Dynamic pricing and assortment under a contextual MNL demand

Vineet Goyal
Columbia University
New York, NY
vgoyal@ieor.columbia.edu

Noemie Perivier
Columbia University
New York, NY
np2708@columbia.edu

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_2 (potentially exponentially small). Our regret upper bound scales as $\tilde{O}(d\sqrt{\kappa_2 T} + \log(T)/\kappa_2)$, which gives a stronger bound than the existing $\tilde{O}(d\sqrt{T}/\kappa_2)$ 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 maximize the seller's profit. 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.

Problem formulation. 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)([34], [30]) which is widely used in modeling customer preferences. In both problems, the objective is to maximize the seller's expected profit, which is equivalent to minimizing its cumulative expected regret (i.e., the difference between the optimal value and the value obtained by following a given policy).

In this paper, 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 (see Sections 2.1 and 3.1 for exact definitions of the problems). 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.

Literature review. Non-feature-based dynamic pricing, where the seller sells identical products to the customers over time, was initially investigated by [24] 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, non-feature-based 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 [38], [39] and [41]). 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.

In this paper, we consider feature-based pricing and assortment problems. Feature-based dynamic pricing has recently received a lot of attention. Most of the existing work study a single product pricing setting under various demand models (linear, binary, generalized linear) and assume that the price sensitivity of each product (i.e., the price coefficient in the demand model) is a known constant (see for instance [21], [22], [5] , [13], [29], [28], [26] and [35]). In the single product setting, the most related work is [6], which is the first to incorporate contextual information about the customers under the form of feature-dependent price sensitivities. The benefit of this assumption is illustrated on real datasets. [6] considers a high dimensional setting with a sparsity assumption under a generalized demand model and propose a policy with near-optimal regret. However, they assume i.i.d. features, whereas we consider a more specific demand model (MNL model), but with adversarial features. In [23], is considered a multiple product version of the problem, with feature-based price sensitivities and under the assumption that the demand follows a MNL model. In this last work, the features are also assumed to be i.i.d. Our multi-product pricing setting directly extends [23], and captures the single product setting with unknown (and feature-based) price sensitivity and adversarial features under a binary demand model with logistic noise, for which, to the best of our knowledge, no algorithm with near-optimal regret is known yet.

The MNL feature-based dynamic assortment problem is a variant of the contextual generalized linear bandits problem (see for instance [18], [27] and [2]) with a more complicated state space, in which multiple arms are pulled at the same time. UCB and TS based policies have recently been proposed for the MNL contextual dynamic assortment problem (see for instance [33], [32] and [12]). However, these work suffer from a dependency in a problem-dependent constant κ_2 (potentially exponentially small), which captures the 'degree of non-linearity' of the problem.

Another related stream of literature is that of combinatorial bandits (see for instance [11], [25] and [36]), and in particular, top k combinatorial bandits ([37]). In this framework, the agent can pull a subset of arms of cardinality less 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.

1.1 Our contributions.

In this work, we introduce 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 [23]. 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 ([20]) 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 [23], 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

2

are drawn i.i.d. from some unknown distribution. Our work considers an adversarial context and uses an ONS method to update the parameters. Note that the connection with the Online Convex Optimization framework has been exploited in the literature (see for example the stochastic online gradient descent in [21] for the single product setting without price sensitivities), however, 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 learn the true model parameters when repeatedly pricing at price p). Note that all prices are informative in [21]. 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 (we note that adding random shocks was first used in the context of dynamic pricing in [31]; however, our work is the first to simultaneously use random price shocks and an ONS based update, and the analysis of our policy differs from the analysis in [31]).

Our results imply a $O(d\log(T))$ regret in the single product setting with adversarial contexts and without price sensitivity considered in [21], under the extra assumption that the noise follows a logistic distribution. This improves over the $O(\sqrt{T})$ regret bound of [21] in this case (note that [21] studies more precisely the effect of drifts in the parameters). Note that in the same setting, but with feature-dependent price coefficients, our results also imply a $\tilde{O}(\sqrt{T})$ regret. We have become aware that a concurrent recent paper [42] also obtained a logarithmic regret for the contextual single product pricing problem without price sensitivity through a variant of the online newton method (for a more general demand model with strictly log-concave noise). We would like to underscore that our theoretical results were obtained independently from this work. Furthermore, the algorithm proposed in [42] uses the exp-concavity parameter in the descent step and the regret upper bound provided 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, which may be better in practical applications. However, our regret upper bound still depends on this constant. 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 [42] 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 $\tilde{O}(dK\sqrt{T})$, which is optimal as a function T. One major limitation of the state-of-the-art algorithms for the MNL contextual bandits problem ([33], [32] and [12]) is that the regret bounds scale with a problem dependent constant, which we will refer to as κ_2 (see Section 3 for a formal definition); κ_2 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 $C_1 d\sqrt{\sum_{t=1}^{T} \kappa_{2,t}^*} + C_2 d^2 \log(T)/\kappa_2$ (where $\kappa_{2,t}^* \leq 1$ are problem dependent constants, whose value decrease as the model is further away from the linear model). Therefore, we give a significantly stronger bound than [33], whose regret term of order $\sqrt{T}$ scales with $1/\kappa_2$. Note that our first order term improves for small values of $\{\kappa_{2,t}^*\}$. Moreover, a prior knowledge of the value of κ_2 is often presupposed in the existing algorithms, which may be a major hindrance to their practical implementation. The quantity κ_2 appears for example in the design of the exploration bonuses of the UCB-MNL algorithm of [33]. Our algorithm does not rely on an a priori knowledge of κ_2. Our policy is based on optimistic parameter search instead of exploration bonuses, as used in [33]. 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 since 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, which 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 $B_p^d(0, W)$ 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 $B^d(0, W)$. 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 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 $S_t \subseteq [N]$, which can be chosen adversarially. 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 $x_{t,1}, ..., x_{t,|S_t|} \in \mathbb{R}^d$. For all $t \in [T]$, let $k_t = |S_t|$ and let $K \le N$ 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. In each period t,

1. The seller observes $x_{t,1}, ..., x_{t,k_t} \in \mathbb{R}^d$ and offers prices $p_{t,j}$ for all $j \in [k_t]$.

2. The customer observes the prices, then purchases one *single* product $j \in [k_t] \cup \{0\}$. The probability to purchase each product j is given by:

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

   where $\theta^*, \alpha^* \in \mathbb{R}^d$ are two model parameters, which are unknown to the policy maker. 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 $j = 0$).

3. The policy maker observes only the customer's purchase decision. 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 $u_{t,j} = \tilde{x}_{t,j}^\top \gamma^*$. Also, for each t and for a given estimator γ_t of the true parameter at time t, we denote by $q_{t,j}(\gamma_t, \vec{p_t})$ the estimated purchase probability for product j (obtained by replacing γ^* by γ_t in (1)).

We make the following assumptions, which are standard in the dynamic pricing literature:

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

Assumption 2.2. $||(\theta^*, \alpha^*)||_2 \le W$ 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 [23], we make the following assumption, which implies that the price-sensitivity of each product is not too close to zero. This assumption may be reasonable in practice: when the price of a product goes to infinity, its utility should decrease significantly.

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

Assumption 2.4. The upper bound K on the number of products in each set S_t is constant.

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 is to design a pricing policy so as to minimize the sellers' cumulative expected regret:

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

4

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.

Let $p_{max} := \frac{1+K\max(W,1)}{L} + \frac{1}{K}$. We show in Lemma B.2 that for the algorithm we propose, the prices posted at each time t satisfy $p_{t,j} \in [0, p_{max}]$ for all $j \in [k_t]$. Hence we can consider policies π that only post prices in $[0, p_{max}]$.

Finally, we define the following constant, which provides information about how much a feasible demand curve can deviate from the linear model (a smaller κ_1 implies a larger deviation from the linear model).

$$\kappa_1 = \min_{\gamma \in B^{2d}(0,W), t \geq 1, j \in \{1,\ldots k_t\}, \vec{p} \in [0,p_{\max}]^{k_t}} q_{t,j}(\gamma, \vec{p}) q_{t,0}(\gamma, \vec{p}).$$

Note that our pricing policy does not directly use the value of κ_1. However, this constant still appears in our regret bound (see Section 2.2 for more discussion).

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. Given our estimator γ_t of the true parameters at time t, we let ℓ_t denote the log loss at time t:

$$\ell_{t,1}(\gamma_t) = -\sum_{j=0}^{k_t} y_{t,j} \log(q_{t,j}(\gamma_t, \vec{p_t})).$$

For a time-dependent sequence of positive regularizers $\{\lambda_t\}_{t \geq 1}$ (the exact value of the regularizers used in our algorithm is given in Theorem 2.5), the estimator obtained before projection after conducting one step of our descent method is:

$$\hat{\gamma}_t = \gamma_{t-1} - \frac{1}{\mu} H_{t-1}^{-1} \nabla \ell_{t-1}(\gamma_{t-1}),$$

where $\mu = \frac{1}{2(1+(1+p_{\max})2\sqrt{6KW})}$ 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, $\ell_{t,1}$ is convex (as can be deduced from Lemma B.5 1.), hence $H_t \succ 0$ for all t.

Let B_t denote the set $B^{2d}(0, W) \cap \{(\theta, \alpha) \mid x_{t,j}^\top \alpha \geq L$ 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.

Running time: The two main computational steps of Algorithm 1 are calculating the inverse of H_t and projecting the parameter back in the set of feasible parameters relative to the norm H_t. The time complexity of the first step is $O(d^3)$, which is reasonable when d is not too large as in the setting we consider. The projection step can also be done efficiently by formulating the problem as a Quadratic Programming problem.

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

---

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 $\vec{p_t}$
Observe feedback $\vec{y_t}$
Set $\gamma_{t+1} = (\theta_{t+1}, \alpha_{t+1})$ as follows: $\gamma_{t+1} = \Pi_{B_{t+1}}^{H_t} \left(\gamma_t - \frac{1}{\mu} H_t^{-1} \nabla \ell_{t,1}(\gamma_t) \right)$

end for

---

in [21] 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 [20], which is the classic online analogue of the Newton method. The ONS method moves into the direction of $A_t^{-1} \nabla \ell_{t,1}$, 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 $\nabla \ell_{t,1}$, 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 parameter $1/\kappa_1$ (which corresponds to the exp-concavity parameter β in the literature) in the descent step, as is done by the ONS algorithm.

We are ready to present our regret bound.

Theorem 2.5. *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 2.5 is presented in Appendix B. We would like to point out that, even though our algorithm does not use the parameter $1/\kappa_1$, it still appears within the constant C.

Under the assumptions of [21] (single product dynamic pricing with adversarial contexts and constant price sensitivity), and assuming that the noise has a logistic distribution, we can show the following regret bound, which contrasts with the $O(\sqrt{T})$ upper bound established in [21].

Corollary 2.6. *If* $k_t = 1$ *for all* t, *and if the price coefficient is a known constant, then setting* $\lambda_1 = 1$, $\lambda_t = d \log(t)$ *for all* $t \geq 1$, *and letting* $\Delta p_{t,1} = 0$ *at each step, there is a constant C depending only on W such that the regret of Algorithm 1 with regularizers* $\{\lambda_t\}$ *and price shocks* $\{\Delta p_{t,1}\}$ *is bounded as:*

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

2.3 High level ideas and sketch of the proof.

We provide here the main ideas in the proof of Theorem 2.5. 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 2.7. *There exist constants C_1, C_2 depending only on W, L, K, such that:*

$$R^\pi(T) \leq \mathbb{E}\left[C_1 C_2 \sum_{t=1}^{T} \sum_{j=1}^{k_t} [((\alpha_t - \alpha^*)^T x_{t,j})^2 + ((\theta_t - \theta^*) x_{t,j})^2] + C_1 \sum_{t=1}^{T} ||\Delta \vec{p_t}||^2 \right]. \quad (2)$$

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 (2) 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 2.8. *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 2.8 is valid for any sequence of prices posted by the seller. However, it is not possible, in general, to derive directly Theorem 2.5 from Lemma 2.8 without an appropriate price experimentation scheme. 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 left-hand side in Lemma 2.8 would be zero and the bound provided would not be useful. However, adding random price shocks allows us to deviate from uninformative prices and to derive an upper bound on the first term of (2) based on Lemma 2.8. This concludes the proof of Theorem 2.5. If there is a single product ($k_t = 1$), and there is no price sensitivity (i.e. for all t, the coefficient in front of $p_{t,1}$ is a known constant, that we assume to be 1 without loss of generality), note that from Lemma 2.8, we get $\sum_t x_t^T(\theta_t - \theta_*) = O(d \log(T))$. Combining this with (2) and the choice of $\Delta p_{t,1} = 0$ for all t, we immediately obtain Corollary 2.6.

In [21], 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 2.8.

The proof of Lemma 2.8 is based on a lower and an upper bound on $\sum_{t=1}^T \ell_{t,1}(\gamma_t) - \ell_{t,1}(\gamma^*)$. 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 $\ell_{t,1}$ and is based on a Bernstein inequality for martingales difference sequences. This is similar to the inequality used in [21]. The main technical hindsight lies in the proof of the upper bound. It mimics the analysis of the Online Newton Step method presented in [20], but relies on the specific structure of the gradient and Hessian of $\ell_{t,1}$ for the MNL model. Moreover, it unically exploits the self concordant-like property of $\ell_{t,1}$. Let's first recall the definition of a self-concordant-like function.

Definition 2.9 (self-concordant-like functions [40])**.** 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 [40], we show in Appendix A the following property.

Proposition 2.10. $\ell_{t,1}$ *is self-concordant-like with* $M_f = (1 + p_{\max})\sqrt{6|S_t|}$.

We also 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 one *single* product $j \in S_t \cup \{0\}$. Each product j is purchased with probability:

$$q_{t,j}(S_t, \theta^*) = \begin{cases} \frac{e^{x_{t,j}^\top \theta^*}}{1 + \sum_{i \in S_t} e^{x_{t,i}^\top \theta^*}} & \text{if } j \in S_t \\ \frac{1}{1 + \sum_{i \in S_t} e^{x_{t,i}^\top \theta^*}} & \text{if } j = 0 \end{cases} \tag{3}$$

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 [33], 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 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 [33], we also introduce the following constant, which typically appears in connection to the link function in the generalized linear bandits ([18], [27]).

$$\kappa_2 := \min_{|S| \leq K, j \in S, t \geq 1} \min_{||\theta^*|| \leq W} q_{t,j}(S, \theta) q_{t,0}(S, \theta) > 0.$$

A smaller value of κ_2 can be interpreted as a bigger deviation from the linear model. As mentioned before, the regret bound of the dynamic assortment policies of [33] and [12] exhibit a harmful dependency in κ_2. Besides, an a priori knowledge of the value of κ_2 is presupposed. Our goal is to design a dynamic assortment policy which does not require prior knowledge of κ_2 and achieves a regret with a better dependency in κ_2. For all $t \in [T]$, let $\kappa_{2,t}^* = \sum_{j \in S_t^*} q_{t,j}(S_t^*, \theta^*) q_{t,0}(S_t^*, \theta^*)$. $\kappa_{2,t}^*$ 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 $\tilde{O}\left(\sqrt{\sum_{t=1}^{T} \kappa_{2,t}^*}\right)$ term. Note that we always have $\kappa_{2,t}^* \leq 1$ hence this is a strict improvement. As a result, a high level of non-linearity at time t induces a smaller $\kappa_{2,t}^*$ 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 $\{\lambda_t\}_{t=1}^T$ (the exact values are given in Theorem 3.4):

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

$\hat{\theta}_t$ satisfies the equation $\nabla \mathcal{L}_t^{\lambda_t}(\theta) = 0$, where the gradient of $\mathcal{L}_t^{\lambda_t}(\theta)$ is given by: $\nabla \mathcal{L}_t^{\lambda_t}(\theta) = \sum_{s=1}^{t-1} \sum_{j \in S_s} (q_{s,j}(S_s, \theta) - y_{s,j}) x_{s,j} + \lambda_t \hat{\theta}_t$. Following [2], for all $\delta \in [0, 1)$, we now define a confidence set $C_t(\delta)$ for θ^* as follows:

$$C_t(\delta) := \{\theta \in \Theta | \|g_t(\theta) - g_t(\hat{\theta}_t)\|_{H_t^{-1}(\theta)} \leq \gamma_t(\delta)\},$$

where $g_t(\theta) := \sum_{s=1}^{t-1} \sum_{j \in S_s} q_{s,j}(S_s, \theta) x_{s,j} + \lambda_t \theta$, where $H_t(\theta)$ is the Hessian of the regularized negative log likelihood evaluated at θ, i.e.,

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

and where

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

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 $\mathbb{P}(\forall t, \theta^* \in C_t(\delta)) \geq 1 - \delta$.*

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 (3) to the case where the model parameters corresponding to each product are uncorrelated. More precisely, we now consider some estimator $\tilde{\theta} \in \mathbb{R}^{d \times N}$ of the true parameters. For all $j \in [N]$, $\tilde{\theta}_j$ represents the parameter associated with product j. The estimated probability that item j is purchased at time t if assortment $S \subseteq [N]$ is offered is computed as:

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

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 $\tilde{\theta}_{t,j} = \arg\max_{\theta \in C_t(\delta)} x_{t,j}^\top \theta$. We offer the set S_t of the K items maximizing the optimistic expected revenue $\tilde{r}_t(S, \tilde{\theta})$:

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

Since we assumed all prices to be unit, this is equivalent to offering the K products with highest $x_{t,j}^\top \tilde{\theta}_{t,j}$. In the case of non-uniform revenues, our algorithm is still valid; however, an extra factor $1/\kappa_2$ would appear in the regret upper bound with our current analysis. We also note that Algorithm 2 is mainly of theoretical interest, since computing each $\tilde{\theta}_{t,j}$ remains computationally expensive.

