Effective Adaptive Exploration of Prices and Promotions in Choice-Based Demand Models

Lalit Jain *
University of Washington

Zhaoqi Li
University of Washington

Erfan Loghmani
University of Washington

Blake Mason
Rice University

Hema Yoganarasimhan
University of Washington

January 16, 2024

Abstract

We consider the problem of setting the optimal prices and promotions for a multi-product category when the firm lacks demand information. At each time, a customer arrives and chooses a product based on a discrete choice model where each product's utility depends on product features, its price and promotion, and the customer's features. Using a Thompson Sampling approach, we develop a regret minimizing, alternatively profit maximizing, algorithm for the retailer. We provide the first adaptive algorithm that simultaneously incorporates pricing and promotions into a discrete choice model. To make our algorithm computationally feasible over an infinite space of prices and promotions, we provide a novel method for learning the optimal price and promotion given a set of demand parameters. We also provide theoretical justification for our results and improve upon existing regret guarantees. Using simulations based on real-life grocery store data, we show that our method significantly outperforms existing approaches. In addition, we extend our methodology to a contextual setting, which allows for consumer heterogeneity and personalized pricing and promotion. Compared to existing works, our approach is agnostic to the parametric specification of the utility model and needs no assumptions on the underlying distribution of customer features.

Keywords: Demand models, pricing, optimization, bandits, Thompson sampling, dynamic pricing

*The empirical exercise contains the researcher(s)' own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researcher(s) and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein. Thanks are due to the participants of the ISMS 2022 Marketing Science Conference, SICS 2023 conference, Yale Junior Quant Marketing Conference 2022, and Kellogg Marketing Seminar 2023 for their feedback. Finally, we would like to thank Shirsho Biswas, Simha Mummalaneni, Omid Rafieian, and Frederico Rossi for their extensive comments and help that have significantly improved the paper. Please address all correspondence to: lalitj@uw.edu and hemay@uw.edu.

1 Introduction

Modern retailers typically sell a large number of brands/products in the same category. In this setting, the two key decisions that these retailers make are how to set – (1) pricing and (2) promotions/marketing mix variables, i.e., they need to decide how to price and promote each product in the category so as to maximize category profits. For example, in a digital setting, promotions could be decisions around which products to display on the retailer's main category page, which products to highlight in the search results for the category, and how much (or if at all) should a product be featured in promotional mailers/emails sent to customers. In a brick-and-mortar retail setting, the promotion variable can capture the standard display and feature variables.

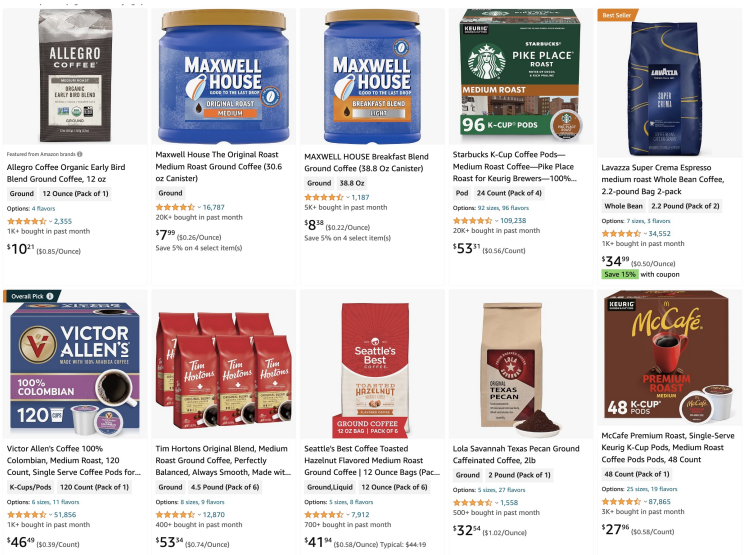


Figure 1: An example page of an online retailer selling ground coffee. The retailer must choose which prices to choose, and how to allocate promotion badges and screen space to the items.

The benefit of a digital (and/or modern retail) environment is that retailers have the ability to change prices and adjust the marketing mix in a way that is customized to each segment or individual customer in real-time. This gives them the opportunity to extract value in ways that were not previously possible. However, to leverage these technological advances, they need to be able to jointly optimize for price and promotions (and potentially personalize these decisions based on customer features) over the entire category of products. As a concrete example, consider an online retailer (e.g., Amazon) selling coffee products as shown in Figure 1. There are several types of promotions on this page – (1) the smaller "Featured from Amazon Brands", and the larger "Overall Pick" and "Best Seller" badges, and 2) several price promotions, e.g. "Save 15%" and "Typical $44.19". Thus, besides just setting prices, the retailer has chosen which items to promote.

In principle, if the retailer has perfect information on the underlying demand model for each consumer-product combination in a category, it could solve a joint-optimization problem to find the optimal price and promotion vectors that maximize profit. However, in practice, firms lack such demand information.[1]

[1]This could be due to several reasons such as entrance into a new category, entrance into new markets, the addition of new products in the category, changes in demand and consumer preferences due to macroeconomic factors (Huang et al., 2022; Hitsch et al., 2021). Note that in order to optimize category-level profits, the retailer needs a good understanding of not only price and promotion elasticities for each brand, but also that of cross-price and cross-promotion elasticities, to capture substitution effects within the

The two standard ways to learn the demand model are to use observational data or to run A/B tests, and then use these data to estimate a category-level demand model (that may incorporate consumer attributes). However, both of these approaches are problematic for different reasons. The former can lead to misleading estimates unless the data satisfy certain exclusion restrictions, which are hard to verify/ensure in many observational settings.[2] Further, this approach does not work when the retailer has new categories, if we have new products in existing categories, or if the demand parameters change. The latter approach does not suffer from endogeneity problems but can be costly because it requires the firm to run A/B tests across a wide range of price and promotion vectors, many of which would be far from the profit-maximizing prices and promotions. Such experimentation can become exponentially costly with a large number of brands involved and the data requirements necessary for inference can be challenging. Further, from a practical perspective, fully randomizing prices and/or promotions is often practically infeasible since managers are reluctant to randomly assign prices/promotions to a large batch of data since consumers may react adversely to large differences in prices due to fairness concerns (Priester et al., 2020).

In this paper, we present a solution to the problem of joint optimization of prices and promotions for a category with multiple products, with unknown demand functions. We first discuss our approach for the setting where consumers are homogeneous or the firm lacks consumer/segment-level features. Later we will expand our approach to allow for consumer- and market-level heterogeneity i.e., consider a setting where the firm has data on market characteristics, consumers' demographics, past purchases, or browsing behavior, that it can leverage to customize prices and promotions.

Our solution builds on two canonical but separate streams of literature – (1) empirical demand estimation in industrial organization (IO) and marketing (Berry and Haile, 2021), and (2) Multi-Arm Bandits, and in particular Thompson sampling (Russo et al., 2018). Our solution concept and algorithm have two components. First, we start with the standard multinomial demand model, commonly employed in empirical demand estimation (McFadden, 1980). We assume that consumers have latent utilities over each product which depend on price, promotion, and product features, and choose the product that maximizes their overall utility. With GEV (generalized extreme value) shocks, this gives rise to analytical choice probabilities.

Under this paradigm, we first derive the firm's optimal price and promotion vector for a given parameter vector, which has not previously been considered in the literature. In general, finding the global optima of the profit function is non-trivial since the profit function is not convex. Past work has considered the problem of finding the optimal price without promotions, and an approximate solution can be effectively found using a binary search procedure (Hanson and Martin, 1996). However, to our knowledge, there is no general procedure to find both the optimal price *and promotion*. To ensure that the promotion decisions are realistic, we consider two types of constraints on promotion variables – box and simplex constraints – that map to the common constraints faced by retailers when making promotion/advertising decisions. Given these constraints, we develop a novel procedure to simultaneously derive the optimal promotions and prices. Effectively, we show that the optimal choice of promotions is restrained to a finite set; so for each of these promotions, we can learn the optimal price, and then choose the combination of promotion and price that has the highest revenue.

category.

[2] A large literature in economics and marketing has focused on demand estimation to address the endogeneity issues prevalent in observational data. We refer readers to Berry (1994) and Berry and Haile (2021) for a comprehensive discussion.

The above procedure works if the firm knows the true parameters of the demand model. However, this is not true in practice. Therefore, the second component of our solution concept focuses on how to simultaneously learn the demand parameters (i.e., *exploring* the parameter space) without incurring a large experimentation cost (i.e., *exploiting* the information learned so far to price/promote effectively). Balancing exploration vs. exploitation tradeoffs, often referred to as the multi-armed bandit framework (Lattimore and Szepesvári, 2020), is a well-considered area in machine learning and adaptive experimentation. We employ one of the most popular Bayesian-based methods, namely *Thompson or Posterior Sampling* (Russo et al., 2018) for our solution. However, naively implementing Thompson sampling would discretize the price–promotion space, and for each price and promotion, we would need to maintain a posterior distribution over the possible demand values at that point. In this case, Thompson sampling explores by sampling a demand for each potential price from the corresponding posterior and then presents a customer with the price–promotion combination with the highest observed demand based on previous data. It then observes the item that the customer chooses to purchase and updates its posterior distribution.

This approach of discretization is problematic for multiple reasons. Firstly, the set of potential price–promotion combinations that need to be considered can be extremely large; exponentially increasing in category size.[3] Secondly and implicitly, the demand function should be smooth in prices and promotions and hence information acquired about the demand function at any one point should inform points around it. Naively estimating the demand function at every single point separately will ignore this structure completely. In addition, since we have no information about price-promotion combinations that have not yet been measured, the algorithm will potentially spend a long time exploring every single combination before it starts optimizing.

To overcome these obstacles, we develop a Thompson sampling approach that builds upon the structure of the economic discrete choice model discussed earlier. Instead of doing Thompson sampling on top of individual price–promotion pairs, we maintain a posterior distribution over the model parameters themselves. As a result, each draw from the posterior corresponds to a realization of a demand curve, and the associated price–promotion we choose maximizes this sampled demand curve (where the maximization is done based on the procedure described earlier).

We now present some results on the theoretical properties and empirical performance of our method. First, as is common in the adaptive experimentation literature (Lattimore and Szepesvári, 2020), we prove a theoretical *regret bound* for our algorithm. The regret measures the loss in total profit due to playing sub-optimal prices and promotions due to exploring price-promotion space. We show that our Thompson Sampling approach has a regret that scales like $O(K\sqrt{\kappa T})$, where T is the time horizon, K is the number of products, and κ is the inverse of the worst-case elasticity of demand. Intuitively, the dependence on κ measures how easy it is to learn about the underlying parameter vectors when we vary prices and promotions. If the elasticity is very small, demand is not very responsive to price or promotion changes which makes learning the underlying parameters more challenging. This in turn, will increase our regret by slowing down the effectiveness of exploration.

Secondly, we present a comprehensive empirical evaluation of our approach based on parameter estimates

[3]For example, if there are twenty different products, and twenty different prices (at .50 increments between \$10 and \$20), there are already 200 arms/combinations to consider – which is prohibitive in any multi-armed bandit even without promotions.

from the NielsenIQ Retail Measurement Services (RMS) data for the ground coffee category. For this exercise, we use data from the largest store in King County, WA that has data on weekly brand-level purchases, prices, feature, and display. This category has 10 major brands (including an *Other* brand encompassing many smaller brands). We specify a latent utility demand model at the user level, aggregate it to the brand level, and then estimate the underlying demand parameters using the standard Berry-inversion (Berry, 1994; Berry et al., 1995). Using these estimates, we perform a series of counterfactual simulations, where we assume that the retailer does not know the demand parameters and instead uses our approach to set prices and promotions.

