemma.

Lemma 2.6. *There is a constant $\tilde{C}$ depending only on W, L, K such that with probability at least $1 - \frac{\log(T)}{T^2}$,*

$$\sum_{t=1}^{T} \sum_{j=1}^{k_t} (x_{t,j}^\top (\theta_t - \theta^*) - x_{t,j}^\top (\alpha_t - \alpha^*) p_{t,j})^2 \leq \tilde{C} d \log(T).$$

We present the proof in Appendix B. Note that the upper bound in Lemma 2.6 is valid for any sequence of prices posted by the seller. However, it is meaningful only with random price experimentation. Suppose at each step we only post the myopic vector of prices $\vec{p_t} = g(X_t\alpha_t, X_t\theta_t)$. If the prices $\{p_{t,j}\}$ happened to be uninformative (i.e. $x_{t,j}^\top (\theta_t - \theta^*) - x_{t,j}^\top (\alpha_t - \alpha^*) p_{t,j} = 0$), then the bound provided in the above lemma would be empty. However, adding random price shocks allows us to deviate from uninformative prices and to derive an upper bound on the first term of (1) based on Lemma 2.6. This concludes the proof of Theorem 2.4. Note that Lemma 2.6 immediately implies a $O(d \log(T))$ regret bound when the model does not involve feature-dependent price sensitivities.

In [20], where no price sensitivity is involved and the utility is simply written under the form $u_t = x_t^\top \theta^*$, the use of a stochastic gradient descent method allows the author to directly obtain a $O(\sqrt{T})$ bound on the sum $\sum_{t=1}^T x_t^\top (\theta_t - \theta^*)^2$. Such a bound, when combined with our random price shocks, would only give us a linear regret. By exploiting the special structure of the MNL function and using our variant of the ONS instead of an online gradient descent, we obtain the stronger $O(\log(T))$ bound of Lemma 2.6.

The proof of Lemma 2.6 is based on a lower and an upper bound on $\sum_{t=1}^T \ell_t(\gamma_t) - \ell_t(\gamma^*)$. Both involve $\sum_{t=1}^T \sum_{j=1}^{k_t} (x_{t,j}^\top (\theta_t - \theta^*) - x_{t,j}^\top (\alpha_t - \alpha^*) p_{t,j})^2$. The proof of the lower bound exploits the convexity of ℓ_t and is based on a Bernstein inequality for martingales difference sequences. This is similar to the inequality used in [20]. The main technical hindsight lies in the proof of the upper bound. It mimics the analysis of the Online Newton Step method presented in [19], but relies on the specific structure of the gradient and Hessian of ℓ_t for the MNL model. Moreover, it unically exploits the self concordant-like property of ℓ_t. Let's first recall the definition of a self-concordant-like function.

Definition 2.7 (self-concordant-like functions [39])**.** A convex function $f \in C^3(\mathbb{R}^n)$ is called a self-concordant-like function if:

$$|\phi'''(t)| \leq M_f \phi''(t) ||u||_2$$

for $t \in \mathbb{R}$ and $M_f > 0$, where $\phi(t) := f(x + tu)$ for any $x \in \text{dom}(f)$ and $u \in \mathbb{R}^n$.

By adapting the proof of Lemma 4 in [39], we can show that ℓ_t satisfies the following property.

Proposition 2.8. *ℓ_t is self-concordant-like with $M_f = X\sqrt{6|S_t|}$.*

We detail in Appendix A some useful properties satisfied by self-concordant-like functions.

3 Improved algorithm for MNL contextual bandits

3.1 Problem formulation

We consider the following MNL dynamic assortment optimization problem, also referred as the MNL contextual bandits. In each period t, the seller observes feature vectors $\{x_{t,j}\}_{j=1}^N \in \mathbb{R}^d$. As before, this represents a combination of customer and product features which can be adversarially chosen.