We now present the our upper bound on the regret of our policy.

Theorem 3.4. *For $\delta = \frac{1}{K^2 T^2}$, $\lambda_1 = 1$, and $\lambda_t = d \log(tK)$ for all $t \geq 1$, the regret of Algorithm 2 satisfies, for some constants $\tilde{C}_1$ and $\tilde{C}_2$ that do not depend on d, K, T and that depend only polynomially on W:*

$$R(T) \leq \tilde{C}_1 K d \log(KT) \left(\sqrt{\sum_{t=1}^T \kappa_{2,t}^*} \right) + \frac{\tilde{C}_2 d^2 K^4}{\kappa_2} \log(KT)^2.$$

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 $\tilde{\theta}_{t,j} = \arg\max_{\theta \in C_t(\delta)} x_{t,j}^\top \theta$
        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.4 (Appendix C.2), we have the following concentration result on the optimistic parameter $\tilde{\theta}_{t,j}$ associated with each product $j \in [N]$:

$$\|\theta^* - \tilde{\theta}_{t,j}\|_{H_t(\theta^*)} \leq (1 + \sqrt{6K}W)\gamma_t(\delta). \tag{4}$$

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 $O(d^2 K^4 \log(KT)^2/\kappa_2)$. To show that the first term is of order $O\left(Kd\log(KT)\sqrt{\sum_{t=1}^T \kappa_{2,t}^*}\right)$, we first use the concentration result stated in (4), 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 $H_t(\theta^*) \succeq \kappa_2 \sum_{s=1}^T \sum_{j \in S_s} x_{s,j} x_{s,j}^\top$, 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 $1/\kappa_2$. 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 J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24, pages 2312–2320. Curran Associates, Inc., 2011.

[2] Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits, 2020.

[3] Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Mnl-bandit: A dynamic learning approach to assortment selection. *CoRR*, abs/1706.03880, 2017.

[4] Shipra Agrawal, Vashist Avadhanula, Vineet Goyal, and Assaf Zeevi. Thompson sampling for the mnl-bandit. *CoRR*, abs/1706.00977, 2017.

[5] Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Repeated contextual auctions with strategic buyers. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 622–630. Curran Associates, Inc., 2014.

[6] GahhYi Ban and N. Bora Keskin. Personalized dynamic pricing with machine learning. *SSRN Electronic Journal*, 01 2017.

[7] Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

[8] Josef Broder and Paat Rusmevichientong. Dynamic pricing under a general parametric choice model. *Operations Research*, 60:965–980, 08 2012.

[9] Felipe Caro and Jérémie Gallien. Dynamic assortment with demand learning for seasonal consumer goods. *Management Science*, 53(2):276–292, 2007.

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

[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, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 151–159. JMLR.org, 2013.

[12] Xi Chen, Yining Wang, and Yuan Zhou. Dynamic assortment optimization with changing contextual information, 2019.

[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. *Surveys in Operations Research and Management Science*, 20, 06 2015.

[15] Arnoud V. Den Boer and Bert Zwart. Mean square convergence rates for maximum quasi-likelihood estimator. *Stochastic Systems*, 4:375–403, 01 2014.

[16] Vivek F. Farias and Benjamin Van Roy. Dynamic pricing with a prior on market response. *Operations Research*, 58(1):16–29, 2010.

[17] Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits, 2020.

[18] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23, pages 586–594. Curran Associates, Inc., 2010.

[19] David A. Freedman. On Tail Probabilities for Martingales. *The Annals of Probability*, 3(1):100 – 118, 1975.

[20] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Mach. Learn.*, 69(2-3):169–192, 2007.

[21] Adel Javanmard. Perishability of data: Dynamic pricing under varying-coefficient models, 2017.

[22] Adel Javanmard and Hamid Nazerzadeh. Dynamic pricing in high-dimensions, 2018.

[23] Adel Javanmard, Hamid Nazerzadeh, and Simeng Shao. Multi-product dynamic pricing in high-dimensions with heterogeneous price sensitivity, 2019.

[24] 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, page 594, USA, 2003. IEEE Computer Society.

[25] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. *CoRR*, abs/1410.0949, 2014.

[26] Renato Paes Leme and Jon Schneider. Contextual search via intrinsic volumes. In Mikkel Thorup, editor, *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 268–282. IEEE Computer Society, 2018.

[27] Lihong Li, Yu Lu, and Dengyong Zhou. Provable optimal algorithms for generalized linear contextual bandits. *CoRR*, abs/1703.00048, 2017.

[28] Allen Liu, Renato Paes Leme, and Jon Schneider. Optimal contextual pricing and extensions. In Dániel Marx, editor, *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021*, pages 1059–1078. SIAM, 2021.

[29] Ilan Lobel, Renato Paes Leme, and Adrian Vladu. Multidimensional binary search for contextual decision-making. *Oper. Res.*, 66(5):1346–1361, 2018.

[30] D. Mcfadden. Modelling the choice of residential location. *Transportation Research Record*, 1977.

[31] Mila Nambiar, David Simchi-Levi, and He Wang. Dynamic learning and pricing with model misspecification. *Management Science*, 65(11):4980–5000, 2019.

[32] Min-hwan Oh and Garud Iyengar. Thompson sampling for multinomial logit contextual bandits. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 3151–3161. Curran Associates, Inc., 2019.

[33] Min-hwan Oh and Garud Iyengar. Multinomial logit contextual bandits: Provable optimality and practicality. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 9205–9213. AAAI Press, 2021.

[34] R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975.

[35] Sheng Qiang and Mohsen Bayati. Dynamic pricing with demand covariates, 2016.

[36] Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In Mohammed Javeed Zaki, Zoran Obradovic, Pang-Ning Tan, Arindam Banerjee, Chandrika Kamath, and Srinivasan Parthasarathy, editors, *Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 24-26, 2014*, pages 461–469. SIAM, 2014.

[37] Idan Rejwan and Yishay Mansour. Combinatorial bandits with full-bandit feedback: Sample complexity and regret minimization. *CoRR*, abs/1905.12624, 2019.

[38] Paat Rusmevichientong, Zuo-Jun Max Shen, and David B. Shmoys. Dynamic assortment optimization with a multinomial logit choice model and capacity constraint. *Operations Research*, 58(6):1666–1680, 2010.

[39] Denis Sauré and Assaf Zeevi. Optimal dynamic assortment planning with demand learning. *Manufacturing & Service Operations Management*, 15(3):387–404, 2013.

[40] Quoc Tran-Dinh, Yen-Huan Li, and Volkan Cevher. Composite convex minimization involving self-concordant-like cost functions, 2018.

[41] Yining Wang, Xi Chen, and Yuan Zhou. Near-optimal policies for dynamic multinomial logit assortment selection models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 3101–3110. Curran Associates, Inc., 2018.

[42] Jianyu Xu and Yu-Xiang Wang. Logarithmic regret in feature-based dynamic pricing. *CoRR*, abs/2102.10221, 2021.

[43] Lijun Zhang, Tianbao Yang, Rong Jin, Yichi Xiao, and Zhi-Hua Zhou. 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]

A Self-concordant properties

In this section, we derive some useful properties of self-concordant-like functions which will be used in the subsequent regret proofs. We first remind the reader of the definition of self-concordant-like functions.

Definition A.1 (self-concordant-like functions [40]). A convex function $f \in C^3(\mathbb{R}^n)$ is called a self-concordant-like function with constant M_f 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 $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^n$.

Our results essentially rely on the following property of self-concordant-like functions.

Proposition A.2 (Theorem 4 in [40]). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a M_f- self-concordant-like function and let $x, y \in dom(f)$, then:*

$$e^{-M_f\|y-x\|_2}\nabla^2 f(x) \preceq \nabla^2 f(y)$$

Now, for any $X > 0, m \geq 0, d' > 0, \{z_j\}_{j\in[m]} \in \mathbb{R}^{d'}$ satisfying $\|z_j\|_2 \leq X$ for all $j \in [m]$, and $\{y_j\}_{j\in[m]} \in \mathbb{R}$, we consider the function $\ell : B^{d'}(0, W) \longrightarrow \mathbb{R}$ defined as:

$$\ell(\beta) = -\sum_{i=1}^m y_i z_i^T \beta + \log\left(1 + \sum_{j=1}^m e^{z_j^T \beta}\right). \tag{5}$$

Note that the negative log likelihood $\ell_{t,1}$ and $\ell_{t,2}$ can be written as

$$\ell_{t,1}(\gamma) = -\sum_{i=0}^{|S_t|} y_{t,i} \log(q_{t,i}(\gamma, \vec{p_t}))$$

$$= -\sum_{i=1}^{|S_t|} y_{t,i}\tilde{x}_{t,i}^\top\gamma + \left(\sum_{i=0}^{|S_t|} y_{t,i}\right) \log\left(1 + \sum_{j=0}^{|S_t|} e^{\tilde{x}_{t,j}^\top\gamma}\right)$$

$$= -\sum_{i=1}^{|S_t|} y_{t,i}\tilde{x}_{t,i}^\top\gamma + \log\left(1 + \sum_{j=1}^{|S_t|} e^{\tilde{x}_{t,j}^\top\gamma}\right), \qquad \text{(since } \sum_{i=0}^{|S_t|} y_{t,i} = 1\text{)}$$

with $\|\tilde{x}_{t,j}\|_2 = \|[x_{t,j}, -p_{t,j}x_{t,j}]\|_2 \leq \sqrt{1 + p_{max}^2}\|x_{t,j}\|_2 \leq 1 + p_{max}$.
Similarly,

$$\ell_{t,2}(\theta) = -\sum_{i\in S_t} y_{t,i} \log(q_{t,i}(S_t, \theta)) = -\sum_{i\in S_t} y_{t,i}x_{t,i}^\top\theta + \log\left(1 + \sum_{i\in S_t} e^{x_{t,i}^\top\theta}\right),$$

with $\|x_{t,j}\|_2 \leq 1$.

Hence for all $t \geq 0$, $\ell_{t,1}$ and $\ell_{t,2}$ are of the form given in 5 with ($z_j = \tilde{x}_{t,j}$ and $X = 1 + p_{max}$) and ($z_j = x_{t,j}$ and $X = 1$), respectively. In particular, $\ell_{t,1}, \ell_{t,2}$ satisfy all properties stated below with the corresponding constants.

Proposition A.3. *The function ℓ is self-concordant-like with $M_\ell = X\sqrt{6m}$.*

Proof. By following the proof of Lemma 4 in [40], we obtain that the for all $a, \mu \in \mathbb{R}^n$, the function $\psi(s) := \log\left(\sum_{i=0}^n e^{a_i s + \mu_i}\right)$ satisfies the inequality:

$$|\psi'''(s)| \leq \sqrt{6}\|a\|_2\psi''(s). \tag{6}$$

Now, let $f(\beta) := \log\left(1 + \sum_{j=1}^m e^{z_{t,j}^T \beta}\right)$.
Let $d \in \mathbb{R}^d$ and let $\phi(s) := f(\beta + sd) = \log\left(\sum_{i=0}^m e^{a_i s + \mu_i}\right)$, where $\mu_i = \beta^T z_i$, $a_i = d^T z_i$ and

$\mu_0, a_0 = 0$. Then using (6), we obtain:

$$|\phi'''(s)| \leq \sqrt{6}\|a\|_2 \phi''(s) = \sqrt{6}\sqrt{\sum_{i=1}^{m}(d^T z_i)^2}\phi''(s)$$

$$\leq \sqrt{6}\sqrt{\sum_{i=1}^{m}\|d\|_2^2\|z_i\|_2^2}\phi''(s) \leq X\sqrt{6m}\|d\|_2\phi''(s),$$

where the second inequality comes from Cauchy-Schwartz and the last one is since $||z_j||_2 \leq X$ for all $j \in [m]$. This shows that f is self-concordant-like with constant $M_f = X\sqrt{6m}$.

Since ℓ is the sum of f and a linear operator (for which the third derivatives are zero), we obtain that ℓ is self-concordant-like with constant $X\sqrt{6m}$.

$\square$

Proposition A.4. *The hessian of ℓ satisfies, for all $u \in \mathbb{R}^d, \beta_1, \beta_2 \in B^{d'}(0, W)$:*

$$u^T\left(\int_0^1\int_0^1 z\nabla^2\ell(\beta_1 + zw(\beta_2 - \beta_1))dzdw\right)u \geq \frac{1}{2(1+X2\sqrt{6m}W)}u^T\nabla^2\ell(\beta_1)u$$

Proof. From Proposition A.2, we obtain:

$$u^T\int_0^1\int_0^1 z\nabla^2\ell(\beta_1 + zw(\beta_2 - \beta_1))dzdwu \geq u^T\nabla^2\ell(\beta_1)u\int_0^1\int_0^1 e^{-M_\ell||zw(\beta_2-\beta_1)||_2}zdzdw$$

Besides,

$$\int_0^1\int_0^1 e^{-M_\ell||zw(\beta_2-\beta_1)||_2}zdzdw = \int_0^1 z\left(\frac{1 - e^{-M_\ell z||\beta_1-\beta_2||_2}}{M_\ell z||\beta_1 - \beta_2||_2}\right)dz$$

Noting that $\frac{1-e^{-x}}{x} \geq \frac{1}{1+x}$ for $x > 0$ (derived from $e^x \geq (1 + x)$) we obtain:

$$\int_0^1\int_0^1 e^{-M_\ell||zw(\beta_2-\beta_1)||_2}zdzdw \geq \int_0^1 z\left(\frac{1}{1+M_\ell z||\beta_1-\beta_2||_2}\right)dz$$

$$\geq \int_0^1 z\left(\frac{1}{1+M_\ell||\beta_1-\beta_2||_2}\right)dz$$

$$= \frac{1}{2(1+M_\ell||\beta_1-\beta_2||_2)}$$

$$\geq \frac{1}{2(1+X2\sqrt{6m}W)}$$

where the last inequality comes from the fact that $\beta_1, \beta_2 \in B^{d'}(0, W)$.

$\square$

Proposition A.5. *The hessian of ℓ satisfies, for all $u \in \mathbb{R}^d, \beta_1, \beta_2 \in B^2(0, W)$:*

$$u^T\int_0^1 \nabla^2\ell_t(\beta_1 + z(\beta_2 - \beta_1))dzu \geq \frac{1}{(1+X2\sqrt{6m}W)}u^T\nabla^2\ell_t(\beta_1)u$$

Proof. From Proposition A.2 and using the inequality $\frac{1-e^{-x}}{x} \geq \frac{1}{1+x}$ for $x \geq 0$, we obtain:

$$u^T\int_0^1\nabla^2\ell_t(\beta_1 + z(\beta_2 - \beta_1))dzu \geq u^T\nabla^2\ell(\beta_1)u\int_0^1 e^{-M_\ell||z(\beta_2-\beta_1)||_2}dz$$

$$\geq u^T\nabla^2\ell(\beta_1)u\left(\frac{1 - e^{-M_\ell||\beta_1-\beta_2||_2}}{M_\ell||\beta_1 - \beta_2||_2}\right)$$

$$\geq u^T\nabla^2\ell(\beta_1)u\left(\frac{1}{1+M_\ell||\beta_1-\beta_2||_2}\right)$$

$$\geq u^T\nabla^2\ell(\beta_1)u\frac{1}{(1+X2\sqrt{6m}W)}$$

where the last inequality comes from the fact that $\beta_1, \beta_2 \in B^{d'}(0, W)$.

$\square$

B Dynamic pricing technical proofs

B.1 Proof of Theorem 2.5

We provide here the full proof of the regret upper bound given in Theorem 2.5. We first state a few technical lemmas.

To begin with, we characterize the greedy price corresponding to the current context and some given estimators of the parameters. The proposition below can be found in [23].

Proposition B.1. *([23] Proposition 3.1) If the true utility model parameters are θ, α, then the optimal prices $\{p_{t,i}\}_{i \in [k_t]}$ are as follows. For all $t \geq 1$ and product $i \in [k_t]$,*

$$p^*_{t,i} = \frac{1}{x^\top_{t,i}\alpha} + B^0_t \equiv g(X_t\alpha, X_t\theta)_i \,, \tag{7}$$

where B^0_t is the unique fixed point B of the following equation:

$$B = \sum_{i=1}^{k_t} \frac{1}{x^\top_{t,i}\alpha} e^{-(1+x^\top_{t,i}\alpha B)} e^{x^\top_{t,i}\theta} \,. \tag{8}$$

The following lemma then guarantees the boundedness of the prices posted by Algorithm 1 as well as the boundedness of the resulting product utilities. The proof is deferred to Appendix B.2

Lemma B.2. *Let $\vec{p_t} = g(X_t\alpha_t, X_t\theta_t) + \vec{\Delta p_t}$ be the vector of prices posted at time t and let $p_{max} := \frac{1+K\max(W,1)}{L} + \frac{1}{W}$. Then, for all $j \in [k_t]$, we have $g(X_t\alpha_t, X_t\theta_t)_j \in [0, p_{max}]$ and $p_{t,j} \in [0, p_{max}]$.*

It follows that the product utilities satisfy $\tilde{x}^\top_{t,j}\gamma_t \leq W(1+p_{max}) \equiv M$ for all $t \geq 1, j \in [k_t]$.