We show that our method significantly outperforms a series of benchmarks. Firstly we consider a firm that is using a pure exploitation strategy. Namely, they estimate demand in each time period and then play prices–promotions according to the optimal value of this estimated demand curve. We show that our algorithm outperforms this *greedy* benchmark. We also compare to the M3P algorithm of Javanmard et al. (2020) (only designed for the setting of prices without promotions), where the firm forces exploration by playing random prices for a proportion of the total time and see similar results.

In the second part of the paper, we extend the baseline homogeneous model to allow for settings where the retailer has data on customer and/or market-specific variables that can be informative of consumer preferences and demand. Prior research has shown that while retailers can benefit significantly by customizing prices based on store- or customer-level demographics, there is not much user/store-level customization in prices and promotions (Hoch et al., 1995; DellaVigna and Gentzkow, 2019; Hitsch et al., 2021). Many theories have been suggested for the lack of customization including managerial inertia, lack of tools and/or the ability to differentiate and learn parameters at more granular levels, and brand image/fairness concerns. We propose an adaptive contextual framework that can help alleviate the former two concerns since they allow the retailer to automate the learning and customization of prices and promotions using adaptive sampling methodologies.

In our framework, at each time period, the retailer observes real-time customer and/or market features. Note that customer- and market-level features can be fairly general – they can be fine-grained and encapsulate demographic features or behavioral browsing/purchase history often available to retailers, or they can be coarse and represent the segment-level information rather than individual-level features. We incorporate these contextual features in our model by allowing the demand parameters to vary as a function of the context vector. We then develop a methodology to quantify the demand parameters as neural networks of the context and extend our Thompson sampling approach to this setting.

We provide both novel theoretical results on regret bounds as well as extensive experimental benchmarks for our contextual Thompson Sampling algorithm. First, in the setting where the demand parameters are known to be linear functions of the context, we are able to theoretically establish a regret of $O(dK\sqrt{\kappa T})$ where d is the context dimension. We then perform a series of simulations where we consider two types of contextual settings – (1) Orthogonal groups, where there each user belongs to an orthogonal segment and the number of contexts is finite, and (2) Weighted average, where users are a weighted combination of orthogonal contexts. The former can be interpreted as a setting where there is no information sharing between segments and the latter is a setting where the population is very heterogeneous (and the number of contexts is infinite). In each of these cases, we allow the retailer to learn the demand parameters using the neural network without assuming that it is a linear function of the context. In both cases, our algorithm outperforms benchmark

approaches. We extend the simulations to a setting based on the NielsenIQ data for two stores from King County WA. Similar to the homogeneous case, first we estimate demand parameters as functions of both store and seasonality features (as context vectors), and then use parameters based on these estimates to run simulations. Again, we show that our approach does well even when the distribution of the context vector changes (e.g., due to seasonality). Finally, we consider a setting where demand parameters are non-linear and unknown functions of the context vector and show that our approach outperforms Greedy and M3P benchmarks. We also present some comparisons of different types of neural network architectures and their relative performance in the non-linear setting.

Overall, our contextual Thompson Sampling approach is quite general and works for a wide range of scenarios commonly faced by marketers and requires no specific assumptions on the form of the demand parameters as functions of the context (user or market features). It works irrespective of – (1) whether the underlying demand parameters are linear in contexts or non-linear, (2) whether the context distribution is known ahead of time or not, and (3) whether the context distribution remains constant or changes over time.

In summary, our paper provides the following main contributions to the literature on adaptive experimentation and demand estimation. We provide the first adaptive algorithm which: (1) sets prices and promotions for a multi-product category, (2) utilizes a multinomial choice model to drive model-based exploration, (3) incorporates customer- and market-level heterogeneity with minimal assumptions on user features and demand parameters. As part of accomplishing these goals, we provide a novel Thompson Sampling method to simultaneously optimize for pricing and promotions which minimizes regret (or maximizes profits).In addition, we give both strong theoretical guarantees and empirical validation based on both synthetic and real-life inspired data sets. From a managerial perspective, our solution is easy to deploy, can flexibly handle batched updates [4], and provides computationally efficient model updates. As such, it can be directly adopted by retailers to set customized prices and promotions at the consumer/segment/market-level and optimize profits for categories with a large number of brands.

2 Related Literature

Our work relates to two broad streams of literature – (1) the adaptive pricing work in operations research and computer science literature (also referred to as dynamic pricing in this literature), and (2) the structural demand estimation literature in economics and marketing.

2.1 Adaptive Pricing

The work on adaptive pricing is extensive and the papers in this area consider the following common setup. In the most common case, there is one retailer/seller who sells a single product for a fixed number of rounds. In each round, the seller first chooses a price for the product and then observes an associated demand. This demand is either a discrete variable representing a single purchase by a single customer, or a continuous variable representing aggregate expected demand over a population. Finally, the seller receives a revenue, which is simply the price times the observed demand. In general, the seller must learn the demand function through price exploration and the goal is to bound the *regret* of the pricing policy, namely the total loss in profit due to playing sub-optimal prices. The hope (in stochastic settings) is to establish pricing policies that

[4] i.e. model update after a batch of customer decisions rather than after each one

suffer no more than $O(\sqrt{T})$ regret where T is the time horizon over which the game is played.

Table 1 summarizes existing work on adaptive pricing on the following dimensions:

- *Single or Multi Product*: Unsurprisingly, the complexity of adaptive pricing in the case of multiple products can be much higher than in the case of a single product. Firstly, in the setting of multiple products, depending on the choice of the model class, the practitioner may have to estimate more parameters. For example, if a linear model is assumed, and there are K products, then K^2 cross-elasticities need to be estimated, and regret will scale with K^2. However, if a multinomial model is assumed, then the cross-elasticities are implicit and do not need to be explicitly estimated. Secondly, in non-parametric settings, discretizing prices is practically impossible. If there are K products, the size of a discretization of the set of prices will grow exponentially in K.
- *Type of Demand model*: In general, all demand functions arise from underlying user-level choice models. However, the form of the demand function is up to the practitioner. Commonly used demand model classes assume that demand is linear, log-linear, or logistic/multinomial function of the prices.
- *Parametric or non-Parametric*: The model classes mentioned above are all parametric demand models. Non-parametric estimation of demand functions has also been considered in the literature. To make this feasible, other assumptions are introduced. For example, the demand function may be assumed to be α-Hölder smooth and monotonic or be drawn from a bounded Gaussian Process with a known kernel.
- *Continuous or Discrete Prices:* In the non-parametric setting, it's common to consider a discrete, finite, and fixed set of prices. This is problematic for several reasons. Firstly, without assuming any structure on the underlying demand function, information about one price may not inform the demand at other prices. This could lead to a setting where the retailer has to experiment at each price to learn the demand curve. Secondly, the regret obtained scales with the number of prices considered which can force over-exploration and be extremely costly to the firm. Thirdly, if the discretization is not sufficiently fine relative to the time horizon, the approximation error can lead to linear regret (Kleinberg and Leighton, 2003).
- *Model-based exploration*: As described above, adaptive pricing depends on the framework of exploration v.s. exploitation. Many of the works in the pricing literature employ "forced/random exploration" methods to ensure sufficient exploration. These methods ensure that the underlying model parameters are sufficiently well learned by experimenting with sufficiently diverse sets of prices. In practice, forced exploration methods are not managerially feasible since they require the firm to set effectively random prices for sufficiently long periods of time. In this work, we introduce the notion of "model-based exploration" which drives exploration intelligently by leveraging gathered information and the underlying model structure. As we will see in our experiments, model-based exploration leads to a far less erratic pricing schedule, which is both managerially appealing and leads to significantly less regret.
- *Marketing Mix:* In practice, firms can drive customer preferences by not just manipulating prices, but also introducing marketing-mix variables such as promotions (e.g., display, feature) and advertising. As far as we know, our model is the first to introduce marketing-mix explicitly in adaptive pricing literature. As we will see in Section 4.1, optimizing marketing-mix variables introduces a new set of challenges.
- *Customer Heterogeneity and Market Features:* Consumers' decisions are heterogeneous across a population and demand models should incorporate this heterogeneity to customize prices and promotions across

segments/individual users. In addition, pricing decisions should capture time-varying market conditions.

Paper	Multi-Product	Multinomial Demand model	Parametric	Continuous Prices	Model-Based Exploration	Marketing Mix	Consumer, Market Features
Our Method	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Weaver and Kumar (2022)	No	No	No	No	Yes	No	No
Miao and Chao (2021)	Yes	Yes	Yes	Yes	No	No	No
Ban and Keskin (2021)	No	Yes	Yes	Yes	No	No	Yes
Wang et al. (2021)	No	No	No	Yes	Yes	No	No
Xu and Wang (2021)	No	Yes	Yes	Yes	No	No	Yes
Bastani et al. (2021)	Yes	No	Yes	Yes	Yes	No	Yes
Javanmard et al. (2020)	Yes	Yes	Yes	Yes	No	No	No
Misra et al. (2019)	No	No	No	No	No	No	No
Mueller et al. (2018)	Yes	No	Yes	Yes	Yes	No	No
Ferreira et al. (2018)	Yes	No	Yes	No	No	No	No
Trovo et al. (2018)	No	No	No	No	Yes(UCB)	No	No
Ganti et al. (2018)	Yes	Yes	Yes	Yes	Yes	No	Yes
Cheung et al. (2017)	No	No	No	Yes	Yes	No	No
Javanmard (2017)	No	Yes	Yes	Yes	No	No	Yes
Javanmard and Nazerzadeh (2016)	No	Yes	Yes	Yes	No	No	Yes
Qiang and Bayati (2016)	No	No	Yes	Yes	Yes	No	Yes
Besbes and Zeevi (2015)	No	No	No	Yes	No	No	No
Besbes and Zeevi (2009)	No	Yes	Both	Yes	No	No	No
Broder and Rusmevichientong (2012)	No	Yes	Yes	Yes	No	No	No
Keskin and Zeevi (2014)	Yes	No	Yes	Yes	Yes	No	No
den Boer and Zwart (2014)	No	Yes	Yes	Yes	Some	No	No
Kleinberg and Leighton (2003)	No	No	No	Yes	No	No	No

Table 1: Prior Literature on Adaptive Pricing

As we can see from Table 1, this is the first paper to provide a unified algorithmic framework that addresses all of these dimensions. We now discuss the most salient of these works here while providing a more extensive description of the works listed in Table 1 and their contributions in Appendix A.

In particular, Javanmard et al. (2020); Miao and Chao (2021) are the closest to our work. Firstly Javanmard et al. (2020), also assumes a multinomial response model, but unlike our setting where we assumed a fixed set of items each round, they assume that the set of items changes each round and that the seller receives covariate information for each item. Their proposed M3P algorithm for pricing in this setting alternates between rounds of playing random prices and then greedily playing the optimal price according to the estimated demand model at that time. The second work, Miao and Chao (2021) proposes an algorithm that runs in cycles, with each cycle restarting when the outside option is selected. During a cycle, a hybrid Thompson Sampling and price shock method chooses the fixed prices played for the whole round. It is not clear how to extend their method to incorporate promotions, or for batched settings. All of these works also rely on a degree of forced exploration, which is managerially impractical in real-life settings.