The seller needs to decide on the set $S_t \subseteq [N]$ to offer, with $|S_t| \leq K$. Given the offered assortment, the customer purchases product $j \in S_t \cup \{0\}$ with probability:

$$q_{t,j}(S_t, \theta^*) = \begin{cases} \frac{e^{x_{t,j}^\top \theta^*}}{1+\sum_{i \in S_t} e^{x_{t,i}^\top \theta^*}} & \text{if } j \in S_t \\ \frac{1}{1+\sum_{i \in S_t} e^{x_{t,i}^\top \theta^*}} & \text{if } j = 0 \end{cases} \tag{1}$$

where $\theta^* \in \mathbb{R}^d$ is an underlying model parameter. As before, the binary variable $y_{t,j}$ indicates whether the customer has purchased product j at time t. Note that our model encompasses the contextual logistic bandit problem with finitely many arms (which corresponds to the case where $K = 1$). The objective is to minimize the cumulative expected regret over the T periods:

$$R(T) = \sum_{t=1}^{T} \left[\sum_{j \in S_t^*} q_{t,j}(S_t^*, \theta^*) - \sum_{j \in S_t} q_{t,j}(S_t, \theta^*) \right],$$

where S_t^* denotes the optimal assortment at time t.

In [31], the authors study a more general model where a reward $r_{t,j} \in [0, 1]$ is also revealed for each product at time t. We first consider the case of uniform rewards ($r_{t,j} = 1$) for all products. Maximizing $\sum_{j \in S_t} q_{t,j}(S_t, \theta)$ over all sets $S \subseteq [N]$ of cardinality at most K is now equivalent to selecting the K products which have the highest utility $x_{t,j}^\top \theta$. Hence the set S_t as well as the optimal set S_t^* always contain exactly K products.

Similarly as in the pricing setting, we make the following two assumptions:

Assumption 3.1. For all $t \in [T]$, $j \in [N]$, $||x_{t,j}||_2 \leq 1$.

Assumption 3.2. $||\theta^*||_2 \leq W$ for some known constant W.

Following [31], we also introduce the following constant, which typically appears in connection to the link function in the generalized linear bandits ([18], [26]).

$$\kappa := \min_{|S| \leq K, j \in S, t \geq 1} \min_{\|\theta^*\| \leq W} q_{t,j}(S, \theta) q_{t,0}(S, \theta) > 0.$$

A smaller value of κ can be interpreted as a bigger deviation from the linear model. As mentioned before, the regret bound of the dynamic assortment policies of [31] and [12] exhibit a harmful dependency in κ. Besides, an a priori knowledge of the value of κ is presupposed. Our goal is to design a dynamic assortment policy which does not require prior knowledge of κ and achieves a regret with a better dependency in κ. For all $t \in [T]$, let

$$\kappa_t^* = \sum_{j \in S_t^*} q_{t,j}(S_t^*, \theta^*) q_{t,0}(S_t^*, \theta^*).$$

κ_t^* represents the degree of non-linearity for the optimal set S_t^* and depends on the unknown parameter θ^* as well as on the feature vectors present at time t. We show that the $\tilde{O}(\sqrt{T})$ term of our regret bound can be replaced by a $\tilde{O}\left(\sqrt{\sum_{t=1}^{T} \kappa_t^*}\right)$ term. Note that we always have $\kappa_t^* \leq 1$ hence this is a strict improvement. As a result, a high level of non-linearity at time t induces a smaller κ_t^* and positively impacts the regret.

3.2 Dynamic assortment policy

We design a tight confidence set for the true parameter θ^* and use it to construct upper confidence bounds on the utility of each product at time t. Our algorithm relies on optimistic parameter search over the confidence interval, as used by [2] in the logistic bandits setting. However, in our setting, the seller's decision at time t involves the choice of multiple products. Hence we cannot build a unique optimistic estimator θ_t as in [2]. The key idea is to do the optimistic parameter search independently for each product, generating a set of parameters $\{\tilde{\theta}_{t,j}\}_{j=1}^N$ such that with high probability, $x_{t,j}^\top \tilde{\theta}_{t,j}$ is an upper bound on the utility $x_{t,j}^\top \theta^*$ of product j.