We now borrow two lemmas from [23]. First, let $h_t(\vec{p}) = \sum_{j=1}^{k_t} q_{t,j}(\gamma^*, \vec{p})p_j$ be the expected revenue at time t after posting prices $\vec{p}$. The following lemma shows the boundedness of $||\nabla^2 h_t(\vec{p})||_2$.

Lemma B.3. *([23], proof of Lemma 5.4) There is a constant C_1 depending only on W, L and K such that the operator norm $||\nabla^2 h_t(\vec{p})||_2$ satisfies $||\nabla^2 h_t(\vec{p})||_2 \leq C_1$ for all $t \geq 0$ and $\vec{p} \in [0, p_{\max}]^{k_t}$.*

The next lemma gives an upper bound in the difference of myopic prices for different values of the estimators.

Lemma B.4 ([23], Lemma 5.2). *Let $\gamma_1 = (\theta_1, \alpha_1) \in B^{2d}(0, W)$ and $\gamma_2 = (\theta_2, \alpha_2) \in B^{2d}(0, W)$ be such that for all $j \in [k_t]$, $x^\top_{t,j}\alpha_i \geq L$. Then there exists a constant C_2 depending only on W, L, K such that:*

$$||g(X_t\alpha_1, X_t\theta_1) - g(X_t\alpha_2, X_t\theta_2)||^2_2 \leq C_2(||X_t(\alpha_t - \alpha^*)||^2_2 + ||X_t(\theta_t - \theta^*)||^2_2)$$

We are now ready to present the proof of Theorem 2.5.

Proof of Theorem 2.5.

Recall that $h_t(\vec{p}) = \sum_{j=1}^{k_t} q_{t,j}(\gamma^*, \vec{p})p_j$ is the expected revenue at time t after posting prices $\vec{p}$. Since for all $t \geq 0$, $\vec{p^*_t} = \arg\max h_t(\vec{p})$, we have that $\nabla h_t(\vec{p^*_t}) = 0$.

Hence, by doing a Taylor expansion of h_t at $\vec{p^*_t}$, we obtain that the regret is bounded as follows:

$$R^\pi(T)$$

$$= \mathbb{E}\left[\sum_{t=1}^T \sum_{j=1}^{k_t} q_{t,j}(\gamma^*, \vec{p^*_t})p^*_{t,j} - q_{t,j}(\gamma^*, \vec{p_t})p_{t,j}\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^T h(\vec{p^*_t}) - h(\vec{p_t})\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T}(\nabla h(\vec{p_t^*})(\vec{p_t} - \vec{p_t^*}) + \frac{1}{2}(\vec{p_t} - \vec{p_t^*})^\top \nabla^2 h(\vec{\tilde{p}})(\vec{p_t} - \vec{p_t^*}))\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T}\frac{1}{2}(\vec{p_t} - \vec{p_t^*})^\top \nabla^2 h(\vec{\tilde{p}})(\vec{p_t} - \vec{p_t^*})\right] \qquad (\nabla h_t(\vec{p_t^*}) = 0)$$

$$\leq \mathbb{E}\left[\frac{C_1}{2}\sum_{t=1}^{T}||\vec{p_t} - \vec{p_t^*}||_2^2\right]. \qquad \text{(Lemma B.3)}$$

$$= \mathbb{E}\left[\frac{C_1}{2}\sum_{t=1}^{T}||g(X_t\alpha_t, X_t\theta_t) + \vec{\Delta p_t} - g(X_t\alpha^*, X_t\theta^*)||_2^2\right]$$

$$\leq \mathbb{E}\left[\frac{C_1}{2}\sum_{t=1}^{T}||g(X_t\alpha_t, X_t\theta_t) - g(X_t\alpha^*, X_t\theta^*)||_2^2 + \frac{C_1}{2}\sum_{t=1}^{T}||\vec{\Delta p_t}||_2^2\right]$$

$$\leq \mathbb{E}\left[\frac{C_1 C_2}{2}\sum_{t=1}^{T}\sum_{j=1}^{k_t}((\alpha_t - \alpha^*)^T x_{t,j})^2 + ((\theta_t - \theta^*)^T x_{t,j})^2\right] + \mathbb{E}\left[\frac{C_1}{2}\sum_{t=1}^{T}||\vec{\Delta p_t}||_2^2\right] \quad \text{(Lemma B.4)}$$

$$\leq \mathbb{E}\left[\frac{C_1 C_2}{2}\sum_{t=1}^{T}\sum_{j=1}^{k_t}((\alpha_t - \alpha^*)^T x_{t,j})^2 + ((\theta_t - \theta^*)^T x_{t,j})^2\right] + C_1\sum_{t=1}^{T}\frac{K}{W^2\sqrt{t}}$$

$$\leq \mathbb{E}\left[\frac{C_1 C_2}{2}\sum_{t=1}^{T}\sum_{j=1}^{k_t}((\alpha_t - \alpha^*)^T x_{t,j})^2 + ((\theta_t - \theta^*)^T x_{t,j})^2\right] + \frac{2C_1 K\sqrt{T}}{W^2}, \qquad (9)$$

where in the third equality, $\vec{\tilde{p}}$ is a linear combination of $\vec{p}$ and $\vec{p}^*$.

Using Lemma 2.8, we obtain that with probability at least $1 - \frac{\log_2(T)}{T^2}$:

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

Besides, since $\alpha_t, \alpha^*, \theta_t, \theta^*, X_t, \vec{p_t}$ are bounded by constants depending only on W, L, K, there is a constant C_0 depending only on W, L, K such that $\sum_{t=1}^{T}\sum_{j=1}^{k_t}((\theta_t - \theta^*)^T x_{t,j} - (\alpha_t - \alpha^*)^T x_{t,j} p_{t,j})^2 \leq C_0 KT$.

Let $\mathbb{A}_T$ denote the event that $\sum_{t=1}^{T}\sum_{j=1}^{k_t}((\theta_t - \theta^*)^T x_{t,j} - (\alpha_t - \alpha^*)^T x_{t,j} p_{t,j})^2 \leq C(T)$.

We have that:

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{k_t}((\theta_t - \theta^*)^T x_{t,j} - (\alpha_t - \alpha^*)^T x_{t,j} p_{t,j})^2\right]$$

$$= \mathbb{E}\left[\mathbf{1}_{\mathbb{A}_T}\sum_{t=1}^{T}\sum_{j=1}^{k_t}((\theta_t - \theta^*)^T x_{t,j} - (\alpha_t - \alpha^*)^T x_{t,j} p_{t,j})^2\right]$$

$$+ \mathbb{E}\left[\mathbf{1}_{\overline{\mathbb{A}_T}}\sum_{t=1}^{T}\sum_{j=1}^{k_t}((\theta_t - \theta^*)^T x_{t,j} - (\alpha_t - \alpha^*)^T x_{t,j} p_{t,j})^2\right]$$

$$\leq C(T) + C_0 KT \times \frac{\log_2(T)}{T^2} \equiv \tilde{C}(T)$$

This implies

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{j=1}^{k_t}((\theta_t - \theta^*)^T x_{t,j} - (\alpha_t - \alpha^*)^T x_{t,j}(g(X_t\alpha_t, X_t\theta_t)_j + \Delta p_{t,j}))^2\right] \leq \tilde{C}(T).$$

18

Given that for all t, j, $\Delta p_{t,j}$ has zero mean, developing the above sum implies the following inequalities:

$$\mathbb{E} \sum_{t=1}^{T} \sum_{j=1}^{k_t} ((x_{t,j}^{\top}(\theta_t - \theta^*)) - g(X_t\alpha_t, X_t\theta_t)_j(x_{t,j}^{\top}(\alpha_t - \alpha^*)))^2 \leq \tilde{C}(T) \qquad (10)$$

and

$$\mathbb{E} \sum_{t=1}^{T} \sum_{j=1}^{k_t} ((\alpha_t - \alpha^*)^T x_{t,j})^2 \Delta p_{t,j}^2 \leq \tilde{C}(T). \qquad (11)$$

We now use the two above inequalities to provide bounds on $\mathbb{E} \sum_{t=1}^{T} \sum_{j=1}^{k_t} ((\alpha_t - \alpha^*)^T x_{t,j})^2$ and on $\mathbb{E} \sum_{t=1}^{T} \sum_{j=1}^{k_t} ((\theta_t - \theta^*)^T x_{t,j})^2$.

First, since for all $t \leq T$, $\Delta p_{t,j}^2 \geq \frac{1}{W^2\sqrt{T}}$, it follows that

$$\mathbb{E} \sum_{t=1}^{T} \sum_{j=1}^{k_t} ((\alpha_t - \alpha^*)^T x_{t,j})^2 \leq W^2 \tilde{C}(T)\sqrt{T}. \qquad (12)$$

Then, using the Cauchy-Schwartz inequality and noting that $g(X_t\alpha_t, X_t\theta_t)_j \leq p_{\max}$ by Lemma B.2, we have

$$\mathbb{E} \left(\sum_{t=1}^{T} \sum_{j=1}^{k_t} 2g(X_t\alpha_t, X_t\theta_t)_j x_{t,j}^{\top}(\theta_t - \theta^*)x_{t,j}^{\top}(\alpha_t - \alpha^*) \right)$$

$$\leq 2p_{\max}\mathbb{E} \left(\sum_{t=1}^{T} \sum_{j=1}^{k_t} x_{t,j}^{\top}(\theta_t - \theta^*)x_{t,j}^{\top}(\alpha_t - \alpha^*) \right)$$

$$\leq 2p_{max} \sqrt{\mathbb{E} \left[\sum_{t=1}^{T} \sum_{j=1}^{k_t} x_{t,j}^{\top}(\alpha_t - \alpha^*)^2 \right]} \sqrt{\mathbb{E} \left[\sum_{t=1}^{T} \sum_{j=1}^{k_t} x_{t,j}^{\top}(\theta_t - \theta^*)^2 \right]}$$

$$\leq 2p_{max}T^{1/4}\sqrt{W^2\tilde{C}(T)} \sqrt{\mathbb{E} \left[\sum_{t=1}^{T} \sum_{j=1}^{k_t} x_{t,j}^{\top}(\theta_t - \theta^*)^2 \right]}.$$

Hence, first noting that (10) implies that

$$\mathbb{E} \left[\sum_{t=1}^{T} \sum_{j=1}^{k_t} x_{t,j}^{\top}(\theta_t - \theta^*)^2 \right] - \mathbb{E} \left(\sum_{t=1}^{T} \sum_{j=1}^{k_t} 2g(X_t\alpha_t, X_t\theta_t)_j x_{t,j}^{\top}(\theta_t - \theta^*)x_{t,j}^{\top}(\alpha_t - \alpha^*) \right) \leq \tilde{C}(T),$$

we obtain

$$\mathbb{E} \left[\sum_{t=1}^{T} \sum_{j=1}^{k_t} x_{t,j}^{\top}(\theta_t - \theta^*)^2 \right] - 2p_{max}T^{1/4}\sqrt{W^2\tilde{C}(T)} \sqrt{\mathbb{E} \left[\sum_{t=1}^{T} \sum_{j=1}^{k_t} x_{t,j}^{\top}(\theta_t - \theta^*)^2 \right]} \leq \tilde{C}(T). \qquad (13)$$

Let $A := \mathbb{E} \left[\sum_{t=1}^{T} \sum_{j=1}^{k_t} x_{t,j}^{\top}(\theta_t - \theta^*)^2 \right]$. Consider the two following cases:

- $2p_{max}T^{1/4}\sqrt{W^2\tilde{C}(T)}\sqrt{A} \leq \frac{A}{2}$. Then, from (13), we obtain that $A \leq 2\tilde{C}(T)$.

- $2p_{max}T^{1/4}\sqrt{W^2\tilde{C}(T)}\sqrt{A} > \frac{A}{2}$. Then $A \leq 16p_{max}^2\sqrt{T}W^2\tilde{C}(T)$.

So $A \leq 16 \max(2, p_{max})^2 \sqrt{T} W^2 \tilde{C}(T)$.

Coming back to the upper bound on the regret given by equation (9), we obtain that the total regret is bounded as follows:

$$R^\pi(T) \leq C_1 C_2 \left[\sqrt{T} W^2 \tilde{C}(T)(1 + 16 \max(2, p_{max})^2) + W^2 \tilde{C}(T)\sqrt{T} \right] + \frac{2 C_1 K \sqrt{T}}{W^2}.$$

Using the definition of $\tilde{C}(T) = \tilde{C} d \log(T)$, we conclude that for some constant $C > 0$ depending on W,K,L,

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

$\square$

B.2 Proof of the main lemmas

Throughout this section, we will use the following closed form expressions of the gradient and hessian of the loss $\ell_{t,1}$.

Remember that $y_{t,j}$ is the binary variable which indicates whether the customer has purchased product j at time t. The loss at time t can be expressed as:

$$\ell_{t,1}(\gamma) = \sum_{j=0}^{k_t} -y_{t,j} \log \left(q_{t,j}(\gamma, \vec{p_t}) \right) = \log \left(1 + \sum_{j=1}^{k_t} e^{\tilde{x}_{t,j}^\top \gamma} \right) - \sum_{j=1}^{k_t} y_{t,i} \tilde{x}_{t,i}^\top \gamma$$

We can then write the gradient and hessian of $\ell_{t,1}$ as:

$$\nabla \ell_{t,1}(\gamma) = \sum_{j=1}^{k_t} (q_{t,j}(\gamma, \vec{p_t}) - y_{t,j}) \tilde{x}_{t,j} \tag{14}$$

and

$$\nabla^2 \ell_{t,1}(\gamma) = \frac{\sum_{j=1}^{k_t} e^{\tilde{x}_{t,j}^\top \gamma} \tilde{x}_{t,j} \tilde{x}_{t,j}^T (1 + \sum_{j=1}^{k_t} e^{\tilde{x}_{t,j}^\top \gamma}) - (\sum_{j=1}^{k_t} e^{\tilde{x}_{t,j}^\top \gamma} \tilde{x}_{t,j})(\sum_{j=1}^{k_t} e^{\tilde{x}_{t,j}^\top \gamma} \tilde{x}_{t,j}^T)}{(1 + \sum_{j=1}^{k_t} e^{\tilde{x}_{t,j}^\top \gamma})^2}$$

$$= \sum_{j=1}^{k_t} q_{t,j}(\gamma, \vec{p_t}) \tilde{x}_{t,j} \tilde{x}_{t,j}^T - \sum_{j=1}^{k_t} \sum_{i=1}^{k_t} q_{t,j}(\gamma, \vec{p_t}) q_{t,i}(\gamma, \vec{p_t}) \tilde{x}_{t,j} \tilde{x}_{t,i}^T. \tag{15}$$

Lemma B.5. *All following properties hold for all $t \geq 0$ and $\gamma \in \mathbb{R}^d$:*

1. $\nabla^2 \ell_{t,1}(\gamma) \succeq \kappa_1 \sum_{j=1}^{k_t} \tilde{x}_{t,j} \tilde{x}_{t,j}^T$,

2. $\nabla^2 \ell_{t,1}(\gamma) \preceq 2 \sum_{j=1}^{k_t} \tilde{x}_{t,j} \tilde{x}_{t,j}^T$,

3. $\nabla \ell_{t,1}(\gamma) \nabla \ell_{t,1}(\gamma)^T \preceq 2 \sum_{j=1}^{k_t} \tilde{x}_{t,j} \tilde{x}_{t,j}^T$,

4. $\nabla \ell_{t,1}(\gamma)^T \nabla \ell_{t,1}(\gamma) \leq 2 \sum_{j=1}^{k_t} \tilde{x}_{t,j}^T \tilde{x}_{t,j}$.