Three other works in the multi-product setting use Thompson sampling for price exploration, namely Ganti et al. (2018); Ferreira et al. (2018); Bastani et al. (2021). The first considers a stylized demand model that does not take product cross-elasticities into account and therefore may fail to properly capture demand in

settings with many products. The second considers a discrete set of prices and focuses on the problem of effective inventory management. The final paper considers the case of a linear demand model and focuses on the problem of prior misspecification where the definition of regret is with respect to an algorithm that has a correctly specified prior. They provide a result for Thompson Sampling in the linear setting but rely on various assumptions on the context distribution. None of these works consider the setting of multiple products with promotions and none of them extend to a setting with non-linear utility.

2.2 Structural Demand Estimation and Price, Promotion Customization

Our paper also relates to the large literature on structural demand estimation in economics and marketing. Starting with the early work by McFadden et al. (1977) and Guadagni and Little (1983), researchers have proposed and estimated structural demand models based on multinomial choice models. These models have been shown to perform well empirically and have the ability to produce credible counterfactuals (Berry and Haile, 2021). A key feature of these models is that they capture the substitution patterns within a product category reasonably well, without the need to explicitly model cross-price elasticities as in the case of linear or log-linear demand models (Train, 2009). In this paper, we build this literature to develop a structural demand model and consider extensions where we allow the model parameters to vary with user features/demographics in a non-linear and high-dimensional way and learn this function using a neural network.

Further, our work also relates to the marketing literature on the customization of prices and promotions using empirical demand models. Starting with the seminal work by Rossi et al. (1996), marketers have leveraged individual-level heterogeneity and purchase histories to customize and target pricing and/or promotions such as coupons (Zhang and Krishnamurthi, 2004; Pancras and Sudhir, 2007; Zhang and Wedel, 2009; Johnson et al., 2013; Howell et al., 2016; Bradlow et al., 2017; Smith et al., 2022). Typically, these papers use a Bayesian approach to modeling and estimation and then evaluate counterfactual targeting policies. Recent work has also considered alternative demand specifications such as deep learning models (Gabel and Timoshenko, 2022; Dubé and Misra, 2023; Wei and Jiang, 2022) and trees (Feldman et al., 2022). The consensus from this literature is that customizing prices and promotions based on user demographics and behavioral purchase data can significantly increase profits. The main difference between our work and this literature is that in our case, we consider a setting where the firm can strategically explore prices and promotions to maximize profits. In contrast, the earlier literature either relies on observational data, which may not have sufficient variation to arrive at optimal prices/promotions or uses a fully randomized experiment, which can be extremely costly from a profit perspective.

Overall, our modeling approach leverages the strengths of the structural demand estimation literature, such as employing a parsimonious demand model that does not require us to estimate cross-elasticities, incorporating user demographics/features, and the ability to produce realistic demand curves. We combine these merits with the adaptive pricing literature to develop a unified framework for effective experimentation along the demand curve to optimize for profits.

3 Choice Based Demand Model

We first consider a setting where consumers have homogeneous preferences and as a result demand parameters are constant across the population. In Section 6 we will extend this model to heterogeneous preferences.

3.1 Problem Formulation

Consider an e-commerce firm or retailer (e.g., Amazon, Instacart, WayFair) that stocks multiple products in each category, e.g. the setting of Figure 1. In each period, the firm has to decide how to price each of these products as well as how to promote them. These promotions could include:

- Feature and Display Promotions: For example, highlighting a product on a search ranking, or the front page of an e-commerce website, placing products on end-of-the-aisle displays in physical grocery stores, or promoting products in mailers/emails sent to customers. In all of these settings, there is a fixed number of items that can be highlighted due to screen or physical space constraints.
- Advertising: If the retailer has a fixed advertising budget they must spend, they need to decide what products to spend the budget on, and how to allocate this budget across items.
- Price Promotions: These are constituted of two parts, a discount on the price along with a featuring aspect (e.g., highlighting using "on sale" or "red stickers"). Thus, from the retailer's perspective, they are just setting prices on items along with deciding which items to feature to consumers and highlight the discount.

A common paradigm in the demand estimation literature in economics and marketing is to *assume* that the firm knows the parameters of the underlying demand function and has optimally chosen the prices to maximize some measure of category profits (Berry et al., 1995; Kadiyali et al., 2001; Sudhir, 2001). However, this may not always be true in practice, especially since today's retailers deal with a large number of products in dynamic market conditions, where many new products are added to the inventory regularly. Our goal is to solve the problem from the perspective of a firm that does not know the true demand parameters but has to choose prices and promotions to optimize category profits over a time horizon.

Formally, suppose that a firm has a set of K products (in a given category) that it offers to a buyer in each of T rounds.[5] In each round $t, 1 \leq t \leq T$, the firm has to choose a price and a marketing mix variable (e.g., promotion) for each of the K products in the category. Let the price vector be $\mathbf{p}_t = (p_{1t}, \cdots, p_{Kt}) \in \mathcal{P} := [\ell, u]^K$ where $\ell, u \in \mathbb{R}_{\geq 0}$ are lower and upper bounds on the prices respectively. The marketing mix allocation is $\mathbf{x}_t := (x_{1t}, \cdots, x_{Kt}) \in \mathcal{X} \subset \mathbb{R}^K$, where the possible set of marketing mix allocations is $\mathcal{X}$. In Section 4.1, we will more precisely discuss the nature of these promotions. In general, we will see that without loss of generality we can take $\mathcal{X}$ to be finite. The buyer then chooses an item $I_t \in \{0, 1, \cdots, K\}$ (with the index 0 corresponding to a *no-purchase* option). Furthermore, we assume that for each item i there is an associated marginal cost, $m_i \geq 0$, for the seller. Thus, in round t, the seller receives reward/profit r_t equivalent to $p_{I_t} - m_{I_t}$ if $I_t \in [K]$ and otherwise receives a reward of 0.

We assume that the seller's policy is *adaptive*. That is, $\mathbf{p}_t, \mathbf{x}_t$ is determined by the history of prices and promotions selected up to time t, $\{(\mathbf{p}_s, \mathbf{x}_s, I_s, r_s)\}_{s=1}^{t-1}$. Consistent with the large literature on demand estimation, we assume that the probability that the user chooses item i is determined by a *multinomial random utility model* (McFadden, 1980). That is, there is a parameter vector $\theta = [\alpha_1, \cdots, \alpha_K, \beta_1, \cdots, \beta_K, \gamma_1, \cdots, \gamma_K] \in \mathbb{R}^K \times \mathbb{R}^K_{>0} \times \mathbb{R}^K_{>0}$ so that the utility that the consumer receives

[5] Assuming a market size or a mass of buyers does not change any theoretical results. In our numerical experiments, we work with batches of potential buyers.

from purchasing product i given a price p and marketing mix variable x is:

$$U(p_{it}, x_{it}) = \alpha_i - \beta_i p_{it} + \gamma_i x_{it} + \epsilon_{it}, \tag{1}$$

where ϵ_{it} is a random shock that follows a Type-1 extreme value distribution. We allow consumers' price sensitivity (β) and response to promotions (γ) to vary by product. From a purely theoretical perspective, one may expect consumers' (dis)utility from changes in prices and promotions to be independent of the product. However, this may not be true in practice and there is empirical evidence that they may vary by product e.g. Luo et al. (2007); therefore, we adopt a fully general specification. In our algorithm, it is trivial to set the βs and γs to be constant across all the products if we believe that they should be the same.

Thus, the probability that the buyer buys item i in period t given by:

$$\mathbb{P}_\theta(I_t = i | \mathbf{p}_t, \mathbf{x}_t, \{(\mathbf{p}_s, \mathbf{x}_s, I_s, r_s)\}_{s=1}^{t-1}) = \frac{\exp(\alpha_i - \beta_i p_{it} + \gamma_i x_{it})}{1 + \sum_{k=1}^{K} \exp(\alpha_k - \beta_k p_{kt} + \gamma_k x_{kt})} \tag{2}$$

where $\mathbb{P}_\theta, \mathbb{E}_\theta$ are the probability law and expectation induced by the filtration with parameter θ.[6]Therefore, we can define the expected profit in round t as:

$$R_\theta(\mathbf{p}_t, \mathbf{x}_t) = \mathbb{E}_\theta[r_t | \mathbf{p}_t, \mathbf{x}_t, \{(\mathbf{p}_s, \mathbf{x}_s, I_s, r_s)\}_{s=1}^{t-1}] = \sum_{i=1}^{K} \frac{(p_{it} - m_i) \exp(\alpha_i - \beta_i p_{it} + \gamma_i x_{it})}{1 + \sum_{k=1}^{K} \exp(\alpha_k - \beta_k p_{kt} + \gamma_k x_{kt})} \tag{3}$$

Note that the promotions considered are assumed to have fixed costs, and as such do not affect the profit equation. This formulation is consistent with most types of promotion in standard digital environments (e.g., which products to feature on the category page, which products to highlight in search rankings or emails) which typically do not have marginal costs. In Section 4.1, we provide a more detailed discussion of the types of promotions and the constraints on the promotion variables considered. Finally, we assume that our parameters are bounded, $\alpha_i \in [-M, M]$, $\beta_i \in [0, M]$, and $\gamma_i \in [0, M]$ for a known $M > 0$. The positivity assumption on β captures the fact that our utility should decrease in price, and the assumption on γ captures our assumption that advertising/promotion has a positive effect on utility. We define the constraint region on the parameters as $\Omega = [-M, M]^K \times [0, M]^{2K}$.

3.2 Firm's goal

The goal of the retailer/seller is to design an algorithm that effectively learns the vector of optimal prices and marketing mix variables for the set of K products. To balance the trade-off between exploration and exploitation, the seller attempts to minimize their *expected regret*, which is the cumulative difference between their expected profit from using the optimal price and marketing mix variable versus the policy that they play. To be precise, for any fixed set of demand parameters, θ, we define the optimal profit-maximizing price and

[6]Note that we allow both price and marketing mix variables such as display and feature to directly enter the latent utility or choice probability, which is the standard practice in the literature (Rossi et al., 1996; Pancras and Sudhir, 2007; Hitsch et al., 2021). It is possible to identify consideration sets using exclusion restrictions, i.e., by restricting a subset of these variables to influence consideration but not utility, e.g., price affects consideration but not purchase conditional on consideration (Goeree, 2008). Our approach can be easily extended to such two-step models as long as the researcher is willing to make such exclusion assumptions.

marketing mix allocation as:

$$\mathbf{p}_*, \mathbf{x}_* := \arg \max_{\mathbf{p}\in\mathbb{R}, \mathbf{x}\in\mathcal{X}} R_\theta(\mathbf{p}, \mathbf{x}). \tag{4}$$

Without loss of generality we will assume that $\mathbf{p}_* \in \mathcal{P} = [\ell, u]^K$.