Confidence set. The main ingredient is the design of a confidence set for θ^*. We classically start by computing the maximum likelihood estimator of θ^*. Let $\hat{\theta}_t$ be the unique minimizer of the following function, for a sequence of time-dependent regularizers $\{\lambda_t\}_{t=1}^T$:

$$\mathcal{L}_t^{\lambda_t}(\theta) = -\sum_{s=1}^{t-1}\sum_{j\in S_s} y_{s,j}\log(q_{s,j}(S_s,\theta)) + \tfrac{\lambda_t}{2}\|\theta\|^2.$$

$\hat{\theta}_t$ satisfies the equation $\nabla\mathcal{L}_t^{\lambda_t}(\theta) = 0$, where the gradient of $\mathcal{L}_t^{\lambda_t}(\theta)$ is given by:

$$\nabla\mathcal{L}_t^{\lambda_t}(\theta) = \sum_{s=1}^{t-1}\sum_{j\in S_t}(q_{s,j}(S_t,\theta) - y_{t,j})x_{t,j} + \lambda_t\hat{\theta}_t.$$

Following [2], for all $\delta \in [0,1)$, we now define a confidence set $C_t(\delta)$ for θ^* as follows:

$$C_t(\delta) := \{\theta \in \Theta | \|g_t(\theta) - g_t(\hat{\theta}_t)\|_{H_t^{-1}(\theta)} \le \gamma_t(\delta)\}, \text{ where } g_t(\theta) := \sum_{s=1}^{t-1}\sum_{j\in S_s} q_{s,j}(S_t,\theta)x_{t,j} + \lambda_t\theta,$$

$H_t(\theta)$ is the Hessian of the regularized negative log likelihood evaluated at θ,

$$H_t(\theta) := \sum_{s=1}^{t-1}\left[\sum_{i\in S_s} q_{t,i}(S_t,\theta)x_{t,i}x_{t,i}^T - \sum_{i\in S_t}\sum_{j\in S_t} q_{t,i}(S_t,\theta)q_{t,j}(S_t,\theta)x_{t,i}x_{t,j}^T\right] + \lambda_t I_d$$

and

$$\gamma_t(\delta) = \sqrt{\lambda_t}(W + \tfrac{1}{2}) + \frac{2d}{\sqrt{\lambda_t}}\log\left(\frac{4}{\delta}\left(1 + \frac{2tK}{d\lambda_t}\right)\right).$$

The following proposition is the analogue, in the multi-product setting, of Proposition 1 in [2] and establishes that $C_t(\delta)$ is a confidence set for θ^*. The details are given in Appendix C.3.

Proposition 3.3. *Let $\delta \in (0,1]$. Then $\mathbb{P}(\forall t, \theta^* \in C_t(\delta)) \ge 1 - \delta$.*

The proof of Proposition 3.3 builds upon the new Bernstein-like tail inequality for self-normalized vectorial martingales presented in [17]. However, Theorem 1 in [17] does not directly apply to our setting because of the correlation between the variables $\{\varepsilon_{t,j} := y_{t,j} - q_{t,j}(S_t,\theta^*)\}_{j\in S_t}$ induced by the presence of multiple purchase options in period t. We thus present a generalization of Theorem 1 in [17] for the multiple products setting that handles such correlation.