Proof. First note that for all $i, j \in [k_t]$:

$$(\tilde{x}_{t,j} - \tilde{x}_{t,i})^T (\tilde{x}_{t,j} - \tilde{x}_{t,i}) = \tilde{x}_{t,j} \tilde{x}_{t,j}^T + \tilde{x}_{t,i} \tilde{x}_{t,i}^T - \tilde{x}_{t,j} \tilde{x}_{t,i}^T - \tilde{x}_{t,i} \tilde{x}_{t,j}^T \succeq 0$$

$$(\tilde{x}_{t,j} + \tilde{x}_{t,i})^T (\tilde{x}_{t,j} + \tilde{x}_{t,i}) = \tilde{x}_{t,j} \tilde{x}_{t,j}^T + \tilde{x}_{t,i} \tilde{x}_{t,i}^T + \tilde{x}_{t,j} \tilde{x}_{t,i}^T + \tilde{x}_{t,i} \tilde{x}_{t,j}^T \succeq 0$$

Hence for all $\alpha \in \mathbb{R}$,

$$|\alpha|(\tilde{x}_{t,j} \tilde{x}_{t,j}^T + \tilde{x}_{t,i} \tilde{x}_{t,i}^T) \succeq \alpha(\tilde{x}_{t,j} \tilde{x}_{t,i}^T + \tilde{x}_{t,i} \tilde{x}_{t,j}^T). \tag{16}$$

20

Similarly, for all $\alpha \in \mathbb{R}$,

$$|\alpha|(\tilde{x}_{t,j}^T \tilde{x}_{t,j} + \tilde{x}_{t,i}^T \tilde{x}_{t,i}) \geq \alpha(\tilde{x}_{t,j}^T \tilde{x}_{t,i} + \tilde{x}_{t,i}^T \tilde{x}_{t,j}). \tag{17}$$

It follows that:

1.
$$
\begin{aligned}
\nabla^2 \ell_{t,1}(\gamma) &= \sum_{j=1}^{k_t} q_{t,j}(\gamma, \vec{p_t})\tilde{x}_{t,j}\tilde{x}_{t,j}^T - \sum_{j=1}^{k_t}\sum_{i=1}^{k_t} q_{t,j}(\gamma, \vec{p_t})q_{t,i}(\gamma, \vec{p_t})\tilde{x}_{t,j}\tilde{x}_{t,i}^T \\
&= \sum_{j=1}^{k_t} q_{t,j}(\gamma, \vec{p_t})\tilde{x}_{t,j}\tilde{x}_{t,j}^T - \frac{1}{2}\sum_{j=1}^{k_t}\sum_{i=1}^{k_t} q_{t,j}(\gamma, \vec{p_t})q_{t,i}(\gamma, \vec{p_t})(\tilde{x}_{t,j}\tilde{x}_{t,i}^T + \tilde{x}_{t,i}\tilde{x}_{t,j}^T) \\
&\succeq \sum_{j=1}^{k_t} q_{t,j}(\gamma, \vec{p_t})\tilde{x}_{t,j}\tilde{x}_{t,j}^T - \frac{1}{2}\sum_{j=1}^{k_t}\sum_{i=1}^{k_t} q_{t,j}(\gamma, \vec{p_t})q_{t,i}(\gamma, \vec{p_t})(\tilde{x}_{t,i}\tilde{x}_{t,i}^T + \tilde{x}_{t,j}\tilde{x}_{t,j}^T) \\
&= \sum_{j=1}^{k_t} q_{t,j}(\gamma, \vec{p_t})\left(1 - \sum_{i=1}^{k_t} q_{t,j}(\gamma, \vec{p_t})\right)\tilde{x}_{t,j}\tilde{x}_{t,j}^T \\
&= \sum_{j=1}^{k_t} q_{t,j}(\gamma, \vec{p_t})q_{t,0}(\gamma, \vec{p_t})\tilde{x}_{t,j}\tilde{x}_{t,j}^T \\
&\succeq \kappa_1 \sum_{j=1}^{k_t} \tilde{x}_{t,j}\tilde{x}_{t,j}^T.
\end{aligned}
$$

2.
$$
\begin{aligned}
\nabla^2 \ell_{t,1}(\gamma_t) &= \sum_{j=1}^{k_t} q_{t,j}(\gamma, \vec{p_t})\tilde{x}_{t,j}\tilde{x}_{t,j}^T - \sum_{j=1}^{k_t}\sum_{i=1}^{k_t} q_{t,j}(\gamma, \vec{p_t})q_{t,i}(\gamma, \vec{p_t})\tilde{x}_{t,j}\tilde{x}_{t,i}^T \\
&\preceq \sum_{j=1}^{k_t} q_{t,j}(\gamma, \vec{p_t})\tilde{x}_{t,j}\tilde{x}_{t,j}^T + \frac{1}{2}\sum_{j=1}^{k_t}\sum_{i=1}^{k_t} q_{t,j}(\gamma, \vec{p_t})q_{t,i}(\gamma, \vec{p_t})(\tilde{x}_{t,i}\tilde{x}_{t,i}^T + \tilde{x}_{t,j}\tilde{x}_{t,j}^T) \\
&= \sum_{j=1}^{k_t} q_{t,j}(\gamma, \vec{p_t})\left(1 + \sum_{i=1}^{k_t} q_{t,j}(\gamma, \vec{p_t})\right)\tilde{x}_{t,j}\tilde{x}_{t,j}^T \\
&\preceq 2\sum_{j=1}^{k_t} \tilde{x}_{t,j}\tilde{x}_{t,j}^T.
\end{aligned}
$$

3.
$$
\begin{aligned}
\nabla \ell_{t,1}(\gamma_t)\nabla \ell_{t,1}(\gamma_t)^T &= \sum_{j=1}^{k_t}\sum_{i=1}^{k_t}(q_{t,j}(\gamma, \vec{p_t}) - y_{t,j})(q_{t,i}(\gamma, \vec{p_t}) - y_{t,i})\tilde{x}_{t,j}\tilde{x}_{t,i}^T \\
&= \frac{1}{2}\sum_{j=1}^{k_t}\sum_{i=1}^{k_t}(q_{t,j}(\gamma, \vec{p_t}) - y_{t,j})(q_{t,i}(\gamma, \vec{p_t}) - y_{t,i})(\tilde{x}_{t,j}\tilde{x}_{t,i}^T + \tilde{x}_{t,i}\tilde{x}_{t,j}^T) \\
&\preceq \frac{1}{2}\sum_{j=1}^{k_t}\sum_{i=1}^{k_t}|q_{t,j}(\gamma, \vec{p_t}) - y_{t,j}||q_{t,i}(\gamma, \vec{p_t}) - y_{t,i}|(\tilde{x}_{t,j}\tilde{x}_{t,j}^T + \tilde{x}_{t,i}\tilde{x}_{t,i}^T)
\end{aligned}
$$
$$\text{by (16)}$$
$$
\begin{aligned}
&= \sum_{j=1}^{k_t}|q_{t,j}(\gamma, \vec{p_t}) - y_{t,j}|\tilde{x}_{t,j}\tilde{x}_{t,j}^T \sum_{i=1}^{k_t}|q_{t,i}(\gamma, \vec{p_t}) - y_{t,i}| \\
&\preceq 2\sum_{j=1}^{k_t} \tilde{x}_{t,j}\tilde{x}_{t,j}^T,
\end{aligned}
$$

where the last inequality uses that for all $i \in [k_t]$, $|q_{t,i}(\gamma, \vec{p_t}) - y_{t,i}| \leq 1$ and that

$$\sum_{i=1}^{k_t} |q_{t,i}(\gamma, \vec{p_t}) - y_{t,i}| = \begin{cases} \sum_{i \in [k_t] \setminus \{i_t\}} q_{t,i}(\gamma, \vec{p_t}) + (1 - q_{t,i_t}(\gamma, \vec{p_t})) & \text{if } i_t \in [k_t] \\ \sum_{i \in [k_t] \setminus \{i_t\}} q_{t,i}(\gamma, \vec{p_t}) & \text{if } i_t = 0 \end{cases} \leq 2.$$

4. Similarly,

$$\nabla \ell_{t,1}(\gamma_t)^T \nabla \ell_{t,1}(\gamma_t)^T = \sum_{j=1}^{k_t} \sum_{i=1}^{k_t} (q_{t,j}(\gamma, \vec{p_t}) - y_{t,j})(q_{t,i}(\gamma, \vec{p_t}) - y_{t,i}) \tilde{x}_{t,j}^T \tilde{x}_{t,i}$$

$$\leq \frac{1}{2} \sum_{j=1}^{k_t} \sum_{i=1}^{k_t} |q_{t,j}(\gamma, \vec{p_t}) - y_{t,j}| |q_{t,i}(\gamma, \vec{p_t}) - y_{t,i}| (\tilde{x}_{t,j}^T \tilde{x}_{t,j} + \tilde{x}_{t,i}^T \tilde{x}_{t,i})$$

by (17)

$$\leq 2 \sum_{j=1}^{k_t} \tilde{x}_{t,j} \tilde{x}_{t,j}^T.$$

$\square$

B.3 Proof of Lemma 2.8

In order to prove Lemma 2.8, we will combine the lower and upper bounds on $\sum_{t=1}^{T}[\ell_{t,1}(\gamma_t) - \ell_{t,1}(\gamma^*)]$ provided in the two following lemmas. The proofs are deferred to Appendices B.4 and B.5, respectively.

Lemma B.6 (Upper bound $\sum_{t=1}^{T} \ell_{t,1}(\gamma_t) - \sum_{t=1}^{T} \ell_{t,1}(\gamma^*)$)**.**

$$\sum_{t=1}^{T}[\ell_{t,1}(\gamma_t) - \ell_{t,1}(\gamma^*)] \leq \frac{2d}{\kappa_1} \log \left(\lambda_T + \frac{2TK}{d} \right) + 2\mu W^2 - \frac{\mu \kappa_1}{2} \sum_{t=1}^{T} \sum_{j=1}^{k_t} ((\gamma_t - \gamma^*)^T \tilde{x}_{t,j})^2$$

Lemma B.7 (Lower bound $\sum_{t=1}^{T} \ell_{t,1}(\gamma_t) - \sum_{t=1}^{T} \ell_{t,1}(\gamma^*)$)**.**

$$\mathbb{P}\left(\sum_{t=1}^{T}[\ell_{t,1}(\gamma_t) - \ell_{t,1}(\gamma^*)] \leq -2 \sqrt{2 \log T \sum_{t=1}^{T} \sum_{j=1}^{k_t} ((\gamma_t - \gamma^*)^T \tilde{x}_{t,j})^2} \right.$$
$$\left. - KW(1 + p_{max}) \left(\frac{4 \log(T)}{3} + 1 \right) \right) \leq \frac{\lceil 2 \log_2(T) + 1 \rceil}{T^2}. \quad (18)$$

We now prove the following expanded version of Lemma 2.8.

Lemma 2.8 1. With probability at least $1 - \frac{\log_2(T)}{T^2}$,

$$\sum_{t=1}^{T} \sum_{j=1}^{k_t} (x_{t,j}^\top (\theta_t - \theta^*) - x_{t,j}^\top (\alpha_t - \alpha^*) p_{t,j})^2 \leq C(T) := \max \left\{ \frac{128 \log(T)}{\kappa_1^2 \mu^2}, \right.$$

$$\left. \frac{4}{\mu \kappa_1} \left(\frac{2d}{\kappa_1} \log \left(\lambda_T + \frac{2TK}{d} \right) + 2\mu W^2 + KW(1 + p_{max}) \left(\frac{4 \log(T)}{3} + 1 \right) \right) \right\}$$

Using that $\lambda_t = d \log(t)$ for all $t \geq 2$, we get that for a constant $\tilde{C}$ depending only on W, L, K:

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

22

Proof. To ease the presentation, let

$$B(T) := \frac{2d}{\kappa_1} \log\left(\lambda_T + \frac{2TK}{d}\right) + 2\mu W^2 + KW(1 + p_{max})\left(\frac{4\log(T)}{3} + 1\right)$$

We obtain by combining Lemmas B.6 and B.7 that with probability at least $1 - \frac{\lceil\log(T)\rceil}{T^2}$:

$$\frac{\mu\kappa_1}{2} \sum_{t=1}^{T} \sum_{j=1}^{k_t} ((\gamma_t - \gamma^*)^T \tilde{x}_{t,j})^2 - 2\sqrt{2\log T}\left\{\sum_{t=1}^{T} \sum_{j=1}^{k_t} ((\gamma_t - \gamma^*)^T \tilde{x}_{t,j})^2\right\}^{1/2} \leq B(T) \quad (19)$$

Now let $A = \sum_{t=1}^{T} \sum_{j=1}^{k_t} ((\gamma_t - \gamma^*)^T \tilde{x}_{t,j})^2$ and consider the two following cases:

- Suppose $\frac{1}{2}\frac{\mu\kappa_1 A}{2} \geq 2\sqrt{2\log(T)}\sqrt{A}$. Then from equation (19), we obtain that: $A \leq \frac{4B(T)}{\mu\kappa_1}$.

- Else, $\frac{1}{2}\frac{\mu\kappa_1 A}{2} < 2\sqrt{2\log(T)}\sqrt{A}$ hence by reorganizing the terms, we get $A \leq \frac{128\log(T)}{\kappa_1^2\mu^2}$.

The proof is complete. $\qquad\square$

B.4 Proof of Lemma B.6

Proof. Consider $\hat{\gamma}_{t+1} = \gamma_t - \frac{1}{\mu}H_{T+1}^{-1}\nabla\ell_{t,1}(\gamma_t)$, so that γ_{t+1} is the projection of $\hat{\gamma}_{t+1}$ in the norm induced by H_t.

By doing a Taylor expansion of $\ell_{t,1}$, we obtain:

$$\ell_{t,1}(\gamma_t) - \ell_{t,1}(\gamma^*) = \nabla\ell_{t,1}(\gamma_t)^T(\gamma_t - \gamma^*)$$
$$- (\gamma_t - \gamma^*)^T \int_0^1 \int_0^1 z\nabla^2\ell_{t,1}(\gamma_t + zw(\gamma^* - \gamma_t))dzdw(\gamma_t - \gamma^*)$$

Using Proposition A.4 and the definition of $\mu = \frac{1}{2(1+(1+p_{max})2\sqrt{6KW})}$, this leads to:

$$\ell_{t,1}(\gamma_t) - \ell_{t,1}(\gamma^*) \leq \nabla\ell_{t,1}(\gamma_t)^T(\gamma_t - \gamma^*) - \mu(\gamma_t - \gamma^*)^T\nabla^2\ell_{t,1}(\gamma_t)(\gamma_t - \gamma^*) \quad (20)$$

Note that in the classical Online Newton Step analysis, a similar equation as the above equation is used, but with the potentially exponentially small exp-concavity parameter β instead of μ. The next part of the proof globally follows the ONS analysis in [20]. We include it here for completeness.

By definition of $\hat{\gamma}_{t+1}$, we can write the following two equalities:

$$\hat{\gamma}_{t+1} - \gamma^* = \gamma_t - \gamma^* - \frac{1}{\mu}H_t^{-1}\nabla\ell_{t,1}(\gamma_t)$$

and

$$H_t(\hat{\gamma}_{t+1} - \gamma^*) = H_t(\gamma_t - \gamma^*) - \frac{1}{\mu}\nabla\ell_{t,1}(\gamma_t).$$

Hence, by multiplying the transpose of the first inequality with the second inequality, we obtain

$$(\hat{\gamma}_{t+1} - \gamma^*)^T H_t(\hat{\gamma}_{t+1} - \gamma^*)$$
$$= (\gamma_t - \gamma^*)^T H_t(\gamma_t - \gamma^*) - \frac{2}{\mu}\nabla\ell_{t,1}^T(\gamma_t)(\gamma_t - \gamma^*) + \frac{1}{\mu^2}\nabla\ell_{t,1}(\gamma_t)^T H_t^{-1}\nabla\ell_{t,1}(\gamma_t)$$

Since γ_{t+1} is the projection of $\hat{\gamma}_{t+1}$ on B_{t+1} relatively to the norm induced by H_t, we have the following inequality:

$$(\hat{\gamma}_{t+1} - \gamma^*)^T H_t(\hat{\gamma}_{t+1} - \gamma^*)^T \geq (\gamma_{t+1} - \gamma^*)^T H_t(\gamma_{t+1} - \gamma^*) \quad (21)$$

Combining equations (??) and (21) gives the following bound on $\nabla\ell_{t,1}(\gamma_t)^T(\gamma_t - \gamma^*)$:

$$\nabla\ell_{t,1}(\gamma_t)^T(\gamma_t - \gamma^*)$$
$$\leq \frac{1}{2\mu}\nabla\ell_{t,1}(\gamma_t)^T H_t^{-1}\nabla\ell_{t,1}(\gamma_t) + \frac{\mu}{2}(\gamma_t - \gamma^*)H_t(\gamma_t - \gamma^*) - \frac{\mu}{2}(\gamma_{t+1} - \gamma^*)^T H_t(\gamma_{t+1} - \gamma^*)$$

Hence, summing from $t = 1$ to T:

$$\sum_{t=1}^{T} \nabla \ell_{t,1}(\gamma_t)^T (\gamma_t - \gamma^*) \le \frac{1}{2\mu} \sum_{t=1}^{T} \nabla \ell_{t,1}(\gamma_t)^T H_t^{-1} \nabla \ell_{t,1}(\gamma_t)$$
$$+ \frac{\mu}{2}(\gamma_1 - \gamma^*) H_1 (\gamma_1 - \gamma^*) - \frac{\mu}{2}(\gamma_{T+1} - \gamma^*) H_{T+1}(\gamma_{T+1} - \gamma^*)$$
$$+ \frac{\mu}{2} \sum_{t=2}^{T} (\gamma_t - \gamma^*)^T (H_t - H_{t-1})(\gamma_t - \gamma^*)$$
$$\le \frac{1}{2\mu} \sum_{t=1}^{T} \nabla \ell_{t,1}(\gamma_t)^T H_t^{-1} \nabla \ell_{t,1}(\gamma_t)$$
$$+ \frac{\mu}{2}(\gamma_1 - \gamma^*)^T (H_1 - \nabla^2 l_1(\gamma_1))(\gamma_1 - \gamma^*)$$
$$+ \frac{\mu}{2} \sum_{t=1}^{T} (\gamma_t - \gamma^*)^T (H_t - H_{t-1})(\gamma_t - \gamma^*)$$

In the last part of the proof, we bring out more specifically the sum $\sum_{t=1}^{T} \sum_{j=1}^{k_t}((\gamma_t - \gamma^*)^T \tilde{x}_{t,j})^2$. Since $H_t - H_{t-1} = \nabla^2 \ell_{t,1}(\gamma_t)$, we obtain, by combining the above inequality with (20) and by using the lower bound on $\nabla \ell_{t,1}^2$ provided in Lemma B.5:

$$\sum_{t=1}^{T} \ell_{t,1}(\gamma_t) - \ell_{t,1}(\gamma^*)$$
$$\le \sum_{t=1}^{T} \nabla \ell_{t,1}(\gamma_t)^T (\gamma_t - \gamma^*) - \mu \sum_{t=1}^{T} (\gamma_t - \gamma^*)^T \nabla^2 \ell_{t,1}(\gamma_t)(\gamma_t - \gamma^*)$$
$$\le \frac{1}{2\mu} \sum_{t=1}^{T} \nabla \ell_{t,1}(\gamma_t)^T H_t^{-1} \nabla \ell_{t,1}(\gamma_t) + \frac{\mu}{2}(\gamma_1 - \gamma^*)^T (H_1 - \nabla^2 l_1(\gamma_1))(\gamma_1 - \gamma^*)$$
$$- \frac{\mu}{2} \sum_{t=1}^{T} (\gamma_t - \gamma^*)^T \nabla^2 \ell_{t,1}(\gamma_t)(\gamma_t - \gamma^*)$$
$$\le \frac{1}{2\mu} \sum_{t=1}^{T} \nabla \ell_{t,1}(\gamma_t)^T H_t^{-1} \nabla \ell_{t,1}(\gamma_t) + \frac{\mu}{2}(\gamma_1 - \gamma^*)^T (H_1 - \nabla^2 l_1(\gamma_1))(\gamma_1 - \gamma^*)$$
$$- \frac{\mu \kappa_1}{2} \sum_{t=1}^{T} \sum_{j=1}^{k_t}((\gamma_t - \gamma^*)^T \tilde{x}_{t,j})^2$$

Applying Lemma B.8 (stated below) and noting that $\frac{\mu}{2}(\gamma_1 - \gamma^*)^T (H_1 - \nabla^2 l_1(\gamma_1))(\gamma_1 - \gamma^*) = \frac{\mu}{2}(\gamma_1 - \gamma^*)^T (\lambda_1 I_d)(\gamma_1 - \gamma^*) = \frac{\mu \lambda_1 \|\gamma_1 - \gamma^*\|^2}{2} \le \frac{\mu \lambda_1 (2W)^2}{2} = 2\mu W^2$ concludes the proof of the lemma.