We take a Bayesian perspective and assume that there exists a prior distribution Π_0 and that $\theta \sim \Pi_0$. The prior can be totally uninformative, i.e. $\Pi_0 \sim N(0, I)$, or can encapsulate any previous information or background knowledge the firm has. We will not place any restrictions on the prior. Our goal is to minimize the total loss in profits due to the exploration of the pricing and promotion space. In the bandit literature, this is known as *regret minimization*. The time T Bayesian-*cumulative regret* is defined as

$$\text{Reg}_T^B := \mathbb{E}_{\theta\sim\Pi_0}\left[\sum_{t=1}^{T} R_\theta(\mathbf{p}_*, \mathbf{x}_*) - R_\theta(\mathbf{p}_t, \mathbf{x}_t)\right] = \mathbb{E}_{\theta\sim\Pi_0}\left[TR_\theta(\mathbf{p}_*, \mathbf{x}_*) - \sum_{t=1}^{T} R_\theta(\mathbf{p}_t, \mathbf{x}_t)\right]. \tag{5}$$

We will casually refer to the Bayesian regret through the rest of our work as the regret.[7] We also consider the *simple regret*. The simple regret at time t is defined as the gap between the optimal revenue and the revenue obtained at time t:

$$sr_t(\theta) := R_\theta(\mathbf{p}_\star, \mathbf{x}_\star) - R_\theta(\mathbf{p}_t, \mathbf{x}_t). \tag{6}$$

Simple regret measures the loss in profit due to the exploration strategy of the learner in a given time period. At any fixed time, we can hope for a policy for which the simple regret is low. However, just considering the simple regret may not be sufficient. Indeed, a learner could randomly explore prices for a long time, incurring a large profit loss before any time for which their simple regret is low. This is why we mostly study cumulative Bayesian regret.

As mentioned above, the regret represents the total profit loss of the firm over T rounds of the game. A successful algorithm will be able to ensure that $\lim_{T\to\infty} \text{Reg}_T^B(\theta)/T = 0$, that is $\text{Reg}_T^B(\theta)$ grows at a rate which is much smaller than the time horizon. In the next section, we will design an algorithm for which $\text{Reg}_T^B = O(\sqrt{T})$. To summarize, the firm's problem can be defined as follows.

Problem Statement: Play a sequence of price-promotion combinations $\{(\mathbf{p}_t, \mathbf{x}_t)\}$ to minimize the expected Bayesian cumulative regret Reg_T^B.

4 Solution Concept

To solve the problem statement in Section 3.2 we focus on two components. Firstly, assuming that we have a true demand model, we discuss how to compute the optimal price and promotion variables. Secondly, we discuss our exploration strategy and demonstrate that it achieves $O(K\sqrt{T})$-regret.

4.1 Characterizing the Optimal Price and Promotion

Our goal is to learn a policy that is playing close to the optimal price and marketing mix allocation:

$$\mathbf{p}_*, \mathbf{x}_* \in \arg \max_{\mathbf{p}\in\mathcal{P}, \mathbf{x}\in\mathcal{X}} R_\theta(\mathbf{p}, \mathbf{x}) \tag{7}$$

[7] We quickly point out that bounds on the Bayesian regret are weaker than bounds on the cumulative regret, $\sum_{t=1}^{T} R_\theta(\mathbf{p}_*, \mathbf{x}_*) - R_\theta(\mathbf{p}_t, \mathbf{x}_t)$. Indeed, a bound on the latter (for any θ) provides a bound on the former, but not vice-versa. For a more in-depth discussion, please see Lattimore and Szepesvári (2020).

In this section, we characterize this optimum. In general, the profit function $R_\theta(\mathbf{p}, \mathbf{x})$ is challenging to optimize since it is not strictly concave or quasi-concave in $(\mathbf{p}, \mathbf{x})$. Therefore, to solve for the optimal $(\mathbf{p}_*, \mathbf{x}_*)$, we adopt a two-step approach. Note that we can rewrite the optimization problem of Equation 7 as

$$\arg \max_{\mathbf{p}\in\mathcal{P}, \mathbf{x}\in\mathcal{X}} R_\theta(\mathbf{p}, \mathbf{x}) = \arg \max_{\mathbf{x}\in\mathcal{X}} \arg \max_{\mathbf{p}\in\mathcal{P}} R_\theta(\mathbf{p}, \mathbf{x})$$

Motivated by this, we first derive the optimal price vector given a fixed $\mathbf{x} \in \mathcal{X}$. Next, we show that only a finite set of promotion vectors can be optimal in our setting - i.e. we can assume $\mathcal{X}$ is finite. This allows us to search for the optimal price and marketing mix combination over a finite set of $\mathbf{x} \in \mathcal{X}$ in the second step. We describe our approach in detail below.

Characterizing the Optimal Price Vector for a Fixed Promotion Vector First, we can show that, for a fixed value of $\mathbf{x}$, there exists a unique global optimum characterized by a fixed point equation.

Lemma 1. *For a fixed value of* $\mathbf{x}$*, the optimal price* $\mathbf{p}_*(\mathbf{x}) := \arg\min_{\mathbf{p}\in\mathbb{R}^K_{\geq 0}} R_\theta(\mathbf{p}, \mathbf{x})$ *satisfies,* $\mathbf{p}_{*,i} = \frac{1}{\beta_i} + R + m_i$ *where* R *is the solution of the fixed point equation*

$$R = \sum_{i=1}^{K} \frac{1}{\beta_i} e^{-(1+\beta_i R+\beta_i m_i)} e^{\alpha_i+\gamma_i x_i} \tag{8}$$

Proof: See Web Appendix B.1.

Lemma 1 implies that we can obtain the optimal price vector (conditional on $\mathbf{x}$ if we can solve for R in Equation (8). Notice that the function on the right-hand side of Equation (8) is strictly decreasing in R, while the left-hand side is increasing. So their difference is monotonic and crosses zero. This immediately gives rise to an algorithm for finding the optimum through binary search on the revenue as the solution to a fixed point problem. In Web Appendix B.2, we present the pseudocode of the binary search algorithm that we use to solve for R in Equation (8). We quickly remark that this approach is not unique and has been observed previously (Aydin and Ryan, 2000; Li and Huh, 2011).[8]

Characterizing the Optimal Promotion Vector for a Fixed Price Vector In many settings, we can think of the promotion variables as coming from a fixed and finite discrete set $\mathcal{X} \subset \{0, 1\}^K$. For example in the case of Figure 1, we could choose a single item to get the "Amazon's Choice" badge. This would correspond to the setting where $\mathcal{X} = \{\mathbf{e}_1, \cdots, \mathbf{e}_K\}$ where $\mathbf{e}_i$ is the i-th standard basis vector; or in other words - the item with the badge corresponds to a promotion variable of 1, and the rest of the products have a promotion variable set to 0. As another example, we could perhaps choose a subset of variables that we choose to promote with "Best Seller" badges. Analogously, this could correspond to a setting with $\mathcal{X} = \{0, 1\}^K$.

However, neither of these discrete settings appropriately captures the *magnitude* of the promotion. As a result, we may want to relax the discrete settings considered above to allow continuous promotion variables under a budget constraint. First, we consider a *simplex* constraint such that $\mathcal{X} = \triangle_K^B = \{\mathbf{x} \in \mathbb{R}^K_{\geq 0} : \sum_{i=1}^K x_i = B\}$. This constraint maps to settings where the firm has a total budget (B) for the marketing mix

[8]Different approaches towards optimization have been taken by Dong et al. (2009) who notice that the revenue is concave as a function of the underlying market shares, and Hanson and Martin (1996) who uses a pathfinding approach.

variable (e.g., advertising spend, promotional spend, or a fixed amount of real-estate space on the front page of a website/mailer) and it has to choose how much of this budget to allocate to each product. Second, we consider a *box* constraint, which implies that each element of the marketing mix vector (x_i) lies within a range $[0, B]$. This constraint captures many different practical settings, e.g., where the marketing mix variable can represent which products are featured in the store or promoted in online/offline displays and by how much. In these cases, B represents the upper bound on how much a single product can be featured/promoted. Note that the box constraint represents a setting where our promotion budget is limited on each item, but we do not have to use it all up, whereas the simplex constraint is a setting where our promotion budget must be fully used up each round.

Theorem 1. *For a fixed $\mathbf{p} \in \mathcal{P}$ consider the problem $\mathbf{x}_*(\mathbf{p}) := \arg\max_{x \in \mathcal{X}} R(\mathbf{p}, \mathbf{x})$. Assume that $p_i \neq p_j$ for some $i, j \in [K]$ and that $\gamma_i > 0$ for all $i \in [K]$. Then $R(\mathbf{p}, \mathbf{x})$, as a function of $\mathbf{x}$, has no critical points in the interior of $\mathcal{B}$. In addition,*

1. *if $\mathcal{X} = [0, B]^K$, $\mathbf{x}_*(\mathbf{p}) \in \{0, B\}^K$;*
2. *if $\mathcal{X} = \triangle_K^B$, $\mathbf{x}_*(\mathbf{p}) \in \{B\mathbf{e}_i : i \in [K]\}$ where $\mathbf{e}_i$ is the i-th standard basis vector in $\mathbb{R}^K$.*

Proof: See Web Appendix B.3.

This result says that, for any given price vector, the optimal promotion is always a corner solution for both types of constraints. In the simplex case, this means that the firm should put all of its marketing budget on one of the products. In the case of the box constraint, it suggests that for each product the firm should completely maximize the amount of promotion or do no promotion at all. To gain some intuition for this result, imagine the case where $\mathbf{p}$ is fixed and $B \to \infty$. In this setting, assuming all the $\gamma_i > 0$, to maximize profits, we should place as much promotion as possible on the product with the highest price and 0 promotion on all the other items. Thus, intuitively, our profit is maximized at a promotion choice that guarantees the highest possible price. The formal proof of this result in the appendix makes this result precise by showing that the profit function has no critical points with respect to the promotion variables $\mathbf{x}$. However, for a finite B, the optimal promotion may not necessarily be a full allocation to the product with the highest price. We refer readers to Web Appendix B.4 for a detailed example and further intuition.

Theorem 1 is an important and fundamental result about optimal promotions in choice models that we have not seen previously in the literature and may be of independent value to choosing optimal promotions. As briefly described previously, it also greatly aids our algorithm. Essentially, it constrains the set of $\mathbf{x}$ vectors that we have to consider when solving for the optimal $\mathbf{p}$. As such, from an algorithmic perspective, this theorem makes the solution computationally feasible. For instance, in the case of a simplex constraint, to solve for the optimal price and marketing mix combination, we simply need to consider K possible vectors for $\mathbf{x}_*$, solve for the optimal price at each of these possibilities using Equation (8), and then pick the combination of $(\mathbf{p}, \mathbf{x})$ that maximizes profit. Note that even in this case, the number of times we have to solve for optimal prices using binary search is linear in K. Nevertheless, it is still finite and feasible to solve for a large assortment of products.

An attractive feature of our solution concept is that it is quite general and allows us to easily consider other settings not represented by the box or simplex constraint. For instance, suppose that the retailer has a set

of rank-ordered spots (say on the front page of their website) where they can feature a subset (say $K' \leq K$) of the products, and they have to decide which products to feature in each spot. This setting can be expressed as a case where there are a set of finitely many promotion vectors which we need to consider when searching for the optimal price. In summary, any setting that can be mathematically represented using a finite set of $\mathcal{X}$ vectors is compatible with our proposed solution concept.

We remark that our solution also provides a link to the literature on selection assortment (Agrawal et al., 2019; Miao and Chao, 2021). There, the goal of the firm is to construct a choice set of K' items, where K' may be much less than K, for the user to choose from. This approach may be overly constraining – in practice, users could consider a much larger set of items than we present; as a result, choosing K' may be difficult. Nevertheless, retailers may want to select a subset of K' items that are promoted and account for those promotions in the customer's choice decision. Theorem 1, provides a solution to find a good subset of items to promote while not making any assumptions about the users' choice behavior over that subset.

Finally, note that our solution concept allows the firm to optimize marketing mix variables. However, in many cases, these variables may be outside the control of the firm (e.g., national advertising set by the manufacturer, coupons provided by the manufacturer, or promotions run directly by the manufacturer). In such cases, the marketing mix variable can be treated as a contextual variable i.e., product or customer feature that is given. We consider this alternative setting in detail in Section 6.