Algorithm. Before describing our algorithm, let's introduce the following notation, which generalizes the choice probabilities given by (1) to the case where the model parameters corresponding to each product are uncorrelated. More precisely, we now consider $\tilde{\theta} \in \mathbb{R}^{d\times N}$. For all $j \in [N]$, $\tilde{\theta}_j$ represents the model parameter associated with product j. The probability that item j is purchased at time t if assortment $S \subseteq [N]$ is offered is then given by:

$$\tilde{q}_{t,j}(S,\tilde{\theta}) = \begin{cases} \frac{e^{x_{t,j}^\top\tilde{\theta}_j}}{1+\sum_{i\in S} e^{x_{t,i}^\top\tilde{\theta}_i}} & \text{if } j \in S \\ \frac{1}{1+\sum_{i\in S} e^{x_{t,i}^\top\tilde{\theta}_i}} & \text{if } j = 0 \end{cases}$$

Now, at time t, our algorithm uses the previous contexts and observations to compute the maximum likelihood estimator $\hat{\theta}_t$ as defined above and constructs the confidence set $C_t(\delta)$. Then, for each product $j \in [N]$, the algorithm finds an optimistic parameter $\tilde{\theta}_{t,j} = \arg\max_{\theta\in C_t(\delta)} x_{t,j}^\top\theta$. Let $z_{t,j}$ denote the resulting upper bound on the utility of product j, i.e., $z_{t,j} = x_{t,j}^\top\tilde{\theta}_{t,j}$. We offer the set S_t of the K items maximizing the optimistic expected revenue $\tilde{r}_t(S,\tilde{\theta})$:

$$\tilde{r}_t(S,\tilde{\theta}) = \sum_{j\in S}\tilde{q}_{t,j}(S,\tilde{\theta}).$$

We present the following upper bound on the regret of our dynamic assortment policy.

Theorem 3.4. *For $\delta = \frac{1}{K^2T^2}$, $\lambda_1 = 1$, and $\lambda_t = d\log(tK)$ for all $t \ge 1$, the regret of Algorithm 2 satisfies, for some universal constants $\tilde{C}_1$ and $\tilde{C}_2$:*

$$R(T) \le \tilde{C}_1 K^{3/2} d\log(KT)\left(\sqrt{\sum_{t=1}^{T}\kappa_t^*}\right) + \frac{\tilde{C}_2 d^2 K^4}{\kappa}\log(KT)^2.$$

Algorithm 2 OFU-MNL

Input: Upper bound K on the size of an assortment, δ, sequence $\{\lambda_t\}_{t=1}^T$.

for all $t \geq 1$ **do**
Observe feature vectors $\{x_{t,1}, ..., x_{t,N}\}$.
Set $\hat{\theta}_t = \arg\min \mathcal{L}_t^{\lambda_t}(\theta)$.
For $j = 1, ..., N$, set

$$\tilde{\theta}_{t,j} = \underset{\theta \in C_t(\delta)}{\arg\max}\, x_{t,j}^\top \theta.$$

Offer set $S_t = \arg\max_{|S| \leq K} \tilde{r}_t(S, \tilde{\theta})$.
Receive feedback $\vec{y_t}$.
end for

3.3 High level ideas and sketch of the proof

We first condition on the event that $\theta^* \in C_t(\delta)$, for all $t \geq 1$. Proposition 3.3 shows that this happens with high probability. From Lemma C.3 (Appendix C.2), we have the following concentration result on the optimistic parameter $\tilde{\theta}_{t,j}$ associated with each product $j \in [N]$:

$$\|\theta^* - \tilde{\theta}_{t,j}\|_{H_t(\theta^*)} \leq (1 + \sqrt{6KW})\gamma_t(\delta). \tag{2}$$

Now, using the optimistic choice of assortment made by the algorithm as well as the fact that $x_{t,j}^\top \tilde{\theta}_{t,j}$ is an upper bound on the true utility for each product, we can first bound the regret as follows:

$$R(T) = \sum_{t=1}^T \left[\sum_{j \in S_t^*} q_{t,j}(S_t^*, \theta^*) - \sum_{j \in S_t} q_{t,j}(S_t, \theta^*) \right] \leq \sum_{t=1}^T \left[\sum_{j \in S_t} \tilde{q}_{t,j}(S_t, \tilde{\theta}_t) - \sum_{j \in S_t} \tilde{q}_{t,j}(S_t, \tilde{\theta}^*) \right]$$
$$\leq \sum_{t=1}^T \sum_{j \in S_t} q_{t,j}(S_t, \theta^*) q_{t,0}(S_t, \theta^*) x_{t,j}^\top (\tilde{\theta}_{t,j} - \theta^*) + R_2(T) := R_1(T) + R_2(T).$$