$\square$

Lemma B.8.

$$\sum_{t=1}^{T} \nabla \ell_{t,1}(\gamma_t)^T H_t^{-1} \nabla \ell_{t,1}(\gamma_t) \le \frac{2d}{\kappa_1} \log \left(\lambda_T + \frac{2TK}{d} \right)$$

Proof. By definition of H_t:

$$H_{t+1} = H_t + \nabla^2 \ell_{t+1,1}(\gamma_{t+1}) + (\lambda_{t+1} - \lambda_t)I_d$$

$$\succeq H_t + \kappa_1 \sum_{j=1}^{k_{t+1}} \tilde{x}_{t+1,j}\tilde{x}_{t+1,j}^T \qquad \text{(by Lemma B.5 and using } \lambda_{t+1} \geq \lambda_t)$$

$$\succeq H_t + \frac{\kappa_1}{2}\nabla \ell_{t+1,1}(\gamma_{t+1})\nabla \ell_{t+1,1}(\gamma_{t+1})^T. \qquad \text{(by Lemma B.5)}$$

The proof now globally uses similar techniques as in the proof of Lemma 11.11 in [10]. From above, we obtain

$$\det(H_{t+1}) \cdot \det\left(I_d - \frac{\kappa_1}{2}H_{t+1}^{-1/2}\nabla \ell_{t+1,1}(\gamma_{t+1})\nabla \ell_{t+1,1}(\gamma_{t+1})^T H_{t+1}^{-1/2}\right) \geq \det(H_t). \qquad (22)$$

Using that for all $z \in \mathbb{R}^d$, $\det(1 - zz^T) = 1 - z^Tz$ (see [10], Lemma 11.11), and that H_{t+1} is symmetric, we get

$$\det\left(I_d - \frac{\kappa_1}{2}H_{t+1}^{-1/2}\nabla \ell_{t+1,1}(\gamma_{t+1})\nabla \ell_{t+1,1}(\gamma_{t+1})^T H_{t+1}^{-1/2}\right)$$

$$= \det\left(I_d - \left(\sqrt{\frac{\kappa_1}{2}}H_{t+1}^{-1/2}\nabla \ell_{t+1,1}(\gamma_{t+1})\right)\left(\sqrt{\frac{\kappa_1}{2}}H_{t+1}^{-1/2}\nabla \ell_{t+1,1}(\gamma_{t+1})\right)^T\right)$$

$$= 1 - \frac{\kappa_1}{2}\nabla \ell_{t+1,1}(\gamma_{t+1})^T H_{t+1}^{-1}\nabla \ell_{t+1,1}(\gamma_{t+1}).$$

Hence, combining this with (22) and reorganizing the terms, we obtain

$$\frac{\kappa_1}{2}\nabla \ell_{t+1,1}(\gamma_{t+1})^T H_{t+1}^{-1}\nabla \ell_{t+1,1}(\gamma_{t+1}) \leq 1 - \frac{\det(H_t)}{\det(H_{t+1})}.$$

Summing from $t = 0$ to $T - 1$ and translating the indices gives:

$$\frac{\kappa_1}{2}\sum_{t=1}^{T}\nabla \ell_{t,1}(\gamma_t)^T H_t^{-1}\nabla \ell_{t,1}(\gamma_t) \leq \sum_{t=1}^{T}\left(1 - \frac{\det(H_{t-1})}{\det(H_t)}\right)$$

$$\leq \sum_{t=1}^{T}\log\left(\frac{\det(H_t)}{\det(H_{t-1})}\right) \qquad (1 - x \leq -\log(x) \text{ for all } x > 0)$$

$$= \log\left(\frac{\det(H_T)}{\det(\lambda_1 I_d)}\right). \qquad (23)$$

We now bound $\log(\det(H_T))$. First note that for all $t \in [T]$, we have by Lemma B.5 that $\nabla^2 \ell_{t,1}(\gamma_t) \preceq 2\sum_{j=1}^{k_t}\tilde{x}_{t,j}\tilde{x}_{t,j}^T$. Since $||x_{t,j}||_2 \leq 1$ for all $t \in [T], j \in [k_t]$, it follows that:

$$\text{trace}(H_T) = \text{trace}\left(\lambda_T I_d + \sum_{t=1}^{T}\nabla^2 \ell_{t,1}(\gamma_t)\right)$$

$$\leq \text{trace}(\lambda_T I_d) + 2\sum_{t=1}^{T}\sum_{j=1}^{k_t}\text{trace}\left(\tilde{x}_{t,j}\tilde{x}_{t,j}^T\right)$$

$$= \text{trace}(\lambda_T I_d) + 2\sum_{t=1}^{T}\sum_{j=1}^{k_t}\text{trace}\left(\tilde{x}_{t,j}^T\tilde{x}_{t,j}\right)$$

$$\leq d\lambda_T + 2TK.$$

Hence, using the determinant-trace inequality $\det(H_T) \leq \left(\frac{\text{trace}(H_T)}{d}\right)^d$ (see [10]) and $\lambda_1 = 1$, we obtain that:

$$\log\left(\frac{\det(H_T)}{\det(\lambda_1 I_d)}\right) = \log\left(\frac{\det(H_T)}{\lambda_1^d}\right) \leq d\log\left(\lambda_T + \frac{2TK}{d}\right) \qquad (24)$$

Putting (23) and (24) together concludes the proof of the lemma. □

B.5 Proof of Lemma B.7

The proof relies on a Bernstein type inequality for martingales difference sequences from [19]:

Proposition B.9. *([19], Th 1.6) Let $X_1, ..., X_n$ be a bounded martingale difference sequence with respect to the filtration $(\mathcal{F}_i)_{1 \leq i \leq n}$ such that $|X_i| \leq M$ for all i.*

Let Σ_n^2 denote the sum of the conditional variances.

$$\Sigma_n^2 = \sum_{i=1}^{n} \mathbb{E}(X_i^2 | \mathcal{F}_{i-1})$$

Then,

$$\mathbb{P}(\sum_{i=1}^{n} X_i \geq \epsilon, \Sigma_n^2 \leq k) \leq exp\left(\frac{-\epsilon^2}{2(k + \frac{\epsilon M}{3})}\right).$$

We are now ready to prove Lemma B.7.

Proof. By convexity of $\ell_{t,1}$:

$$\ell_{t,1}(\gamma^*) - \ell_{t,1}(\gamma_t) \leq \nabla \ell_{t,1}(\gamma^*)^T (\gamma^* - \gamma_t),$$

Hence, letting $D_t = \nabla \ell_{t,1}(\gamma^*)^T (\gamma^* - \gamma_t)$, we have

$$\mathbb{P}\left(\sum_{t=1}^{T} \ell_{t,1}(\gamma_t) - \sum_{t=1}^{T} \ell_{t,1}(\gamma^*) \leq -2\sqrt{2 \log T \sum_{t=1}^{T} \sum_{j=1}^{k_t} ((\gamma_t - \gamma^*)^T \tilde{x}_{t,j})^2} - \frac{4B \log(T)}{3} - B\right)$$

$$\leq \mathbb{P}\left(\sum_{t=1}^{T} D_t \geq 2\sqrt{2 \log T \sum_{t=1}^{T} \sum_{j=1}^{k_t} ((\gamma_t - \gamma^*)^T \tilde{x}_{t,j})^2} + \frac{4B \log(T)}{3} + B\right). \quad (25)$$

Let $\mathcal{F}_t$ be the filtration generated by $\{X_1, \Delta \vec{p}_1, z_1, ..., X_t, \Delta \vec{p}_t, z_t, X_{t+1}, \Delta \vec{p}_{t+1}\}$. Since $\gamma_t, \vec{p}_t$ are $\mathcal{F}_{t-1}$ measurable, and using the expression of $\nabla \ell_t$ in (14), and the fact that for all $j \in [k_t]$, $\mathbb{E}[y_{t,j}] = q_{t,j}(\gamma^*, \vec{p}_t)$, we have that:

$$\mathbb{E}(D_t | \mathcal{F}_{t-1}) = \mathbb{E}\left(\sum_{j=1}^{k_t} (q_{t,j}(\gamma^*, \vec{p}_t) - y_{t,j})\tilde{x}_{t,j} \Bigg| \mathcal{F}_{t-1}\right)^T (\gamma^* - \gamma_t) = 0$$

Therefore, $\{D_t\}_{t=1}^{T}$ is a martingale difference sequence adapted to the filtration $\mathcal{F}_t$.

Moreover, using the Cauchy-Schwartz inequality, we have that D_t is uniformly bounded: $|D_t| = |\sum_{j=1}^{k_t} (q_{t,j}(\gamma^*, \vec{p}_t) - y_{t,j})\tilde{x}_{t,j}^T(\gamma_t - \gamma^*)| \leq \sum_{j=1}^{k_t} |(\gamma_t - \gamma^*)^T \tilde{x}_{t,j}| \leq KW(1 + p_{max})$. We let $B = KW(1 + p_{max})$.

Now, consider Σ_n^2 the sum of the conditional variances:

$$\Sigma_t^2 = \sum_{i=1}^{t} \mathbb{E}(D_i^2 | \mathcal{F}_{i-1}).$$

By Lemma B.5, we can bound Σ_t^2 as follows:

$$\Sigma_t^2 = \sum_{s=1}^{t} (\gamma_s - \gamma^*)^T \nabla \ell_{s,1}(\gamma^*) \nabla \ell_{s,1}(\gamma^*)^T (\gamma_s - \gamma^*) \leq \sum_{s=1}^{t} 2 \sum_{j=1}^{k_s} ((\gamma_s - \gamma^*)^T \tilde{x}_{s,j})^2 \equiv A_t.$$

Note that we cannot directly apply here the Bernstein inequality from Proposition B.9 with $k = A_t$ since A_t is also a random variable. We address this issue as in [43], making use of a peeling process. First note that by using the Cauchy-Scwhartz inequality, we have $A_t = \sum_{s=1}^{t} 2 \sum_{j=1}^{k_s} ((\gamma_s - \gamma^*)^T \tilde{x}_{s,j})^2 \leq 2KTW^2(1 + p_{max})^2 \leq 2B^2 T$.

Now, consider two cases:

- $A_T < \frac{B^2}{T}$. Then, using the Cauchy-Schwartz inequality, we get

$$\sum_{t=1}^{T} D_t \leq \sqrt{T \sum_{t=1}^{T} D_t^2} \leq \sqrt{T(\sum_{t=1}^{T} 2 \sum_{j=1}^{k_t} ((\gamma_t - \gamma^*)^T \tilde{x}_{t,j})^2} < \sqrt{\frac{TB^2}{T}} = B.$$

Thus,

$$\mathbb{P}\left(\sum_{t=1}^{T} D_t \geq \sqrt{2A_T \log(T^2)} + \frac{4B \log(T)}{3} + B \middle| A_T < \frac{B^2}{T} \right) = 0.$$

- $A_T \geq \frac{B^2}{T}$. Then, since by definition of B and Σ_T^2, we always have the upper bounds $A_T \leq 2B^2 T$ and $\Sigma_T^2 \leq A_T$, we have

$$\mathbb{P}\left(\sum_{t=1}^{T} D_t \geq \sqrt{2A_T \log(T^2)} + \frac{2B}{3} \log(T^2) \middle| A_T \geq \frac{B^2}{T} \right)$$

$$= \mathbb{P}\left(\sum_{t=1}^{T} D_t \geq \sqrt{2A_T \log(T^2)} + \frac{2B}{3} \log(T^2), \Sigma_T^2 \leq A_T, \frac{B^2}{T} \leq A_T \leq 2B^2 T \right)$$

$$\leq \sum_{i=1}^{m} \mathbb{P}\left(\sum_{t=1}^{T} D_t \geq \sqrt{\frac{2B^2 2^i log(T^2)}{T}} + \frac{2B}{3} \log(T^2), \Sigma_T^2 \leq A_T, \frac{B^2}{T} 2^{i-1} \leq A_T \leq \frac{B^2}{T} 2^i \right)$$

$$\leq \sum_{i=1}^{m} \mathbb{P}\left(\sum_{t=1}^{T} D_t \geq \sqrt{\frac{2B^2 2^i log(T^2)}{T}} + \frac{2B}{3} \log(T^2), \Sigma_T^2 \leq \frac{B^2}{T} 2^i \right)$$

$$\leq m e^{-log(T^2)}$$

with $m = \lceil \log_2(T^2) \rceil + 1$. The last inequality follows from the Bernstein's inequality for martingales (Proposition B.9) with $k = \frac{B^2}{T} 2^i$ and $\epsilon = \sqrt{2k \log(T^2)} + \frac{2B \log(T^2)}{3}$.

Hence, combining this with (25) we obtain:

$$\mathbb{P}\left(\sum_{t=1}^{T} \ell_{t,1}(\gamma_t) - \sum_{t=1}^{T} \ell_{t,1}(\gamma^*) \leq -2\sqrt{2 \log T \sum_{t=1}^{T} \sum_{j=1}^{k_t} ((\gamma_t - \gamma^*)^T \tilde{x}_{t,j})^2} - \frac{4B \log(T)}{3} - B \right)$$

$$\leq \mathbb{P}\left(\sum_{t=1}^{T} D_t \geq \sqrt{2A_T \log(T^2)} + \frac{4B \log(T)}{3} + B \right)$$

$$\leq \mathbb{P}\left(\sum_{t=1}^{T} D_t \geq \sqrt{2A_T \log(T^2)} + \frac{4B \log(T)}{3} + B \middle| A_T \geq \frac{B^2}{T} \right) \cdot \mathbb{P}(A_T \geq \frac{B^2}{T}) + 0$$

$$\leq \frac{\lceil \log_2(T^2) \rceil + 1}{T^2}. \tag{26}$$

$\square$

B.6 Proof of Lemmas B.2

Proof. Remember that the myopic prices are set as follows: for all $t \geq 1, j \in [k_t], g(X_t \alpha_t, X_t \theta_t)_j = \frac{1}{x_{t,j},^\top \alpha_t} + B_t^0$, where B_t^0 is the unique fixed of the following equation:

$$B = \sum_{j=1}^{k_t} \frac{1}{x_{t,j}^\top \alpha_t} e^{-(1+x_{t,j}^\top \alpha_t B)} e^{x_{t,j}^\top \theta_t}. \tag{27}$$

Define the functions $f_1(B) := \sum_{j=1}^{k_t} \frac{1}{x_{t,j}^\top \alpha_t} e^{-(1+x_{t,j}^\top \alpha_t B)} e^{x_{t,j}^\top \theta_t}$ and $f_2(B) := K \frac{1}{L} e^{-(1+LB)+W}$.

27

By Assumptions 2.2 and 2.3, we have that for for all $B \geq 0$, $f_1(B) \leq K\frac{1}{L}e^{-(1+LB)+W} = f_2(B)$. Now, let B^u be the solution of the following equation:

$$B = f_2(B).$$

Since both f_1 and f_2 are strictly decreasing, we have that for all $B > B^u$,

$$f_1(B) \leq f_2(B) < f_2(B^u) = B^u < B,$$

thus B is not solution of (27). Hence, we deduce that $B_t^0 \leq B^u$. It follows that the myopic prices are bounded above by $\frac{1}{L} + B^u$.