4.2 Thompson Sampling for Adaptive Pricing and Marketing Mix Allocation

Now that we have an algorithm to solve for the optimal $(\mathbf{p}_*, \mathbf{x}_*)$ given a parameter vector, we return to the main problem of how to dynamically choose prices and marketing allocations recognizing that exploration is costly but can lead to better decisions in the future. To this end, we propose the following Thompson Sampling approach for a general multinomial demand model that includes both prices and marketing mix variables.

Algorithm 1 Thompson Sampling for Multinomial Demand Model

Input: $\mathcal{P} \times \mathcal{X}$, Π_0 - prior distribution for θ
  for $t = 0, 1, 2, \cdots$, **do**
    Sample $\widetilde{\theta}_t \sim \Pi_t$
    Set $\mathbf{p}_t, \mathbf{x}_t = \arg\max_{\mathbf{p}\in\mathcal{P},\mathbf{x}\in\mathcal{X}} R_{\widetilde{\theta}_t}(\mathbf{p}, \mathbf{x})$
    Observe I_t and $r_t := p_{I_t} - m_{I_t}$
    Update $\Pi_{t+1} \leftarrow \mathbb{P}(\theta \in \cdot | \mathcal{H}_t)$
  end for

Thompson Sampling is a popular framework to navigate explore vs. exploit tradeoff problems in sequential decision making (Chapelle and Li, 2011; Russo et al., 2018). In Thompson sampling, the learner assumes a prior distribution over the unknown problem parameters (θ in our case). Mapping parameters to their optimal actions ($\mathcal{P} \times \mathcal{X}$ in our case) induces a density over the space of each possible action. Thompson sampling then plays an action drawn from this resulting distribution. Thompson sampling has a rich history in the multi-armed bandits literature, and we refer the reader to the survey Russo et al. (2018) for more details. In the marketing literature, Thompson sampling-based methods have been applied to a variety of settings successfully, including ad recommendations (Schwartz et al., 2017; Aramayo et al., 2022). More broadly, adaptive experimentation techniques/bandits for website design and learning consumer preferences have been

considered in the marketing literature; please see Hauser et al. (2009) and Liberali and Ferecatu (2022).

In practice, there are two challenges to implementing Thompson Sampling in this setting. Firstly, given a posterior distribution over the parameters, computing the density over actions is a potentially in-tractable multi-dimensional integral. Though this might be possible for finitely many actions (Russo et al., 2018), in our case the problem is further compounded since our action space $\mathcal{P} \times \mathcal{X}$ is infinite. In practice, *posterior sampling* avoids the need to compute this integral. Posterior sampling uses the realization that sampling from the induced density on actions is equivalent to sampling $\widetilde{\theta}_t$ from the posterior distribution Π_t, computing the price, and marketing mix variable assuming that $\widetilde{\theta}_t$ is the true parameter and then playing these actions.

The second main challenge of implementing Thompson Sampling is computing the posterior update Π_{t+1}. In general, there is no closed-form expression for Π_{t+1} in the multinomial setting we describe. Hence, we have to resort to techniques that do approximate Posterior sampling. We consider two methods, firstly Laplace approximation (Chapelle and Li, 2011; Russo et al., 2018; Kveton et al., 2020), and secondly the MCMC method of Langevin Dynamics.

Laplace Approximation: To motivate this, consider a regularized log-likelihood function,

$$\mathcal{L}_t(\{(\mathbf{p}_s, \mathbf{x}_s, I_s)\}_{s=1}^t, \theta) = \sum_{s=1}^{t} \log(\mathbb{P}_\theta(I_s|\mathbf{p}_s, \mathbf{x}_s)) + \frac{\lambda}{2}\|\theta\|_2^2$$

and the Maximum Likelihood Estimate (MLE)

$$\widehat{\theta}_t = \arg\min_{\theta \in \Theta} -\mathcal{L}_t(\{(\mathbf{p}_s, \mathbf{x}_s, I_s)\}_{s=1}^t, \theta). \tag{9}$$

Here λ is our regularization parameter and $\|\cdot\|_2$ is the L^2 norm. By the Delta Method (Casella and Berger, 2021), asymptotically in distribution $\sqrt{t}(\hat{\theta} - \theta_*) \overset{t\to\infty}{\to} N(0, V_t^{-1})$ where $V_t = \nabla^2_\theta \mathcal{L}_t(\{(\mathbf{p}_s, \mathbf{x}_s, I_s)\}_{s=1}^t, \theta)$. This allows us to approximate the posterior distribution as a multivariate normal distribution $\Pi_t \approx N(\hat{\theta}_t, V_t^{-1})$. We sample from this approximate posterior in Thompson sampling in lieu of the true posterior.

Langevin Dynamics When the true posterior is not well approximated by a Gaussian, which may be true after a small number of samples, Thompson sampling using Langevin Dynamics has been shown to achieve smaller regret empirically compared to a Laplace approximation (Russo et al., 2018; Mazumdar et al., 2020; Xu et al., 2022). Langevin dynamics, also referred to as Langevin Monte Carlo in some works, is a Markov Chain Monte Carlo (MCMC) method whose stationary distribution is the true posterior. Updates can be performed via gradient descent on the negative log-likelihood of the posterior plus injected Gaussian noise (Roberts and Tweedie, 1996; Bakry et al., 2014). In the limit, Langevin dynamics produces exact samples according to the posterior Welling and Teh (2011). As we will see in Section 7.3, Langevin dynamics can also generalize to non-linear models such as neural networks allowing us to apply Thompson sampling to contextual settings. Our precise formulation of Thompson Sampling with Langevin dynamics is given in Algorithm 2.

Langevin dynamics is fairly straightforward to implement. In each round, Langevin dynamics consists of N steps of gradient descent with additional noise whose variance is proportional to the product of the learning rate η_t with the inverse of the temperature parameter. In practice, the learning rate is chosen to either be constant or decay as $1/t$. The inverse temperature parameter allows us to scale the variance of the noise

Algorithm 2 Langevin-based Thompson Sampling with Multinomial Demand Model

Input: $\mathcal{P} \times \mathcal{X}$, Π_0 - prior distribution for θ, step sizes $\{\eta_t\}_{t\geq 1}$, inverse temperature parameters $\{\psi_t\}_{t\geq 1}$

$\theta_{1,0} \leftarrow [0, \ldots, 0] \in \mathbb{R}^K$

for $t = 1, 2, \cdots,$ **do**

  $\theta_{t,0} = \theta_{t-1,N}$

  for $n = 1, \ldots, N$ **do**

    Sample $\epsilon_{t,n} \sim \mathcal{N}(0, \mathbf{I})$

    $\theta_{t,n} \leftarrow \theta_{t,n-1} - \eta_t \nabla \mathcal{L}_t(\{(\mathbf{p}_s, \mathbf{x}_s, I_s)\}_{s=1}^t, \theta_{t,n-1}) + \sqrt{2\eta_t \psi_t^{-1}}\epsilon_{t,n}$

  end for

  Set $\mathbf{p}_t, \mathbf{x}_t = \arg\max_{\mathbf{p}\in\mathcal{P}, \mathbf{x}\in\mathcal{X}} R_{\theta_{t,N}}(\mathbf{p}, \mathbf{x})$

  Observe I_t and $r_t := p_{I_t} - m_{I_t}$

end for

independently from the learning rate. It is normally set to a constant.

4.3 Theoretical Guarantees

Our main result of this section is the following guarantee bound on the regret of our Thompson sampling algorithm for dynamic pricing and promotion. To begin, define the demand function,

$$\mu_\ell(\theta, \mathbf{p}, \mathbf{x}) := \mathbb{P}_\theta(I_t = i | \mathbf{p}_t, \mathbf{x}_t, \mathcal{H}_{t-1}) = \frac{\exp(\alpha_i - \beta_i p_{it} + \gamma_i x_{it})}{1 + \sum_{k=1}^K \exp(\alpha_k - \beta_k p_{kt} + \gamma_k x_{kt})},$$

to denote the probability that item i is purchased. A critical problem parameter is

$\kappa := \frac{1}{\min_{\theta\sim\Pi_0, \overline{p}, i\in[K]} \mu_i(\theta,\mathbf{p},\mathbf{x})(1-\mu_i(\theta,\mathbf{p},\mathbf{x}))}$ where $\mu_i(\theta, \mathbf{p}, \mathbf{x})(1 - \mu_i(\theta, \mathbf{p}, \mathbf{x}))$ can be interpreted the elasticity of the demand function along the i^{th} direction and captures the difficulty of learning $\alpha_i, \beta_i, \gamma_i$. Hence, κ represents a pessimistic lower bound on the elasticity of demand valid for any set of parameters θ and any price and marketing mix vectors $\mathbf{p}, \mathbf{x}$.

The full statement and proof of the following theorem are in Appendix C. We now present a simplified result that captures the key dependencies on the problem parameters.

Theorem 2. *Assume that* $\max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_\infty \leq B$*, the largest price is bounded by* u*, and* $\|\theta\| \leq S$*. Then the regret of Algorithm 1 is bounded in expectation by* $\widetilde{O}\left(SuK\sqrt{\kappa T}\right)$ *where* $\widetilde{O}(\cdot)$ *consists of constants and doubly logarithmic factors.*

Proof: See Appendix C.

As far as we know, this is the first theoretical guarantee for Thompson Sampling in the setting with both prices and promotions. The result implies that the Bayes regret of the algorithm is of order $\tilde{O}(SuK\sqrt{\kappa T})$. The dependence on κ intuitively implies that if the demand function is particularly "flat" for some parameter vector, then that underlying parameter is more difficult to learn which may impact our regret.

Remark. This matches existing results due to Javanmard et al. (2020) in the case of pricing only (without promotions). The algorithm of Miao and Chao (2021) is able to achieve a Bayesian regret of $O(\sqrt{KT})$. Their paper focuses on the problem of simultaneously choosing an optimal assortment and pricing. Their method is a hybrid of Thompson sampling and forced exploration and does not take promotion variables

into account. In contrast, our method uses model-based exploration (as discussed in Section 2, can handle promotions, and extends to a contextual setting as we will see in Section 6. In general, Thompson sampling is known to have an optimal regret in terms of T but potentially be loose in terms of the underlying parameter dimension. We quickly remark that in the traditional multi-armed bandit setting with finitely many arms the "gap", i.e. difference in mean rewards between any arm and the best arm, arms governs the regret of the underlying problem, (e.g. see Chapter 2 of Lattimore and Szepesvári (2020)). The advantage of these "gap-based" or "instance-dependent" regret bounds is that they provide a logarithmic dependence on T and look like $O(c_{gap}\log(T))$ where c_{gap} is a constant dependent on gaps and can be significantly tighter than the standard $O(\sqrt{T})$ bound. However, in our setting, since we are considering a continuum of possible prices, it's not clear that the notion of "gap" even makes sense. In addition, we suspect that a $O(\log(T))$ bound may not even be possible in general. Indeed Theorem 5.5 of Javanmard et al. (2020) provides an instance in a multinomial pricing setting where any pricing policy must incur $O(\sqrt{T})$ regret.