where $R_2(T)$ is a second order term which we will prove to be of order $O(d^2 K^4 \log(KT)^2/\kappa)$. The second inequality follows from the fact that $\tilde{r}_t(S_t, \tilde{\theta}_t) \geq \tilde{r}_t(S_t^*, \tilde{\theta}^*)$ and the last one results from a Taylor expansion. To show that the first term is of order $O\left(K^{3/2} d \log(KT) \sqrt{\sum_{t=1}^T \kappa_t^*}\right)$, we first use the concentration result stated in (2), which implies the following upper bound, for some $C > 0$:

$$R_1(T) \leq C\sqrt{d \log(KT)} \sum_{t=1}^T \sum_{j \in S_t} \left[q_{t,j}(S_t, \theta^*) q_{t,0}(S_t, \theta^*) ||x_{t,j}||_{H_t(\theta^*)^{-1}} \right].$$

Using that $H_t(\theta^*) \succeq \kappa \sum_{s=1}^T \sum_{j \in S_s} x_{s,j} x_{s,j}^\top$, we could already show that this term is of order $\tilde{O}(\sqrt{T})$ by applying the Elliptical Potential Lemma from [1]. However, we would then obtain a term with linear dependency in $1/\kappa$. We show that the local information given by the terms $\{q_{t,j}(S_t, \theta^*) q_{t,0}(S_t, \theta^*)\}$ and the self-concordance-like property of the log loss can be used to derive a tighter bound on the above sum. The complete version of the proof is provided in Appendix C.2.

4 Conclusion

In this paper, we study contextual dynamic pricing and assortment optimization problems under a MNL choice model. We present a dynamic pricing policy based on a variant of the Online Newton Step method combined with random price shocks that achieves near-optimal regret for the MNL model with adversarial contexts and feature-dependent price sensitivities. We also propose a new optimistic algorithm for the adversarial MNL contextual bandits problem. Both our algorithms leverage the self-concordant property of the MNL log likelihood function to achieve better dependency on potentially exponentially small parameters than existing algorithms. An interesting research direction would be to extend our results to other choice models, such as the nested logit model, which is another widely used model in the Revenue Management literature.

References

[1] Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. "Improved Algorithms for Linear Stochastic Bandits". In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor et al. Vol. 24. Curran Associates, Inc., 2011, pp. 2312–2320.

[2] Marc Abeille, Louis Faury, and Clément Calauzènes. *Instance-Wise Minimax-Optimal Algorithms for Logistic Bandits*. 2020. arXiv: 2010.12642 [cs.LG].

[3] Shipra Agrawal et al. "MNL-Bandit: A Dynamic Learning Approach to Assortment Selection". In: *CoRR* abs/1706.03880 (2017). arXiv: 1706.03880.

[4] Shipra Agrawal et al. "Thompson Sampling for the MNL-Bandit". In: *CoRR* abs/1706.00977 (2017). arXiv: 1706.00977.

[5] Kareem Amin, Afshin Rostamizadeh, and Umar Syed. "Repeated Contextual Auctions with Strategic Buyers". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc., 2014, pp. 622–630.

[6] GahhYi Ban and N. Bora Keskin. "Personalized Dynamic Pricing with Machine Learning". In: *SSRN Electronic Journal* (Jan. 2017). DOI: 10.2139/ssrn.2972985.

[7] Omar Besbes and Assaf Zeevi. "Dynamic Pricing Without Knowing the Demand Function: Risk Bounds and Near-Optimal Algorithms". In: *Operations Research* 57.6 (2009), pp. 1407–1420. DOI: 10.1287/opre.1080.0640. eprint: https://doi.org/10.1287/opre.1080.0640.