Next, recall that B^u satisfies

$$B^u = K\frac{1}{L}e^{-(1+LB^u)+W},$$

which, by reorganizing the terms, is equivalent to

$$B^u e^{LB^u} = K\frac{1}{L}e^{-1+W}.$$

Since for $B = K\max(W,1)/L$, we have that $Be^{LB} = \frac{K\max(W,1)}{L}e^{K\max(W,1)} \geq K\frac{1}{L}e^{-1+W}$ and since $B \longmapsto Be^{LB}$ is nondecreasing, we get that $B^u \leq K\max(W,1)/L$. Hence for all $t \geq 1, j \in [k_t]$, $g(X_t\alpha_t, X_t\theta_t)_j \leq \frac{1}{L} + B^u \leq \frac{1+K\max(W,1)}{L}$. Using that $|\Delta p_{t,j}| \leq \frac{1}{W}$, we deduce that $p_{t,j} = g(X_t\alpha_t, X_t\theta_t)_j + \Delta p_{t,j} \leq \frac{1+K\max(W,1)}{L} + \frac{1}{W}$. $\qquad \square$

C   MNL bandits technical proofs

C.1   Proof of Theorem 3.4

Before presenting the proof of Theorem 3.4, we need to introduce a few useful lemmas, whose proofs can be found in Appendix C.2. Lemma C.1 is the analogue of the elliptical potential lemma appearing in [1], but uses the local curvature information provided by the terms $\{q_{t,j}(S_t, \theta^*)q_{t,0}(S_t, \theta^*)\}$ to obtain an upper bound that no longer depends on the exponential constant $1/\kappa_2$. In particular, the proof uses the self-concordance-like property of the log-loss.

Lemma C.1.

$$\sum_{t=1}^{T}\sum_{j\in S_t} q_{t,j}(S_t, \theta^*)q_{t,0}(S_t, \theta^*)||x_{t,j}||_{H_t(\theta^*)^{-1}}^2 \leq 2dK\log\left(\lambda_{T+1} + \frac{2TK}{d}\right)$$

and

$$\sum_{t=1}^{T}\sum_{j\in S_t} ||x_{t,j}||_{H_t(\theta^*)^{-1}}^2 \leq 2d(K + \frac{1}{\kappa_2})\log\left(\lambda_{T+1} + \frac{2TK}{d}\right)$$

Lemma C.2.
$$\sum_{t=1}^{T}\sum_{j\in S_t} q_{t,j}(S_t, \theta^*)q_{t,0}(S_t, \theta^*) \leq \sum_{t=1}^{T}\kappa_{2,t}^* + R(T)$$

Note that we always have $\sum_{t=1}^{T}\sum_{j\in S_t} q_{t,j}(S_t, \theta^*)q_{t,0}(S_t, \theta^*) \leq T$. In Lemma C.2 we give a tighter upper bound on this sum when the instance is further away from linearity (i.e., when the parameters $\{\kappa_{2,t}^*\}_{t=1}^{T}$ are small).

Lemma C.3. *Define* $Q : \mathbb{R}^K \longrightarrow \mathbb{R}$, *such that for all* $u = \{u_1 \dots, u_K\} \in \mathbb{R}^n$, $Q(u) = \sum_{i=1}^{n}\frac{e^{u_i}}{1+\sum_{j=1}^{n}e^{u_j}}$. *Then for all* $i, j \in [K] \times [K]$,

$$\left|\frac{\partial^2 Q}{\partial i \partial j}\right| \leq 5.$$

Lemma C.4. *For all* $\theta_1, \theta_2 \in B(0, W)$

$$||\theta_1 - \theta_2||_{H_t(\theta_1)} \leq (1 + \sqrt{6K}W)||g_t(\theta_1) - g_t(\theta_2)||_{H_t^{-1}(\theta_1)}.$$

We now give the proof of Theorem 3.4.

Proof. Set $\delta = \frac{1}{K^2 T^2}$ and let A_δ denote the event that $\theta^* \in C_t(\delta)$ for all $t \in [T]$. We know from Proposition 3.3 that A_δ occurs with probability at least $1 - \delta$. We first assume that A_δ is satisfied.

Let $\tilde{\theta}^* = (\theta^*, ..., \theta^*) \in \mathbb{R}^{d \times N}$. Since $\theta^* \in C_t(\delta)$, we have by definition of $\tilde{\theta}_{t,j}$ that for all $j \in [k_t]$, $x_{t,j}^\top \tilde{\theta}_{t,j} \geq x_{t,j}^\top \theta^*$, from which we deduce $\tilde{r}_t(S_t^*, \tilde{\theta}_t) \geq \tilde{r}_t(S_t^*, \tilde{\theta}^*)$. Then, by definition of Algorithm 2, the assortment S_t offered at time t satisfies $\tilde{r}_t(S_t, \tilde{\theta}_t) \geq \tilde{r}_t(S_t^*, \tilde{\theta}_t)$ for all t. Hence we obtain $\sum_{t=1}^T \sum_{j \in S_t} \tilde{q}_{t,j}(S_t, \tilde{\theta}_t) = \tilde{r}_t(S_t, \tilde{\theta}_t) \geq \tilde{r}_t(S_t^*, \tilde{\theta}^*) = \sum_{t=1}^T \sum_{j \in S_t^*} \tilde{q}_{t,j}(S_t^*, \tilde{\theta}^*) = \sum_{t=1}^T \sum_{j \in S_t^*} q_{t,j}(S_t^*, \theta^*)$, where the last inequality follows by noting that $\tilde{q}_{t,j}(S_t^*, \tilde{\theta}^*) = q_{t,j}(S_t^*, \theta^*)$ for all $t \geq 1, j \in [k_t]$. Hence, by noting that for all $t \geq 1, j \in [k_t]$, we also have that $\tilde{q}_{t,j}(S_t, \tilde{\theta}^*) = q_{t,j}(S_t, \theta^*)$, we can bound the regret as follows:

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

$$\leq \sum_{t=1}^T \left[\sum_{j \in S_t} \tilde{q}_{t,j}(S_t, \tilde{\theta}_t) - \sum_{j \in S_t} \tilde{q}_{t,j}(S_t, \tilde{\theta}^*) \right].$$

Now, define $Q : \mathbb{R}^K \longrightarrow \mathbb{R}$, such that for all $u = \{u_1 \ldots, u_K\} \in \mathbb{R}^K$, $Q(u) = \sum_{i=1}^K \frac{e^{u_i}}{1 + \sum_{j=1}^K e^{u_j}}$. Noting that S_t always contain K elements, we write $S_t = \{i_1, \ldots, i_K\}$ where for all $j, i_j \in [N]$. Finally, for all $t \in [T]$, we let $u_t = (x_{t,i_1}^\top \theta_{t,1}, \ldots, x_{t,i_K}^T \theta_{t,K})^T$ and $u_t^* = (x_{t,i_1}^\top \theta^*, ..., x_{t,i_K}^\top \theta^*)^T$.

We obtain, by a second order Taylor expansion for all $t \geq 1$, that for some convex combination $\bar{u}_t$ of u_t and u_t^*, we have:

$$\sum_{t=1}^T \left[\sum_{j \in S_t} \tilde{q}_{t,j}(S_t, \tilde{\theta}_t) - \sum_{j \in S_t} \tilde{q}_{t,j}(S_t, \tilde{\theta}^*) \right]$$

$$= \sum_{t=1}^T Q(u_t) - Q(u_t^*)$$

$$= \sum_{t=1}^T \nabla Q(u_t^*)^T (u_t - u_t^*) + \frac{1}{2} \sum_{t=1}^T (u_t - u_T^*)^\top \nabla^2 Q_t(\bar{u}_t)(u_t - u_T^*)$$

$$= \sum_{t=1}^T \nabla Q(u_t^*)^T (u_t - u_t^*) + R_2(T), \tag{28}$$

where $R_2(T)$ is a second order term that we will explicit later. Now,

$$\sum_{t=1}^T \nabla Q(u_t^*)^T (u_t - u_t^*)$$

$$= \sum_{t=1}^T \left[\frac{\sum_{j \in S_t} e^{x_{t,j}^\top \theta^*} (u_j - u_j^*)}{1 + \sum_{j \in S_t} e^{x_{t,j}^\top \theta^*}} - \frac{\sum_{j \in S_t} (e^{x_{t,j}^\top \theta^*} \sum_{i \in S_t} e^{x_{t,i}^\top \theta^*} (u_j - u_j^*))}{\left(1 + \sum_{j \in S_t} e^{x_{t,j}^\top \theta^*} \right)^2} \right]$$

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

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

$$= \sum_{t=1}^T \sum_{j \in S_t} q_{t,j}(S_t, \theta^*) q_{t,0}(S_t, \theta^*) x_{t,j}^\top (\theta_{t,j} - \theta^*)$$

$$\leq \sum_{t=1}^{T} \sum_{j \in S_t} q_{t,j}(S_t, \theta^*) q_{t,0}(S_t, \theta^*) ||x_{t,j}||_{H_t(\theta^*)^{-1}} ||\theta_{t,j} - \theta^*||_{H_t(\theta*)}$$

$$\overset{(a)}{\leq} (1 + \sqrt{6K}W) \sum_{t=1}^{T} \gamma_t(\delta) \sum_{j \in S_t} q_{t,j}(S_t, \theta^*) q_{t,0}(S_t, \theta^*) ||x_{t,j}||_{H_t(\theta^*)^{-1}}$$

$$\overset{(b)}{\leq} (1 + \sqrt{6K}W) \bar{\gamma}_T(\delta) \sqrt{\sum_{t=1}^{T} \sum_{j \in S_t} q_{t,j}(S_t, \theta^*) q_{t,0}(S_t, \theta^*)}$$

$$\cdot \sqrt{\sum_{t=1}^{T} \sum_{j \in S_t} q_{t,j}(S_t, \theta^*) q_{t,0}(S_t, \theta^*) ||x_{t,j}||_{H_t(\theta^*)^{-1}}^2}, \quad (29)$$

where $\bar{\gamma}_T(\delta) := \max_{t \in [T]} \gamma_t(\delta)$. Since $\tilde{\theta}_{t,j} \in C_t(\delta)$ for all $t \geq 1$, inequality (a) is a consequence of Lemma C.4 and the assumption that A_δ is satisfied. Inequality (b) results from the Cauchy-Schwartz inequality.

Noting that, since $\delta = \frac{1}{K^2 T^2}$, we have that for some constant C which depends only polynomially on W and does not depend on T, d, K, $\bar{\gamma}_T(\delta) \leq C\sqrt{d \log(KT)}$. Thus, by combining Lemmas C.1 and C.2 with inequality (29), we get that for some constant C_1 which depends only polynomially on W and does not depend on T, d, K:

$$\sum_{t=1}^{T} \nabla Q(u_t^*)^T (u_t - u_t^*) \leq C_1 K d \log(KT) \left(\sqrt{\sum_{t=1}^{T} \kappa_{2,t}^* + R(T)} \right)$$

$$\leq C_1 K d \log(KT) \left(\sqrt{\sum_{t=1}^{T} \kappa_{2,t}^*} + \sqrt{R(T)} \right) \quad (30)$$

We now provide a crude upper bound on the second order term $R_2(T)$.

$$R_2(T) = \frac{1}{2} \sum_{t=1}^{T} (u_t - u_T^*)^\top \nabla^2 Q(\bar{u}_t)(u_t - u_T^*)$$

$$\leq \frac{5}{2} \sum_{t=1}^{T} \sum_{j=1}^{K} \sum_{i=1}^{K} x_{t,j}^\top (\tilde{\theta}_{t,j} - \theta^*) x_{t,i}^\top (\tilde{\theta}_{t,i} - \theta^*)$$

$$\leq \frac{5}{2} \sum_{t=1}^{T} \frac{1}{2} \sum_{j=1}^{K} \sum_{i=1}^{K} [(x_{t,j}^\top (\tilde{\theta}_{t,j} - \theta^*))^2 + (x_{t,i}^\top (\tilde{\theta}_{t,i} - \theta^*))^2]$$

$$= \frac{5}{2} K \sum_{t=1}^{T} \sum_{j=1}^{K} (x_{t,j}^\top (\tilde{\theta}_{t,j} - \theta^*))^2$$

$$\leq \frac{5}{2} K \sum_{t=1}^{T} \sum_{j=1}^{K} ||x_{t,j}||_{H_t(\theta^*)^{-1}}^2 ||\tilde{\theta}_{t,j} - \theta^*||_{H_t(\theta^*)}^2$$

$$\leq \frac{5}{2} K \bar{\gamma}_T(\delta)^2 (1 + \sqrt{6K}W)^2 2d(K + \frac{1}{\kappa_2}) \log \left(\lambda_{T+1} + \frac{2TK}{d} \right) \quad (31)$$

where the first inequality results from Lemma C.3 and the last one from Lemmas C.1 and C.4 and the fact that $\tilde{\theta}_{t,j} \in C_t(\delta)$ for all $j \in [K]$.

Using again that $\bar{\gamma}_T(\delta) \leq C\sqrt{d \log(KT)}$, we obtain that for some constant C_2 which depends only polynomially on W and does not depend on T, d, K:

$$R_2(T) \leq \frac{C_2 d^2 K^3}{\kappa_2} \log(KT)^2$$

Coming back to equation (28) and using the upper bounds given by (30) and (31), we obtain:

$$R(T) - C_1 K d \log(KT) \sqrt{R(T)} \leq \frac{C_2 d^2 K^3}{\kappa_2} \log(KT)^2 + C_1 K d \log(KT) \left(\sqrt{\sum_{t=1}^{T} \kappa_{2,t}^*} \right).$$

Consider the two following cases:

- $C_1 K d \log(KT) \sqrt{R(T)} \leq \frac{R(T)}{2}$.
  Then $R(T) \leq 2 \left(\frac{C_2 d^2 K^3}{\kappa_2} \log(KT)^2 + C_1 K d \log(KT) \left(\sqrt{\sum_{t=1}^{T} \kappa_{2,t}^*} \right) \right)$

- Otherwise, $C_1 K d \log(KT) \geq \frac{\sqrt{R(T)}}{2}$, hence $R(T) \leq 4K^2 C_1^2 d^2 \log(KT)^2$.

Hence,

$$R(T) \leq \max \left\{ \frac{2C_2 d^2 K^3}{\kappa_2} \log(KT)^2 + 2C_1 K d \log(KT) \left(\sqrt{\sum_{t=1}^{T} \kappa_{2,t}^*} \right), 4K^2 C_1^2 d^2 \log(KT)^2 \right\}.$$

To finish the proof, we consider the case where A_δ is not satisfied. In this case, $R(T)$ is still upper bounded by KT.

Hence, using that $\delta = \frac{1}{K^2 T^2}$ and by using the law of total probabilities, we conclude that there are some constants $\tilde{C}_1, \tilde{C}_2$ which depends only polynomially on W and do not depend on T, d, K such that:

$$R(T) \leq \tilde{C}_1 K d \log(KT) \left(\sqrt{\sum_{t=1}^{T} \kappa_{2,t}^*} \right) + \frac{\tilde{C}_2 d^2 K^3}{\kappa_2} \log(KT)^2.$$

$\square$

C.2 Proofs of the main lemmas

Proof of Lemma C.1. The proof is similar in spirit to the proof of Lemma B.8 and once again is inspired by the proof of the elliptical potential lemma in [1], while incorporating the local curvature information given by the terms $\{q_{t,j}(S_t, \theta^*)q_{t,0}(S_t, \theta^*)\}$.

$$H_t(\theta^*) = H_{t-1}(\theta^*) + \sum_{i \in S_t} q_{t,i}(S_t, \theta^*) x_{t,i} x_{t,i}^T$$

$$- \sum_{i \in S_t} \sum_{j \in S_t} q_{t,i}(S_t, \theta^*) q_{t,j}(S_t, \theta^*) x_{t,i} x_{t,j}^T + (\lambda_t - \lambda_{t-1}) I_d$$

$$\succeq H_{t-1}(\theta^*) + \sum_{i \in S_t} q_{t,i}(S_t, \theta^*) q_{t,0}(S_t, \theta^*) x_{t,i} x_{t,i}^T, \tag{32}$$

where we used in the last inequality that $\lambda_t \geq \lambda_{t-1}$ and a similar argument as used before. Hence we obtain that

$$\det(H_t(\theta^*)) = \det(H_{t-1}(\theta^*)) \left(1 + \sum_{i \in S_t} q_{t,i}(S_t, \theta^*) q_{t,0}(S_t, \theta^*) \|x_{t,i}\|_{H_t(\theta^*)^{-1}}^2 \right)$$

Taking the log on both sides and summing from $t = 1$ to T, we get:

$$\sum_{t=1}^{T} \log \left(1 + \sum_{i \in S_t} q_{t,i}(S_t, \theta^*) q_{t,0}(S_t, \theta^*) \|x_{t,i}\|^2_{H_t(\theta^*)^{-1}} \right)$$

$$\leq \sum_{t=1}^{T} \log(\det(H_t(\theta^*))) - \log(\det(H_{t-1}(\theta^*)))$$

$$= \log \left(\frac{\det(H_{T+1}(\theta^*))}{\det(H_1(\theta^*))} \right)$$

$$= \log(\det(H_{T+1}(\theta^*))) \qquad\qquad (\lambda_1 = 1))$$

$$\leq \log \left(\frac{(\text{trace}(H_{T+1}))^d}{d} \right) \qquad\text{(determinant-trace inequality (see [10]))}$$

$$\leq d \log \left(\lambda_{T+1} + \frac{2TK}{d} \right). \qquad\qquad\text{(similarly as in C.1)}$$