For more on this topic, we refer the reader to (Abeille and Lazaric, 2017; Hamidi and Bayati, 2020). We also believe that the dependence on κ may not be necessary. Recent work in the bandits literature has removed the dependence on κ for UCB style algorithms (Faury et al., 2020). It is an open problem to remove it in the case of Thompson sampling. Finally, we remark that most of the considerations in this remark are theoretical. In particular, they do not impact the resulting algorithm - just the analysis. This is very unlike other methods such as ILS (Keskin and Zeevi, 2014) or UCB (Faury et al., 2020), where the amount of exploration or the form of confidence bounds is very tied to the underlying problem statement and the theoretical guarantees desired. This plug-and-play nature of Thompson sampling lends to its popularity in practice. As we will see in the next section, our Thompson Sampling approach has excellent empirical performance in our setting.

5 Numerical Experiments

We now present a series of simulations that validate our proposed method. In Section 5.1, we consider a stylized setting to ground ideas and explain the advantage of exploration. Next, in Section 5.2, we follow up with experiments that demonstrate the performance of our method using parameter estimates from a real-life setting that consists of retail purchase data on a commonly purchased product category (ground coffee). Even though this is not a digital retailer that can change prices in real-time, the parameter estimates from this data are representative of the range of true values we may see in a real-life setting providing additional validity to our simulations.

5.1 Motivating Price and Promotion Exploration

Consider a retailer in a market with three products ($K = 3$), who is deciding how to set prices and promotions for these products. In each period t, the firm can set prices in the range $p_{it} \in [0, \$30.00] \ \ \forall \ \ 1 \leq i \leq K$ and choose a binary promotion variable x_{it} that satisfies a simplex constraint, i.e. $\sum_{i=1}^{3} x_{it} = 1, x_{it} \in \{0, 1\} \ \ \forall \ \ t \geq 1$. More simply said, the firm can choose to promote at most one of the products or none at all (following the results from Theorem 1).

The demand for each product follows the choice model outlined in Section 3.1, the parameters for the demand parameters are shown in the top panel of Table 2, and the marginal cost is set to zero. We also record the optimal price and promotion for each product and the market share for each product under the

Parameters	Product 1	Product 2	Product 3
α	1	1	1
β	.1	.2	.3
γ	.8	.3	.5
$\mathbf{p}_*$	\$20.50	\$15.50	\$13.83
$\mathbf{x}_*$	1	0	0

Table 2: Parameters and outcomes at the optimal pricing/promotion for a three-product case. Optimal revenue $R(\mathbf{p}_*, \mathbf{x}_*) = \$\mathbf{10.5}$.

set of optimal prices/promotions. Then, the optimal revenue of the firm, i.e., the revenue if it plays prices and promotions optimally, is \$10.50. We now compare Thompson Sampling in the next two sections to two baselines that existing firms may use, namely a "Greedy" strategy and an "Explore-then-commit" approach.

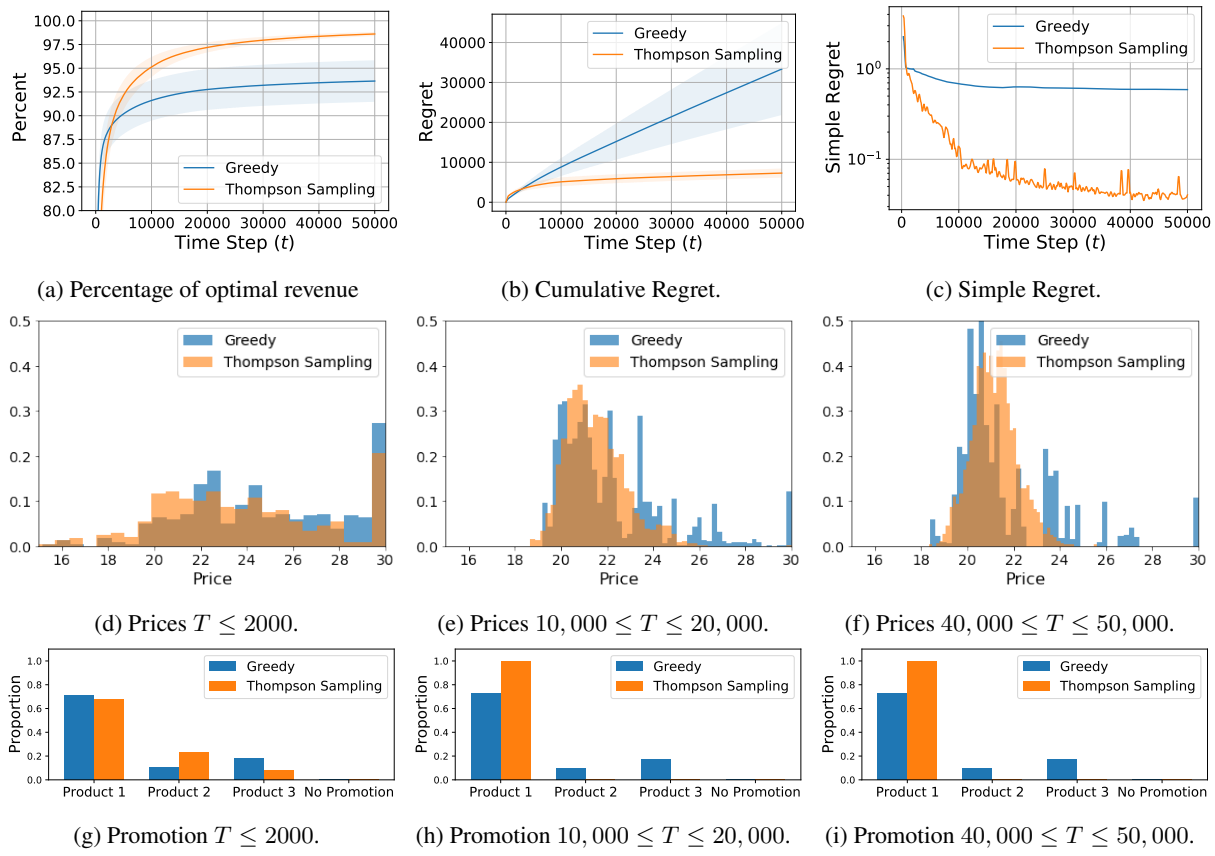
5.1.1 Adaptive Exploration vs. No exploration

We now consider two different strategies that the firm could play – (1) Thompson Sampling (TS) and (2) Greedy. The first strategy uses the approach sketched out in the previous section. However, Greedy is a more myopic approach: under this strategy, in any given period, the firm estimates the parameters of the demand model based on the prices/promotions and corresponding market shares (choice outcomes) observed up to that point. Then, the price and promotion variables are chosen by optimizing the revenue associated with this empirically estimated demand model. Thus, Greedy does not actively explore new prices or marketing mix allocations; rather it simply exploits the observed data (more details on Greedy in Section 5.2.2).

We run 40 replications of a Monte Carlo simulation with 50,000 purchase decisions each, average the results over these replications, and present the results from this exercise in Figure 2. The optimal revenue at time t that the firm would have earned playing the optimal price and promotion is $\$10.5 \times t$. In Figure 2a, we plot the percentage of the optimal revenue that the firm recovers under each strategy at any given time period t. We see that there is roughly a 6% difference between the two methods, after $t = 50,000$ periods/purchases. At face value a 6% profit loss may not seem large; however, this loss is being observed *at each time*. To make this clearer, we plot the cumulative regret of both methods in Figure 2b. Practically, this is the total revenue lost by each method over the time horizon. Notice that TS has more or less constant regret, matching, or perhaps exceeding the guarantees of Theorem 2. In contrast, the regret of Greedy appears linear. To understand why this is the case, in Figure 2c we plot the simple regret for both methods, that is the loss in revenue at each period. The simple regret of TS is an order of magnitude smaller than that of Greedy. In addition, whereas the simple regret flatlines for Greedy, the exploration of TS allows it to continuously improve its simple regret. This discussion highlights the main managerial takeaway of this work – the exploration of Thompson Sampling allows it to more effectively play the optimal prices and promotions.

In the second row of Figure 2, we dive a bit deeper into the prices played by each method. We show the distribution of prices played for the first product over the 40 runs during the first 2000 time steps, then from step 10,000 to 20,000, and then from time step 40,000 to 50,000. As can be seen, initially both Thompson Sampling and Greedy play fairly uniformly over the acceptable price range[9]. However, we already start to see a difference in behavior between time steps 10,000 and 40,000 samples. The distribution of prices

[9]We discuss implementation details more thoroughly in the next section, but if an algorithm chooses to play a price which is larger than \$30 (the maximum acceptable price in this simulation), we round it down to \$30.



(a) Percentage of optimal revenue

(b) Cumulative Regret.

(c) Simple Regret.

(d) Prices $T \leq 2000$.

(e) Prices $10,000 \leq T \leq 20,000$.

(f) Prices $40,000 \leq T \leq 50,000$.

(g) Promotion $T \leq 2000$.

(h) Promotion $10,000 \leq T \leq 20,000$.

(i) Promotion $40,000 \leq T \leq 50,000$.

Figure 2: We compare the performance of Thompson Sampling and a myopic Greedy baseline for 50000 purchase decisions in a setting with three products over 40 simulations. Each line in the top row represents the average over the 40 simulations and 95% confidence regions on the average are shown. In the second row, we plot the played price distributions for product 1 for the three time horizons (for each run). Finally, in the last row, we show the proportion of time that we promoted each product.

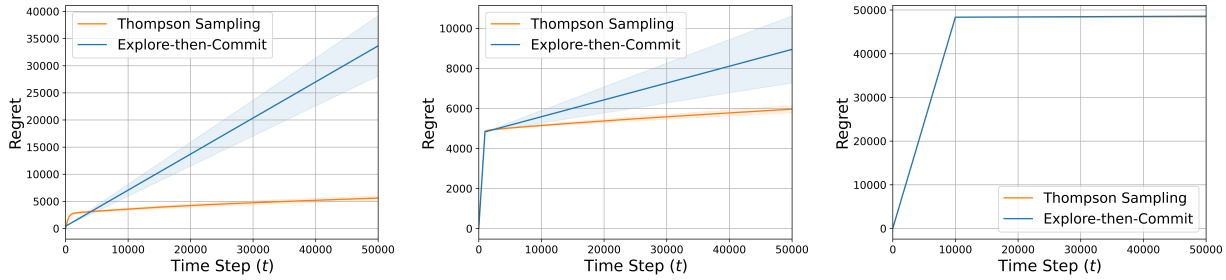
Thompson sampling plays is centered around the optimal price of \$20.50, whereas the distribution of prices that Greedy plays seems to have several peaks, and does not vary smoothly. This phenomenon is exacerbated by time 40,000 – instead of recovering, Greedy doubles down on its bad pricing choices in Figure 2f including noticeable spikes at prices close to \$24, and \$30. On the other hand, the distribution of Thompson sampling is just more peaked around the optimal price. A similar phenomenon happens in the case of promotions. In the last row of the figure, we see that TS chooses to consistently promote Product 1 by 10000 time steps. However, the distribution of products promoted by Greedy does not change in any real way over the whole run. Greedy chooses what to promote early on, and then just sticks with it throughout. In Appendix G, we delve deeper into the significance of determining the optimal promotion strategy by conducting a comparative analysis. Specifically, we assess the performance of Thompson Sampling in contrast to its variants, where the promotion component is intentionally disabled.

Effectively, the lack of strategic exploration in prices for Greedy renders it unable to escape sub-optimal prices and promotions. As a result, the data it gathers is not as informative. We emphasize that the histograms in Figure 2 are not of a specific replicate, but rather accumulate the data from all 40 replicates of the simulation. There is a chance that any given replicate may behave fine, but the average behavior is worse.