[8] Josef Broder and Paat Rusmevichientong. "Dynamic Pricing Under a General Parametric Choice Model". In: *Operations Research* 60 (Aug. 2012), pp. 965–980. DOI: 10.2307/23260288.

[9] Felipe Caro and Jérémie Gallien. "Dynamic Assortment with Demand Learning for Seasonal Consumer Goods". In: *Management Science* 53.2 (2007), pp. 276–292. DOI: 10.1287/mnsc.1060.0613. eprint: https://doi.org/10.1287/mnsc.1060.0613.

[10] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. DOI: 10.1017/CBO9780511546921.

[11] Wei Chen, Yajun Wang, and Yang Yuan. "Combinatorial Multi-Armed Bandit: General Framework and Applications". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research 1. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, pp. 151–159.

[12] Xi Chen, Yining Wang, and Yuan Zhou. *Dynamic Assortment Optimization with Changing Contextual Information*. 2019. arXiv: 1810.13069 [econ.EM].

[13] Maxime C. Cohen, Ilan Lobel, and Renato Paes Leme. "Feature-based Dynamic Pricing". In: *EC '16*. 2016.

[14] Arnoud V. Den Boer. "Dynamic pricing and learning: Historical origins, current research, and new directions". In: *Surveys in Operations Research and Management Science* 20 (June 2015). DOI: 10.1016/j.sorms.2015.03.001.

[15] Arnoud V. Den Boer and Bert Zwart. "Mean square convergence rates for maximum quasi-likelihood estimator". In: *Stochastic Systems* 4 (Jan. 2014), pp. 375–403. DOI: 10.1214/12-SSY086.

[16] Vivek F. Farias and Benjamin Van Roy. "Dynamic Pricing with a Prior on Market Response". In: *Operations Research* 58.1 (2010), pp. 16–29. DOI: 10.1287/opre.1090.0729. eprint: https://doi.org/10.1287/opre.1090.0729.

[17] Louis Faury et al. *Improved Optimistic Algorithms for Logistic Bandits*. 2020. arXiv: 2002.07530 [cs.LG].

[18] Sarah Filippi et al. "Parametric Bandits: The Generalized Linear Case". In: *Advances in Neural Information Processing Systems*. Ed. by J. Lafferty et al. Vol. 23. Curran Associates, Inc., 2010, pp. 586–594.

[19] Elad Hazan et al. "Logarithmic Regret Algorithms for Online Convex Optimization". In: June 2006, pp. 499–513. DOI: 10.1007/11776420_37.

[20] Adel Javanmard. *Perishability of Data: Dynamic Pricing under Varying-Coefficient Models*. 2017. arXiv: 1701.03537 [cs.GT].

[21] Adel Javanmard and Hamid Nazerzadeh. *Dynamic Pricing in High-dimensions*. 2018. arXiv: 1609.07574 [stat.ML].

[22] Adel Javanmard, Hamid Nazerzadeh, and Simeng Shao. *Multi-Product Dynamic Pricing in High-Dimensions with Heterogeneous Price Sensitivity*. 2019. arXiv: 1901.01030 [stat.ML].

[23] Robert Kleinberg and Tom Leighton. "The Value of Knowing a Demand Curve: Bounds on Regret for Online Posted-Price Auctions". In: *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*. FOCS '03. USA: IEEE Computer Society, 2003, p. 594. ISBN: 0769520405.

[24] Branislav Kveton et al. "Tight Regret Bounds for Stochastic Combinatorial Semi-Bandits". In: *CoRR* abs/1410.0949 (2014). arXiv: 1410.0949.

[25] Renato Paes Leme and Jon Schneider. "Contextual Search via Intrinsic Volumes". In: *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*. Ed. by Mikkel Thorup. IEEE Computer Society, 2018, pp. 268–282. DOI: 10.1109/FOCS.2018.00034. URL: https://doi.org/10.1109/FOCS.2018.00034.