Since

$$\sum_{i \in S_t} q_{t,i}(S_t, \theta^*) q_{t,0}(S_t, \theta^*) \|x_{t,i}\|^2_{H_t(\theta^*)^{-1}}$$

$$\leq \frac{1}{\lambda_{min}(H_t(\theta^*))} \sum_{i \in S_t} q_{t,i}(S_t, \theta^*) q_{t,0}(S_t, \theta^*) \|x_{t,i}\|^2_2 \leq \frac{K}{\lambda_1} = K,$$

we get:

$$\sum_{t=1}^{T} \log \left(1 + \sum_{i \in S_t} q_{t,i}(S_t, \theta^*) q_{t,0}(S_t, \theta^*) \|x_{t,i}\|^2_{H_t(\theta^*)^{-1}} \right)$$

$$\geq \sum_{t=1}^{T} \log \left(1 + \frac{1}{K} \sum_{i \in S_t} q_{t,i}(S_t, \theta^*) q_{t,0}(S_t, \theta^*) \|x_{t,i}\|^2_{H_t(\theta^*)^{-1}} \right)$$

$$\geq \sum_{t=1}^{T} \frac{1}{2K} \sum_{i \in S_t} q_{t,i}(S_t, \theta^*) q_{t,0}(S_t, \theta^*) \|x_{t,i}\|^2_{H_t(\theta^*)^{-1}} \qquad (\log(1+x) \geq \frac{x}{2} \text{ for } x \in [0,1])$$

We deduce:

$$\sum_{t=1}^{T} \sum_{i \in S_t} q_{t,i}(S_t, \theta^*) q_{t,0}(S_t, \theta^*) \|x_{t,i}\|^2_{H_t(\theta^*)^{-1}} \leq 2dK \log \left(\lambda_{T+1} + \frac{2TK}{d} \right).$$

To show the second inequality, we come back to equation (32) and further lower bound it using the definition of κ_2:

$$H_t(\theta^*) \succeq H_{t-1}(\theta^*) + \kappa_2 \sum_{i \in S_t} x_{t,i} x_{t,i}^T.$$

We then conclude on the same way:

$$d \log \left(\lambda_{T+1} + \frac{2TK}{d} \right) \geq \sum_{t=1}^{T} \log \left(1 + \sum_{i \in S_t} \kappa_2 \|x_{t,i}\|^2_{H_t(\theta^*)^{-1}} \right)$$

$$\geq \sum_{t=1}^{T} \log \left(1 + \frac{1}{\max(1, K\kappa_2)} \sum_{i \in S_t} \kappa_2 \|x_{t,i}\|^2_{H_t(\theta^*)^{-1}} \right)$$

$$\geq \frac{1}{2(1 + K\kappa_2)} \sum_{i \in S_t} \kappa_2 \|x_{t,i}\|^2_{H_t(\theta^*)^{-1}}$$

Hence,

$$\sum_{i \in S_t} \|x_{t,i}\|_{H_t(\theta^*)^{-1}}^2 \le \frac{1}{\kappa_2} 2d(1 + \kappa_2 K) \log \left(\lambda_{T+1} + \frac{2TK}{d} \right)$$

$\square$

Proof of Lemma C.2. Since S_t^* and S_t both contain K elements, we write $S_t = \{i_1, ..., i_K\}$, $S_t^* = \{j_1, ..., j_K\}$, and we define $u_t := (x_{t,i_1}^\top \theta^*, ..., x_{t,i_K}^\top \theta^*)^T$, $u_t^* := (x_{t,j_1}^\top \theta^*, ..., x_{t,j_K}^\top \theta^*)^T$ the vectors of the true utilities from products in S_t and S_t^*, respectively.

Without loss of generality, we assume that the elements of u_t and u_t^* are sorted by ascending order. Since S_t^* contains the products with the K top utilities, we thus have that for all $i \in [k_t]$, $u_{ti}^* \ge u_{ti}$.

Now, let:

$$g(u) := \sum_{j=1}^{K} \frac{e^{u_j}}{(1 + \sum_{j=1}^{N} e^{u_j})^2}$$

Note that $g(u_t) = \sum_{j \in S_t} q_{t,j}(S_t, \theta^*) q_{t,0}(S_t, \theta^*)$.

Using the mean value theorem, we obtain that:

$$g(u_t) \tag{33}$$

$$= g(u_t^*) + \int_0^1 \nabla g(u_t^* + z(u_t - u_t^*)) dz^\top (u_t - u_t^*)$$

$$= g(u_t^*) + \sum_{i=1}^{K} \int_0^1 \frac{e^{u_{ti}^* + z(u_{ti} - u_{ti}^*)}}{(1 + \sum_{j=1}^{K} e^{u_{tj}^* + z(u_{tj} - u_{tj}^*)})^2} \left(1 - 2 \sum_{k=1}^{K} \frac{e^{u_{tk}^* + z(u_{tk} - u_{tk}^*)}}{1 + \sum_{j=1}^{K} e^{u_{tj}^* + z(u_{tj} - u_{tj}^*)}} \right) (u_{ti} - u_{ti}^*) dz$$

$$\le g(u_t^*) + \sum_{i=1}^{K} \left| \int_0^1 \frac{e^{u_{ti}^* + z(u_{ti} - u_{ti}^*)}}{(1 + \sum_{j=1}^{K} e^{u_{tj}^* + z(u_{tj} - u_{tj}^*)})^2} \left(1 - 2 \sum_{k=1}^{K} \frac{e^{u_{tk}^* + z(u_{tk} - u_{tk}^*)}}{1 + \sum_{j=1}^{K} e^{u_{tj}^* + z(u_{tj} - u_{tj}^*)}} \right) (u_{ti} - u_{ti}^*) dz \right|$$

$$\overset{(a)}{\le} g(u_t^*) + \sum_{i=1}^{K} \int_0^1 \left| \frac{e^{u_{ti}^* + z(u_{ti} - u_{ti}^*)}}{(1 + \sum_{j=1}^{K} e^{u_{tj}^* + z(u_{tj} - u_{tj}^*)})^2} \left(1 - 2 \sum_{k=1}^{K} \frac{e^{u_{tk}^* + z(u_{tk} - u_{tk}^*)}}{1 + \sum_{j=1}^{K} e^{u_{tj}^* + z(u_{tj} - u_{tj}^*)}} \right) \right| dz (u_{ti}^* - u_{ti})$$

$$\overset{(b)}{\le} g(u_t^*) + \sum_{i=1}^{K} \int_0^1 \frac{e^{u_{ti}^* + z(u_{ti} - u_{ti}^*)}}{(1 + \sum_{j=1}^{K} e^{u_{tj}^* + z(u_{tj} - u_{tj}^*)})^2} dz (u_{ti}^* - u_{ti})$$

$$\overset{(c)}{=} g(u_t^*) + \sum_{i=1}^{K} \int_0^1 \frac{e^{u_{ti} + z(u_{ti}^* - u_{ti})}}{(1 + \sum_{j=1}^{K} e^{u_{tj} + z(u_{tj}^* - u_{tj})})^2} dz (u_{ti}^* - u_{ti}), \tag{34}$$

where inequality (a) comes from the fact that $u_{ti}^* \ge u_{ti}$ for all i, (b) uses the inequality $\left| 1 - 2 \sum_{k=1}^{K} \frac{e^{u_{tk}^* + z(u_{tk} - u_{tk}^*)}}{1 + \sum_{j=1}^{K} e^{u_{tj}^* + z(u_{tj} - u_{tj}^*)}} \right| \le 1$, and equality (c) comes from a change of variable.

We will now link this last term to the regret at time t. For $u \in \mathbb{R}^K$, recall the definition:

$$Q(u) := \sum_{i=1}^{K} \frac{e^{u_i}}{1 + \sum_{j=1}^{K} e^{u_j}}$$

We can express the regret as follows:

$$R(T) = \sum_{t=1}^{T} Q(u_t^*) - Q(u_t)$$

By doing a Taylor expansion at u_t^*:

$$R(T) = \sum_{t=1}^{T} \int_0^1 \nabla Q(u_t + z(u_t^* - u_t))dz^\top (u_t^* - u_t)$$

$$= \sum_{t=1}^{T}\sum_{i=1}^{K} \int_0^1 \frac{e^{u_{ti}+z(u_{ti}^*-u_{ti})}}{1+\sum_{j=1}^{K} e^{u_{tj}+z(u_{tj}^*-u_{tj})}}\left(1 - \sum_{k=1}^{K} \frac{e^{u_{tk}+z(u_{tk}^*-u_{tk})}}{1+\sum_{j=1}^{K} e^{u_{tj}+z(u_{tj}^*-u_{tj})}}\right)dz(u_{ti}^* - u_{ti})$$

$$= \sum_{i=1}^{K} \int_0^1 \frac{e^{u_{ti}+z(u_{ti}^*-u_{ti})}}{(1+\sum_{j=1}^{K} e^{u_{tj}+z(u_{tj}^*-u_{tj})})^2}dz(u_{ti}^* - u_{ti}).$$

Noting the correspondence of this last term with the second term of (34), we can complete the proof of the lemma as follows:

$$\sum_{t=1}^{T}\sum_{j\in S_t} q_{t,j}(S_t, \theta^*)q_{t,0}(S_t, \theta^*) = \sum_{t=1}^{T} g(u_t)$$

$$\leq \sum_{t=1}^{T} g(u_t^*) + \sum_{t=1}^{T}\sum_{i=1}^{K} \int_0^1 \frac{e^{u_{ti}^*+z(u_{ti}-u_{ti}^*)}}{(1+\sum_{j=1}^{K} e^{u_{tj}^*+z(u_{tj}-u_{tj}^*)})^2}dz(u_{ti}^* - u_{ti})$$

$$= \sum_{t=1}^{T} g(u_t^*) + R(T)$$

$$= \sum_{t=1}^{T}\sum_{j\in S_t^*} q_{t,j}(S_t^*, \theta^*)q_{t,0}(S_t^*, \theta^*) + R(T)$$

$$= \sum_{t=1}^{T} \kappa_{2,t}^* + R(T). \qquad \text{(by definition of } \kappa_{2,t}^*)$$

$\square$

Proof of Lemma C.3. Let $i, k \in [K]$. We first write:

$$\frac{\partial Q}{\partial i} = \frac{e^{u_i}}{1+\sum_{j=1}^{K} e^{u_j}} - \frac{e^{u_i}(\sum_{j=1}^{K} e^{u_j})}{(1+\sum_{j=1}^{K} e^{u_j})^2}.$$

Then,

$$\frac{\partial^2 Q}{\partial i \partial k} = -\frac{e^{u_i}e^{u_k}}{(1+\sum_{j=1}^{K} e^{u_j})^2}$$

$$- \frac{[e^{u_i}e^{u_k} + \mathbf{1}_{i=k}e^{u_i}\sum_{j=1}^{K} e^{u_j}](1+\sum_{j=1}^{K} e^{u_j})^2 - e^{u_i}(\sum_{j=1}^{K} e^{u_j})2e^{u_k}(1+\sum_{j=1}^{K} e^{u_j})}{(1+\sum_{j=1}^{K} e^{u_j})^4}$$

$$= -\frac{e^{u_i}e^{u_k}}{(1+\sum_{j=1}^{K} e^{u_j})^2} - \frac{e^{u_i}e^{u_k} + \mathbf{1}_{i=k}e^{u_i}\sum_{j=1}^{K} e^{u_j}}{(1+\sum_{j=1}^{K} e^{u_j})^2} + \frac{2e^{u_i}(\sum_{j=1}^{K} e^{u_j})e^{u_k}}{(1+\sum_{j=1}^{K} e^{u_j})^3}.$$

Thus,

$$\left|\frac{\partial^2 Q}{\partial i \partial k}\right| \leq \left|\frac{e^{u_i}e^{u_k}}{(1+\sum_{j=1}^{K} e^{u_j})^2}\right| + \left|\frac{e^{u_i}e^{u_k}}{(1+\sum_{j=1}^{K} e^{u_j})^2}\right| + \left|\frac{e^{u_i}\sum_{j=1}^{K} e^{u_j}}{(1+\sum_{j=1}^{K} e^{u_j})^2}\right| + 2\left|\frac{e^{u_i}(\sum_{j=1}^{K} e^{u_j})e^{u_k}}{(1+\sum_{j=1}^{K} e^{u_j})^3}\right|$$

$$\leq 5.$$

Proof of Lemma C.4.

By the multivariate mean value theorem:

$$g_t(\theta_1) - g_t(\theta_2) = \nabla\mathcal{L}_t^{\lambda_t}(\theta_2) - \nabla\mathcal{L}_t^{\lambda_t}(\theta_1)$$

$$= \int_0^1 \nabla^2\mathcal{L}_t^{\lambda_t}(\theta_1 + z(\theta_2 - \theta_1))dz(\theta_2 - \theta_1)$$

Hence

$$\|g_t(\theta_1) - g_t(\theta_2)\|_{G_t^{-1}(\theta_1,\theta_2)} = \|\theta_1 - \theta_2\|_{G_t(\theta_1,\theta_2)} \tag{35}$$

where $G_t(\theta_1,\theta_2) := \int_0^1 \nabla^2 \mathcal{L}_t^{\lambda_t}(\theta_1 + z(\theta_2 - \theta_1))dz$.

Using Proposition A.5, we have that:

$$G_t(\theta_1,\theta_2) \succeq \frac{1}{(1+\sqrt{6K}W)} H_t(\theta_1) \tag{36}$$

As a result, by combining (36) and (35):

$$
\begin{aligned}
\|\theta_1 - \theta_2\|_{H_t(\theta_1)} &\leq (1 + \sqrt{6K}W)^{1/2}\|\theta_1 - \theta_2\|_{G_t(\theta_1,\theta_2)} \\
&= (1 + \sqrt{6K}W)^{1/2}\|g_t(\theta_1) - g_t(\theta_2)\|_{G_t^{-1}(\theta_1,\theta_2)} \\
&\leq (1 + \sqrt{6K}W)\|g_t(\theta_1) - g_t(\theta_2)\|_{H_t^{-1}(\theta_1)}.
\end{aligned}
$$

$\square$

C.3   Construction of the confidence set

In this section, we build upon the new Bernstein-like tail inequality for self-normalized vectorial martingales developed in [17] to derive a confidence set on θ^*.

Remember that:

$$C_t(\delta) := \{\theta \in \Theta \| \|g_t(\theta) - g_t(\hat{\theta}_t)\|_{H_t^{-1}(\theta)} \leq \gamma_t(\delta)\}$$

Our objective is to prove the following proposition.

Proposition 3.3 2. Let $\delta \in (0,1]$. Then $\mathbb{P}(\forall t, \theta^* \in C_t(\delta)) \geq 1 - \delta$.

We start by a few technical considerations and auxiliary lemmas. The proof of this result relies on a Bernstein concentration inequality which is a variant of the following theorem:

Theorem C.5 (Theorem 4 in [2])**.** *Let $\{\mathcal{F}_t\}_{t=1}^\infty$ be a filtration. Let $\{x_t\}_{t=1}^\infty$ be a stochastic process in $\mathcal{B}_2(d)$ such that x_t is $\mathcal{F}_t$ measurable. Let $\{\varepsilon_t\}_{t=2}^\infty$ be a martingale difference sequence such that ε_t is $\mathcal{F}_{t-1}$ measurable. Furthermore, assume that conditionally on $\mathcal{F}_t$ we have $|\varepsilon_t| \leq 1$ almost surely, and note $\sigma_t^2 = \mathbb{E}\left[\varepsilon_t^2 | \mathcal{F}_t\right]$. Let $\{\lambda_t\}_{t=1}^\infty$ be a predictable sequence of non-negative scalars. For any $t \geq 1$ define:*

$$H_t = \sum_{s=1}^{t-1} \sigma_s^2 x_s x_s^T + \lambda_t I_d, \qquad S_t = \sum_{s=1}^{t-1} \varepsilon_s x_s.$$

Then for any $\delta \in (0,1]$:

$$\mathbb{P}\left(\exists t \geq 1, \|S_t\|_{H_t^{-1}} \geq \frac{\sqrt{\lambda_t}}{2} + \frac{2}{\sqrt{\lambda_t}} \log\left(\frac{2^d \det(H_t)^{\frac{1}{2}} \lambda^{-\frac{d}{2}}}{\delta}\right)\right) \leq \delta.$$

The above Bernstein inequality is of the same flavor as Theorem 1 in [1], but is taking into account information on the local curvature of the reward function.

In our setting, we consider:

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

$$U_t = \sum_{s=1}^{t-1}\sum_{j \in S_s} \varepsilon_{s,j} x_{s,j}$$

where $\varepsilon_{s,j} = y_{s,j} - q_{s,j}(S_s,\theta^*)$.

Note that we cannot directly write H_t, U_t under the form required in Theorem C.5 since for all s, the variables $\{\varepsilon_{s,j}\}_{j \in S_s}$ are correlated. We show below that we can still prove similar concentration guarantees on $\|U_t\|_{H_t^{-1}}$.

Theorem C.6. *for any $\delta \in (0, 1]$:*

$$\mathbb{P}\left(\exists t \geq 1, \|U_t\|_{H_t^{-1}} \geq \frac{\sqrt{\lambda_t}}{4} + \frac{4}{\sqrt{\lambda_t}} \log\left(\frac{2^d \det(H_t)^{\frac{1}{2}} \lambda_t^{-\frac{d}{2}}}{\delta}\right)\right) \leq \delta.$$

Note that this expression is almost identical to the one in Theorem C.5 except for some minor constant modification.