5.1.2 Adaptive Exploration vs Fixed Random Exploration

An alternative approach frequently employed by firms to determine an optimal set of prices and promotions is initially experiment on random prices and subsequently fix their pricing and promotional strategies based on the findings of that experiment. In this section, we explore the viability of such a strategy and assess the potential losses incurred by adopting this approach. This strategy is commonly referred to in the experimentation literature as "Explore-then-Commit." We undertake a comparative analysis, pitting the proposed Thompson Sampling method against the Explore-then-Commit algorithm while varying the number of experimentation steps $\tau_{explore}$. For experiments in this section, we use the setting from Section 5.1.

Figure 3 presents the cumulative regret results. When the initial experimentation steps $\tau_{explore}$ are relatively small (Figure 3a), even after the initial exploratory phase, the Explore-then-Commit approach remains far from attaining optimal pricing and promotion strategies, resulting in a linear regret thereafter. However, Thompson Sampling demonstrates its ability to achieve modest regrets, even with as few as ten initial exploratory samples. Conversely, with a substantially larger initial exploration steps (Figure 3c), Explore-then-Commit and Thompson Sampling perform similarly. However, the prolonged exploratory phase leads to a cumulative regret of $50,000 - much larger than with any of the other smaller exploration periods. Overall, even in this very simple example, using Thompson sampling provides significant wins over a strategy that uses a fixed model, even after a large amount of price exploration using completely random prices.



(a) $\tau_{explore} = 100$ (b) $\tau_{explore} = 1000$ (c) $\tau_{explore} = 10000$

Figure 3: A comparison of Thompson Sampling and the Explore-then-Commit algorithm, illustrating cumulative regret for different numbers of initial exploration phase samples.

5.2 Simulations based on Nielsen Data

Next, we present simulation results based on data from NielsenIQ Retail Measurement Services. In Section 5.2.1, we briefly describe the dataset and parameter estimation, and then in Sections 5.2.2 and 5.2.4, we describe our simulation framework and results.

5.2.1 Parameter Estimation

We focus on the ground coffee category for our empirical exercise due to a few nice features. First, it is a regularly purchased product with a sufficiently large volume. Secondly, it has a number of well-known brands and we expect significant differences in brand preferences and price sensitivities across brands. Finally, because the product sizes are in ounces, we are able to pool data across all UPCs within a brand, obtain the price per ounce for each brand, and perform the estimation exercise at the brand level.

We use NielsenIQ weekly retail scanner data for 2019. The data set contains weekly prices, quantity, and product characteristics for all the ground coffee products sold across a number of stores in the US. The data are aggregated at the store-week-UPC level, i.e., each observation consists of a product UPC code, the brand and size (ounces) of the product, and whether the product was featured/displayed that week. For our analysis, we use data from the largest store in King County, WA that has data on prices, as well as feature and display information. We start by selecting top 9 brands out of 50 that had the most revenue in that store ($\sim 90\%$ of revenue) and group all the remaining brands under an *Other* brand. This gives us a total of 10 brands and one outside option (no purchase). Next, using weekly sales and average price data for each UPC, we obtain the average weekly price per ounce (p_{it}) and the average weekly feature and display variables (w_{it} and o_{it}), as well as the total quantity sold (in ounces) for each of the 10 brands. Table A2 in Appendix H.1 shows the summary statistics of the weekly prices, feature, display, and sales for the 10 brands we consider. We see that the top five brands in this market (based on total volume over the one-year period) are Peet's Coffee, Starbucks, Seattle's Best, Stumptown, and Tony's Coffee, which together account for approximately 70% of the sales. There is also significant across-brand variation in the feature and display variables. While products under brands such as Peets and Starbucks are featured and displayed quite often, many of the brands in the data are never featured or displayed (e.g., Zoka, Lavazza). Overall, we see both within- and across-brand variation in prices, and sufficient across brand-variation in the promotion variables.

We specify the latent utility model at the brand level as:

$$U_i(p_{it}, w_{it}, o_{it}) = \alpha_i - \beta_i p_{it} + \gamma_w w_{it} + \gamma_o o_{it} + \gamma_t \mathbb{I}(week = t) + \epsilon_{it}.$$

We allow the price sensitivity parameters to vary across brands but only estimate one parameter each for feature and display on the whole data instead of having brand-level parameters. This is because the data for feature/display promotions is sparse and many brands are never featured/displayed. We also add a week-fixed effect since we see significant seasonality in coffee purchases (e.g., coffee is more popular in Fall/Winter but less so in the hot summer months). The above specification implies that the probability that the buyer purchases brand i in period t is:

$$\mathbb{P}_\theta(I_t = i | \mathbf{p}_t, \mathbf{w}_t, \mathbf{o}_t, \mathcal{H}_{t-1}) = \frac{\exp(\alpha_i - \beta_i p_{it} + \gamma_w w_{it} + \gamma_o o_{it} + \gamma_t \mathbb{I}(week = t))}{1 + \sum_{k=1}^{K} \exp(\alpha_k - \beta_k p_{kt} + \gamma_w w_{kt} + \gamma_o o_{kt} + \gamma_t \mathbb{I}(week = t))}.$$

Following Berry (1994), we then specify the outside good as the decision to not buy ground coffee (no category purchase), take the log of the market/volume share (in ounces) for each brand, and subtract from it the log of the outside option share.[10] This transformation gives us a linear relationship between the log market share ratio of the brands and brand dummies, brand prices, marketing mix variables, and week fixed effects:

$$\log(Q_{i,t}) - \log(Q_{outside,t}) = \alpha_i - \beta_i p_{it} + \gamma_w w_{it} + \gamma_o o_{it} + \gamma_t \mathbb{I}(week = t), \tag{10}$$

where $Q_{i,t}$ is the market share (in volume) for brand i in week t and $Q_{outside,t}$ is outside option market share.

We then estimate Equation (10) using a least squares model and present the parameter estimates in Table

[10] We use 1.2 times the maximum weekly observed demand (in ounces) during the one-year observation period as the total market size. This allows us to quantify the share of outside option in each week.

A3 in Web Appendix H.2. We find significant differences in the brand-specific intercept terms as well as the price sensitivity parameter across brands. Among the top five brands, consumers are most sensitive to the price of Seattle's Best and least sensitive to that of Stumptown and Tony's Coffee. Next, we see that coefficient for feature ($\gamma_{feature}$) is insignificant and only the display coefficient is significant. Therefore, in Section 5.2.4, we only consider display in our analysis. We ignore the week fixed effects as well. In Section 6 we will extend our model to handle time-varying settings. Overall, all the parameter estimates seem reasonable and form the basis of our simulation results in Section 5.2.4.

5.2.2 Algorithms Compared

We now describe the algorithms that we consider – (1) Thompson Sampling, (2) Greedy method, and (3) M3P based on Javanmard et al. (2020). Each method uses the MLE, Equation 9, to create a corresponding estimate of the demand model. The main difference is in how the estimate is utilized to explore prices. Precise descriptions of these algorithms and parameter selection procedures are given in Web Appendices E, H.6.

Thompson Sampling We considered two variants of Thompson Sampling as described in Section 4.2 – posterior sampling using Laplace approximation and Langevin dynamics. In the case of Laplace approximation, at each round, we sample a price from $N(0, aV_t^{-1})$, where V_t is the Hessian of the MLE up to time t and a is a tunable exploration bonus which we set as $a = .5$. For Langevin dynamics, we set the number of Langevin steps it takes as $N = 50$. We also set the learning rate to be $\eta_t = .01/t$ and $\psi_t = \psi = .5$.

Greedy, the simplest algorithm, initially plays some random prices for a period $\tau_{explore}$. At each time after this, it uses the MLE to estimate the model based on the data up to that point and then plays the optimal price and marketing mix according to the estimated model. Thus, Greedy does not actively explore new prices or marketing mix allocations; rather it simply exploits the observed data. The main problem with this approach is that it could get stuck in a sub-optimal region since new data points don't necessarily add information about the parameters. In general, a greedy policy may lead to an estimation strategy that is not $O(T^{-1/2})$ consistent, and as a result, could lead us to incur a linear regret (Keskin and Zeevi, 2014). This is precisely the phenomenon we observed in Figure 2b.

M3P Javanmard et al. (2020) method builds on greedy and allows for more systematic exploration. The method divides the time horizon into a series of blocks with increasing length. In block b, the method first does random forced exploration, where uniformly random prices from $[\ell, u]$ are played, for K periods (where K is the number of products). This is followed by b periods of exploitation. Namely, similar to Greedy, the MLE is estimated using all data up to that point and the optimal price and promotion for the resulting estimated demand model is played. Thus there are a total of $b + K$ time steps in block b. Since the method increases the number of exploitation rounds in each iteration, (in fact in a time horizon of T it achieves $O(\sqrt{T})$ exploration steps), it is able to achieve a $O(K\sqrt{T} + K)$ regret (Javanmard et al., 2020). Even though M3P does not consider the setting of costs and promotions, and has no theoretical guarantees for this setting, modifying it for our setting is fairly straightforward and we include it in our simulations. Our modifications are discussed in Web Appendix E. For both the Greedy and M3P methods, to learn the MLE we utilize gradient descent, as discussed in detail in Appendix H.6

5.2.3 Simulation and Implementation Description

Using the parameter estimates from the NielsenIQ data for the ground coffee category, we now conduct simulations comparing our Thompson sampling approach to the baseline algorithms discussed above. We consider a setting where, at each time, a customer arrives with the intent to buy 32 ounces (2lb bags) of coffee from the brands we consider.[11] At this point, we decide on the price for each brand and then observe the customer's choice (which is one of the brands or the outside option). The parameters and optimal prices and promotions for this setting are discussed in Appendix H.3.

There are several implementation details and parameters whose discussion we have deferred to Web Appendix H.5. These include choices of the parameter ranges u, l, the initial exploration phase $\tau_{explore}$ in which we play totally random prices to initialize all algorithms, and the number of Monte-Carlo repeats (set to 40). In our simulations, for the display variable, we restrict to $\mathcal{X} = \{\mathbf{e}_i : 1 \leq i \leq K\} \cup \{\mathbf{0}\}$. Again, this corresponds to a setting where we can choose to promote one item or promote none at all.