[26] Lihong Li, Yu Lu, and Dengyong Zhou. "Provable Optimal Algorithms for Generalized Linear Contextual Bandits". In: *CoRR* abs/1703.00048 (2017). arXiv: 1703.00048.

[27] Allen Liu, Renato Paes Leme, and Jon Schneider. "Optimal Contextual Pricing and Extensions". In: *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021*. Ed. by Dániel Marx. SIAM, 2021, pp. 1059–1078. DOI: 10.1137/1.9781611976465.66. URL: https://doi.org/10.1137/1.9781611976465.66.

[28] Ilan Lobel, Renato Paes Leme, and Adrian Vladu. "Multidimensional Binary Search for Contextual Decision-Making". In: *Oper. Res.* 66.5 (2018), pp. 1346–1361. DOI: 10.1287/opre.2018.1722. URL: https://doi.org/10.1287/opre.2018.1722.

[29] D. Mcfadden. "Modelling the Choice of Residential Location". In: *Transportation Research Record* (1977).

[30] Mila Nambiar, David Simchi-Levi, and He Wang. "Dynamic Learning and Pricing with Model Misspecification". In: *Management Science* 65.11 (2019), pp. 4980–5000. DOI: 10.1287/mnsc.2018.3194. eprint: https://doi.org/10.1287/mnsc.2018.3194.

[31] Oh and Iyengar. "Multinomial Logit Contextual Bandits". In: 2019.

[32] Min-hwan Oh and Garud Iyengar. "Thompson Sampling for Multinomial Logit Contextual Bandits". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019, pp. 3151–3161.

[33] R. L. Plackett. "The Analysis of Permutations". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 24.2 (1975), pp. 193–202. ISSN: 00359254, 14679876.

[34] Sheng Qiang and Mohsen Bayati. *Dynamic Pricing with Demand Covariates*. 2016. arXiv: 1604.07463 [stat.ML].

[35] Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. "Contextual Combinatorial Bandit and its Application on Diversified Online Recommendation". In: Apr. 2014, pp. 461–469. ISBN: 978-1-61197-344-0. DOI: 10.1137/1.9781611973440.53.

[36] Idan Rejwan and Yishay Mansour. "Combinatorial Bandits with Full-Bandit Feedback: Sample Complexity and Regret Minimization". In: *CoRR* abs/1905.12624 (2019). arXiv: 1905.12624.

[37] Paat Rusmevichientong, Zuo-Jun Max Shen, and David B. Shmoys. "Dynamic Assortment Optimization with a Multinomial Logit Choice Model and Capacity Constraint". In: *Operations Research* 58.6 (2010), pp. 1666–1680. DOI: 10.1287/opre.1100.0866. eprint: https://doi.org/10.1287/opre.1100.0866.

[38] Denis Sauré and Assaf Zeevi. "Optimal Dynamic Assortment Planning with Demand Learning". In: *Manufacturing & Service Operations Management* 15.3 (2013), pp. 387–404. DOI: 10.1287/msom.2013.0429. eprint: https://doi.org/10.1287/msom.2013.0429.

[39] Quoc Tran-Dinh, Yen-Huan Li, and Volkan Cevher. *Composite convex minimization involving self-concordant-like cost functions*. 2018. arXiv: 1502.01068 [math.OC].

[40] Yining Wang, Xi Chen, and Yuan Zhou. "Near-Optimal Policies for Dynamic Multinomial Logit Assortment Selection Models". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018, pp. 3101–3110.

[41] Jianyu Xu and Yu-Xiang Wang. "Logarithmic Regret in Feature-based Dynamic Pricing". In: *CoRR* abs/2102.10221 (2021). arXiv: 2102.10221. URL: https://arxiv.org/abs/2102.10221.

[42] Lijun Zhang et al. "Online Stochastic Linear Optimization under One-bit Feedback". In: *ICML*. 2016.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix for the complete proofs.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?[Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]