The proof follows the same line as the proof of Theorem 4 in [2], but the analysis differs because of the non independence of the variables $\{\varepsilon_{s,j}\}_{j \in S_s}$. In particular, we analyse the behavior of the global variable $z_s := \sum_{j \in S_s} \varepsilon_{s,j} \xi^\top x_{s,j}$.

As in [2], we consider the non regularized hessian

$$\bar{H}_t = \sum_{s=1}^{t-1}\left[\sum_{i \in S_s} q_{s,i}(S_s, \theta^*) x_{s,i} x_{s,i}^T - \sum_{i \in S_s}\sum_{j \in S_s} q_{s,i}(S_s, \theta^*) q_{s,j}(S_s, \theta^*) x_{s,i} x_{s,j}^T\right]$$

and for all $\xi \in B(0, 1/2)$, we let:

$$M_0(\xi) = 1 \qquad \text{and} \qquad M_t(\xi) = exp(\xi U_t - \|\xi\|_{\bar{H}_t}^2).$$

Note that we define only $M_t(\xi)$ for $\xi \in B(0, 1/2)$ whereas $\xi \in B(0, 1)$ in [17] and [2]. In the following, we consider the filtration $\mathcal{F}_t$ engendered by $\{\{\vec{x_s}, \vec{\epsilon_s}\}_{s=1}^{t-1}, \vec{x_t}\}$. The main ingredient of the proof is to show that relatively to $\mathcal{F}_t$, $M_t(\xi)$ is still a super martingale. The rest of the proof follows immediately from [17] and [2].

To bound $\mathbb{E}\left[\exp(\xi^T U_t) | \mathcal{F}_{t-1}\right]$, we first state the following lemma, whose proof can be found in [2]:

Lemma C.7. *Let ε be a centered random variable of variance σ^2 and such that $|\varepsilon| \leq 1$ almost surely. Then for all $\lambda \in [-1, 1]$:*

$$\mathbb{E}\left[\exp(\lambda\epsilon)\right] \leq 1 + \lambda^2 \sigma^2.$$

Lemma C.8. *For all $\xi \in B(0, 1/2)$, $\{M_t(\xi)\}_{t=0}^\infty$ is a nonnegative super martingale.*

Proof. Note that for all $s \geq 1$, there is a single index $i \in S_s \cup \{0\}$ for which $y_{s,i} = 1$, and $y_{s,j} = 0$ for all $j \in S_s \cup \{0\} \setminus \{i\}$. Besides, we have $\mathbb{P}(y_{s,i} = 1) = q_{s,i}(S_s, \theta^*)$. Hence, conditional on $\mathcal{F}_s$, the variance of $\xi^\top z_s$ can be expressed as:

$$\sigma^2(\xi^\top z_s | \mathcal{F}_s)$$

$$= \mathbb{E}\left(\left(\sum_{j \in S_s}(y_{s,j} - q_{s,j}(S_s, \theta^*))\xi^\top x_{s,j}\bigg|\mathcal{F}_s\right)^2\right) - \left(\mathbb{E}\left(\sum_{j \in S_s}(y_{s,j} - q_{s,j}(S_s, \theta^*))\xi^\top x_{s,j}\bigg|\mathcal{F}_s\right)\right)^2$$

$$= \mathbb{E}\left(\left(\sum_{j \in S_s}(y_{s,j} - q_{s,j}(S_s, \theta^*))\xi^\top x_{s,j}\bigg|\mathcal{F}_s\right)^2\right)$$

$$= \mathbb{E}\left(\sum_{j \in S_s}\sum_{i \in S_s}(y_{s,j}\xi^\top x_{s,j})(y_{s,i}\xi^\top x_{s,i})\bigg|\mathcal{F}_s\right)$$

$$- 2\mathbb{E}\left(\sum_{j \in S_s} y_{s,j}\xi^\top x_{s,j}\bigg|\mathcal{F}_s\right) \cdot \left(\sum_{j \in S_s}\xi^\top x_{s,j} q_{s,j}(S_s, \theta^*)\right) + \left(\sum_{j \in S_s}\xi^\top x_{s,j} q_{s,j}(S_s, \theta^*)\right)^2$$

$$= \mathbb{E}\left(\sum_{j \in S_s} y_{s,j}(\xi^\top x_{s,j})^2\bigg|\mathcal{F}_s\right) - 2\left(\sum_{j \in S_s}\xi^\top x_{s,j} q_{s,j}(S_s, \theta^*)\right)^2 + \left(\sum_{j \in S_s}\xi^\top x_{s,j} q_{s,j}(S_s, \theta^*)\right)^2$$

$$= \sum_{j \in S_s}(\xi^\top x_{s,j})^2 q_{s,j}(S_s, \theta^*) - \left(\sum_{j \in S_s}\xi^\top x_{s,j} q_{s,j}(S_s, \theta^*)\right)^2. \tag{37}$$

Now, noting that U_{t-1} is $\mathcal{F}_{t-1}$-measurable, we have that for all $t \geq 1$:

$$\mathbb{E}\left[\exp(\xi^T U_t)|\mathcal{F}_{t-1}\right] = \exp(\xi^T U_{t-1})\mathbb{E}\left[\exp(\xi^\top z_{t-1})|\mathcal{F}_{t-1}\right].$$

Let $i \in S_{t-1} \cup \{0\}$ be the index for which $y_{t-1,j} = 1$. We have $y_{t-1,j} = 0$ for all $j \in S_{t-1} \setminus \{i\}$. If $i \in S_{t-1}$, using that $\|x_{t,j}\| \leq 1$ for all t,j and $\|\xi\| \leq 1/2$, we have the following inequality:

$$|z_{t-1}| \leq (1 - q_{t-1,i}(S_{t-1}, \theta^*))|\xi^\top x_{t-1,i}| + \sum_{j \in S_{t-1} \setminus \{i\}} q_{t-1,j}(S_{t-1}, \theta^*)|\xi^\top x_{t-1,j}|$$

$$\leq \frac{1}{2}\left(1 + \sum_{j \in S_{t-1}} q_{t-1,j}(S_{t-1}, \theta^*)\right)$$

$$\leq 1$$

Otherwise, $i = 0$ and we have that:

$$|z_{t-1}| \leq \sum_{j \in S_{t-1}} q_{t-1,j}(S_{t-1}, \theta^*)|\xi^\top x_{t-1,j}| \leq 1$$

Since $|z_{t-1}| \leq 1$, we can apply Lemma C.7 and obtain:

$$\mathbb{E}\left[\exp(\xi^T U_t)|\mathcal{F}_{t-1}\right] = \exp(\xi^T U_{t-1})\mathbb{E}\left[\exp(\xi^\top z_{t-1})|\mathcal{F}_{t-1}\right]$$
$$\leq \exp(\xi^T U_{t-1})(1 + \sigma^2(\xi^\top z_{t-1}|\mathcal{F}_{t-1})^2)$$
$$\leq \exp(\xi^T U_{t-1} + \sigma^2(\xi^\top z_{t-1})^2|\mathcal{F}_{t-1}) \qquad (1 + x \leq e^x)$$

Noting that from (37), $\sigma^2(\xi^\top z_{t-1}\big|\mathcal{F}_{t-1})^2$ is exactly equal to $\|\xi\|_{\bar{H}_t}^2 - \|\xi\|_{\bar{H}_{t-1}}^2$, it leads to:

$$\mathbb{E}\left[M_t(\xi)|\mathcal{F}_{t-1}\right]$$
$$= \mathbb{E}\left[\exp\left(\xi^T U_t - \|\xi\|_{\bar{H}_t}^2\Big|\mathcal{F}_{t-1}\right)\right]$$
$$= \mathbb{E}\left[\exp\left(\xi^T U_t\right)|\mathcal{F}_{t-1}\right]\exp\left(-\|\xi\|_{\bar{H}_t}^2\right) \qquad (\bar{H}_t \text{ is } \mathcal{F}_{t-1}\text{- measurable})$$
$$\leq \exp\left(\xi^T U_{t-1} + \sigma^2(\xi^\top z_{t-1}|\mathcal{F}_{t-1})^2 - \|\xi\|_{\bar{H}_t}^2\right)$$
$$= \exp\left(\xi^T U_{t-1} - \|\xi\|_{\bar{H}_{t-1}}^2\right)$$
$$= M_{t-1}(\xi)$$

which shows that $\{M_t(\xi)\}_{t=0}^\infty$ is a super martingale. $\qquad\square$

Proof of Theorem C.6. Using that $\{M_t(\xi)\}_{t=0}^\infty$ is a super martingale by Lemma C.8, the proof follows the proof of Theorem 4 in [2] and Theorem 1 in [17], with some minor modification since ξ now belongs to $B(0, 1/2)$ instead of $B(0, 1)$ to guarantee that $\{M_t(\xi)\}_{t=0}^\infty$ is a super martingale. In the proof of Theorem 1 in [17], for any scalar β, we now define h to be the density of an isotropic normal distribution of precision β^2 truncated on $B(0, 1/2)$ (instead of $B(0, 1)$), and g the density of the normal distribution of precision $2H_t$ truncated on the ball $B(0, 1/4)$ (instead of $B(0, 1/2)$). The upper bound on the ratio of the normalisation constants $\frac{N(g)}{N(h)}$ given by Lemma 6 of [17] remains identical, hence following [2], [17] and taking $\xi_0 = \frac{H_t^{-1}U_t}{\|U_t\|_{H_t^{-1}}}\frac{\beta}{4\sqrt{2}}$ instead of $\xi_0 = \frac{H_t^{-1}U_t}{\|U_t\|_{H_t^{-1}}}\frac{\beta}{2\sqrt{2}}$, we finally obtain that:

$$\mathbb{P}\left(\exists t \geq 1, \|U_t\|_{H_t^{-1}} \geq \frac{\sqrt{\lambda_t}}{4} + \frac{4}{\sqrt{\lambda_t}}\log\left(\frac{2^d \det(H_t)^{\frac{1}{2}}\lambda_t^{-\frac{d}{2}}}{\delta}\right)\right) \leq \delta.$$

$\qquad\square$

We are now ready to complete the proof of Proposition 3.3.

Proof of Proposition 3.3. Since, $\hat{\theta}_t$ minimizes $\mathcal{L}_t^{\lambda_t}(\theta)$, we have that $\nabla \mathcal{L}_t^{\lambda_t}(\hat{\theta}_t) = 0$. Hence,

$$g_t(\hat{\theta}_t) = \sum_{s=1}^{t-1} \sum_{j \in S_s} q_{s,j}(S_s, \theta) x_{t,j} + \lambda_t \hat{\theta}_t = \sum_{s=1}^{t-1} \sum_{j \in S_s} y_{s,j} x_{s,j}$$

As a result,

$$g_t(\hat{\theta}_t) - g_t(\theta^*) = \sum_{s=1}^{t-1} \sum_{j \in S_s} (y_{s,j} - q_{s,j}(S_s, \theta^*)) x_{s,j} - \lambda_t \theta^*$$

$$= U_t - \lambda_t \theta^*.$$

Therefore, since $\|\theta^*\| \leq W$ and $H_t(\theta^*)^{-1} \preceq \frac{1}{\lambda_t} I_d$, we get

$$\|g_t(\hat{\theta}_t) - g_t(\theta^*)\|_{H_t(\theta^*)^{-1}} \leq \|U_t\|_{H_t(\theta^*)^{-1}} + \sqrt{\lambda_t} W.$$

The proof concludes with a straightforward application of Theorem C.6, combined with the following upper bound on $\det(H_t)$ resulting from the determinant-trace inequality:

$$\det(H_t) \leq \left(\frac{\text{trace}(H_t)}{d} \right)^d \leq \left(\lambda_t + \frac{2tK}{d} \right)^d.$$

D   Numerical experiments - Comparison to the ONSP policy from [42]

In this section, we numerically compare the performance of our ONS based pricing policy for self concordant functions (ONSSC) and the ONSP policy from [42] when only a single product needs to be priced and the price sensitivity is unitary (which is the setting of ([42]), and the noise has a logistic distribution.

We study the performance of these two algorithms for different values of W, d and different distributions of the contexts $\{x_t\}$. The optimal parameter θ^* is set as $\theta^* = Y \times Z/\|Z\|$ where $Y \sim \mathcal{U}([0, W])$ and Z is sampled from a multivariate Gaussian distribution $\mathcal{N}(0, I_d)$. In the two first set of experiments, we assume that the contexts $\{x_t\}$ are generated independently at each period according to a multivariate Gaussian distribution $\mathcal{N}(0, I_d)$, then renormalized so that $\|x_t\| = 1$. In the third set of experiments, we assume that the product feature vectors $\{x_t\}$ are generated independently at each period according to a multivariate exponential distribution with scale parameter $\beta = 1$, then renormalized so that $\|x_t\| = 1$. In the fourth and last set of experiments, we consider adversarial contexts constructed similarly as in [42]: we set $d = 2$ and we divide the time horizon into $\log(T)$ epochs, such that each epoch $\mathcal{E}_t$ is constituted of time steps $\{2^{k-1}, \dots, 2^k - 1\}$. For all $t \in \mathcal{E}_k$, we then set $x_t = [0, 1]^T$ if $k \equiv 0 \ [2]$ and $x_t = [1, 0]^T$ if $k \equiv 1 \ [2]$.

The results of our experiments are displayed on Figure 1. We compare the cumulative regret obtained by the two algorithms for $T = 10^4$ steps. As W grows, the parameter γ in the policy from [42] becomes exponentially small. We observe that in this case, our policy achieves a significantly better regret in all sets of experiments. This experimentally supports our claim that avoiding to explicitly use κ in the descent step may lead to more practical algorithms.
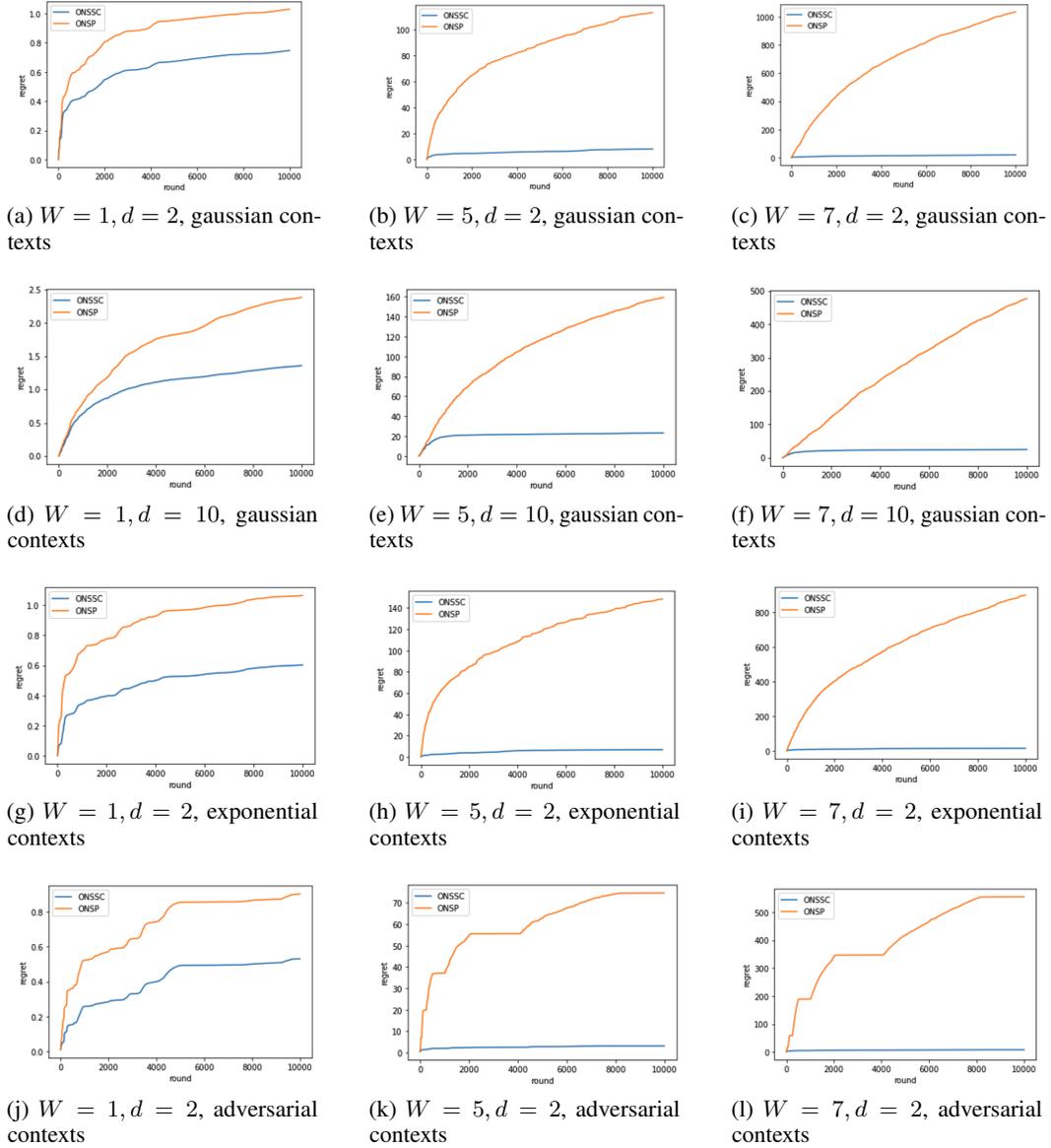
Figure 1: Cumulative regret of the ONSSC and the ONSP policies for different values of W, d and different distributions of the contexts $\{x_t\}$.