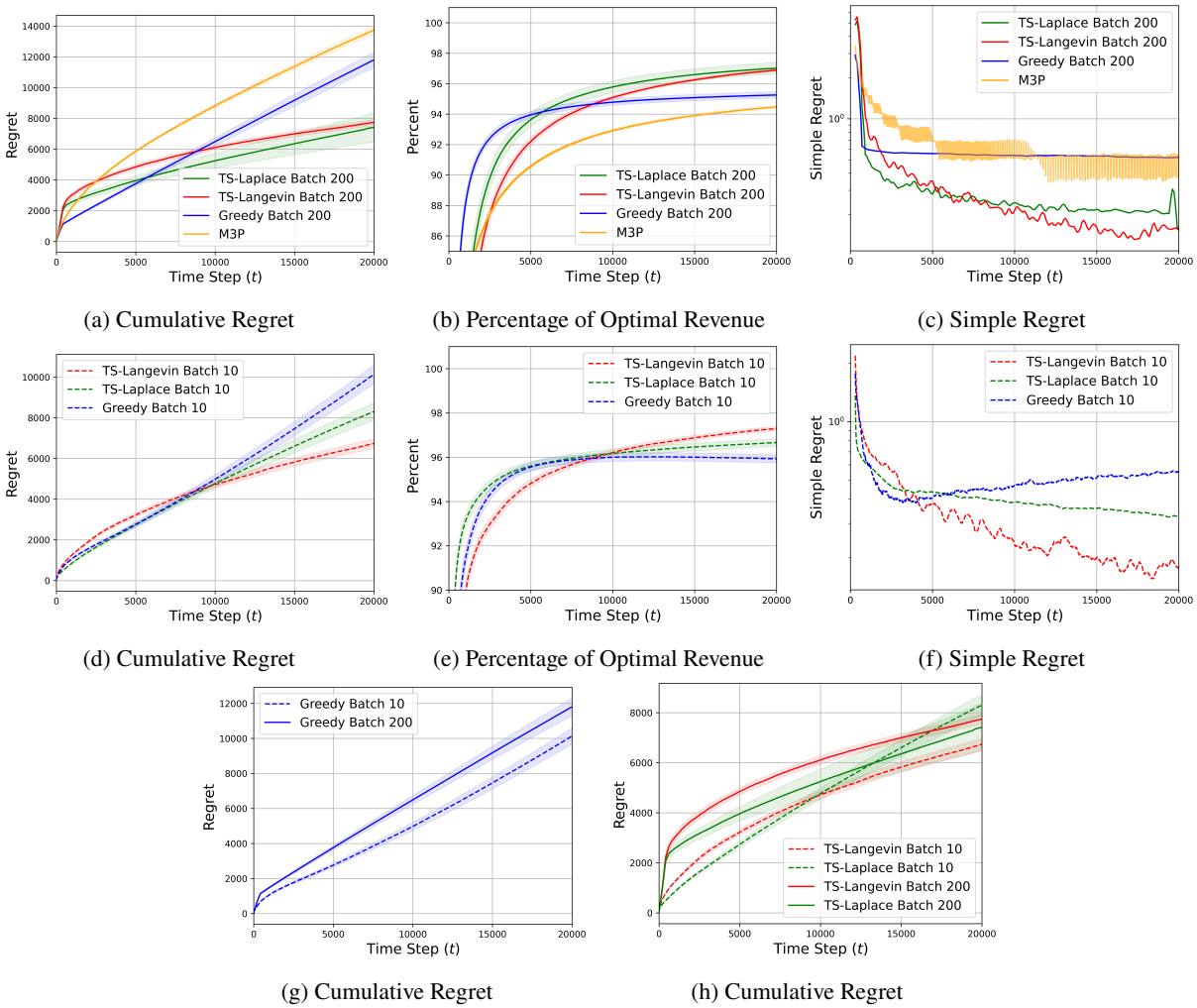
Secondly, we also have to choose the *batch-size*. In practice, it is extremely difficult for firms to do real-time updates of adaptive experimentation algorithms due to computational and engineering costs. Instead, it's common to do updates in fixed time-intervals or in fixed batches. For example, if the batch size is 200, we assume that there are 200 purchase decisions made per period, and all the parameter estimates and the $\{\mathbf{p}_t, \mathbf{x}_t\}$ are updated at the start of each batch and then fixed for the rest of the batch. Depending on the retailer's volume, each batch could translate to a few hours, a day, or a week. [12] We consider two settings, a batch size of 10 (frequent updating) and a batch size of 200.

5.2.4 Simulation Results

Figure 4 shows the summary of our results. In the first row, we consider the setting with a batch size of 200 and plot the regret for our algorithm, using Thompson Sampling implemented with both Laplace approximation (TS-Laplace) and Langevin Dynamics (TS-Langevin), Greedy and M3P. For each algorithm, we show the mean and 95%-confidence bands on the mean cumulative regret for each batch (as defined in Equation (5)) based on 40 runs. Both versions of Thompson Sampling consistently perform better than all the benchmarks. Interestingly, though Greedy shows improved performance initially, it eventually loses out to Thompson Sampling. M3P does particularly poorly. Recall that the extent of exploration here is directly proportional to the dimensionality of the choice set: the bth block of M3P is $b + K$ long, where K is the number of products/brands. Thus, as K increases, the algorithm spends a significant amount of time in exploration, which contributes significantly to the regret (especially since prices in these exploration rounds are chosen randomly). This demonstrates an important practical point, though the regret guarantee of M3P matches ours, i.e. $O(K\sqrt{T})$, the theoretical guarantee is not reflected in its actual performance. We will return to this point when discussing contextual models.

[11] Note that we set all marginal costs to zero since it was not observed in the data. However, they are very easy to include when the retailer has access to them, and doing so does not qualitatively affect our results. As a result, revenue and profits are the same in all of our results and plots.

[12] If we were to make a very conservative assumption that there are only 200 consumers in this category in a given day at the retailer, then the timespan of the data is 20000/200 = 100 days. In most large retailers, the number of daily customers is much larger, and this would represent anyone from just one or two days (e.g., Amazon) to a couple of weeks (for smaller retailers with one or two stores). Note that these numbers reflect the anecdote that Amazon changes prices hourly Mehta et al. (2018).



(a) Cumulative Regret (b) Percentage of Optimal Revenue (c) Simple Regret

(d) Cumulative Regret (e) Percentage of Optimal Revenue (f) Simple Regret

(g) Cumulative Regret (h) Cumulative Regret

Figure 4: We compare the performance of Thompson Sampling (TS-Laplace and TS-Langevin), M3P, and a myopic Greedy baseline for the top ten products with the highest market share on the Nielsen MarketIQ dataset. In the top row we consider a setting with a batch size of 200 and the second row shows a batch size of 10. Finally, in the last row, we compare the Greedy/Thompson Sampling algorithms at the two different batch sizes. All plots show the mean over 40 simulations with 95% confidence intervals on the mean.

To put the gains of Figure 4a in perspective, in Figures 4b we consider the percentage of revenue recovered relative to playing the optimal price/promotion of the true demand model in each customer interaction. As can be seen, by the end of the experiment, Thompson Sampling performs $\approx 2\%$ better than Greedy and about 3% better than M3P. Next, in Figure 4c we plot a smoothed version of the simple regret (Equation (6)), i.e., per period regret for the three algorithms.Again, we find that Thompson Sampling performs well. Finally, we remark that the simple regret of M3P has high variance because it involves periods of completely random exploration even much later in the horizon (these periods are not shown due to our smoothing). From a practical perspective, this makes M3P a poor choice in real empirical applications (quite apart from the high regret for high dimensional settings) – wildly varying prices can be challenging to implement and lead to customer confusion/dissatisfaction. Further, since this leads to largely unpredictable profits/revenues, firms and managers may be wary to adopt it. In contrast, we see that Thompson sampling has low variance in addition to low regret, which makes it appealing from both profitability and practicality standpoints (i.e.,

acceptable in real business settings).

In the second row of Figure 4, we consider the setting of a batch size of 10 and we consider Thompson Sampling with the Laplace Approximation (TS-Laplace), and Thompson Sampling with Langevin Dynamics (TS-Langevin), and Greedy. We did not test against M3P, since it is a method that fundamentally batches on a fixed schedule, and can't be adapted for smaller batch sizes. Most of our observations in the case of a batch size of 200 also hold in this setting where we are updating more frequently. In addition, the proportion of revenue recovered is more than 1% higher for our adaptive methods (Figure 4e). In the last row, we compare Greedy and Thompson Sampling for the two different batch sizes.

Thompson Sampling demands minimal computational resources, and its runtime is comparable to that of baseline methods M3P and Greedy. Further details on the runtime are provided in Web Appendix H.4.

6 Learning Optimal Prices and Promotions in Contextual Settings

In the previous sections, we only considered settings where all consumers and markets were homogeneous or a case where the retailer did not have data on customer-specific features (e.g., segment, demographics, or behavioral variables) or market-level features. However, in most realistic settings, retailers have additional information about the consumer and the market that can be informative of their brand preferences (αs), and price/promotion sensitivities (βs and γs). For instance, retailers typically have zip-code-level, store-level, or individual-level information on the demographics of consumers. More recently, online retailers have access to a large amount of user-level behavioral features, e.g. prior-browsing data, purchase data, consumer location, and device/browser type. Prior research has shown that both demographic and behavioral features can affect users' preference parameters. For example, higher-income customers may have lower price sensitivity compared to lower-income customers (Allenby and Rossi, 1998; Horsky et al., 2006). Similarly, we know that user-level behavioral and contextual features are predictive of user behavior in digital settings (Yoganarasimhan, 2020). In addition, retailers often have data on market conditions and/or seasonality, that can be informative of consumer behavior. Henceforth, we use the term *contextual features* to refer to both consumer-specific and market-specific variables that can be informative of demand.[13]

Ideally, the retailer should use these contextual variables to customize prices and promotions at the individual-, or market-level since doing so can increase profits. However, these features may be high-dimensional, and it is difficult to a priori. specify a model of how these features will affect the user-level demand parameters. Thus, the problem becomes one of learning to serve personalized prices and promotions, when the firm is simultaneously learning the demand model. Note that prior research has considered the problem of personalizing prices after the firm has run a large-scale experiment to learn the demand parameters (Kallus and Zhou, 2021; Dubé and Misra, 2023). In contrast, we consider a setting where the firm is simultaneously learning the demand parameters and optimizing pricing and promotions. Further, in Section 7, we expand our analysis to cases where the demand parameters are unknown and potentially non-linear functions of the context.

[13]Prior research in marketing often refers to contextual features as non-user-level variables that are informative of user-behavior, e.g., time of the day, seasonality, location, etc. (Rafieian and Yoganarasimhan, 2021). However, for simplicity, here we use the phrase *contextual features* to denote both user/market-level variables that are informative of user preferences.

6.1 Demand Model and Algorithm in Contextual Settings

To incorporate user and market features, we consider the following model referred to as the contextual setting. At each time t, we receive a context vector $\mathbf{c}_t \in \mathbb{R}^d$ capturing information about the customer who arrives at time t. The context vector can capture user demographics, past purchase and behavioral history, and even the time of day, the day of the week, or the macroeconomic conditions at which the customer arrives. As discussed earlier, incorporating these features into the demand model can improve the firm's ability to customize prices and promotions per user.

Therefore, we now assume that the parameters $\alpha : \mathbb{R}^d \to \mathbb{R}^K$, $\beta : \mathbb{R}^d \to \mathbb{R}^K$, and $\gamma : \mathbb{R}^d \to \mathbb{R}^K$ are functions of the user-level context vector ($\mathbf{c}_t$). Then, the probability that a customer chooses item i is:

$$\mathbb{P}_{\alpha,\beta,\gamma}(I_t = i|\mathbf{p}_t, \mathbf{x}_t, \mathbf{c}_t) = \frac{e^{\alpha_i(\mathbf{c}_t)-\beta_i(\mathbf{c}_t)p_{ti}+\gamma_i(\mathbf{c}_t)x_{ti}}}{1+\sum_{j=1}^{K} e^{\alpha_j(\mathbf{c}_t)-\beta_j(\mathbf{c}_t)p_{tj}+\gamma_j(\mathbf{c}_t)x_{tj}}}, \tag{11}$$

where for context $\mathbf{c}$, $\alpha_i(\mathbf{c})$, $\beta_i(\mathbf{c})$, and $\gamma_i(\mathbf{c})$ are the i^{th} coordinates of $\alpha(\mathbf{c})$, $\beta(\mathbf{c})$, and $\gamma(\mathbf{c})$. This formulation has previously been considered in Goli et al. (2021) to model the heterogeneous effect of ad-load sensitivity as a function of user attributes, and in Dubé and Misra (2023) to model demand in A/B testing. Notably, neither of these papers considers active data collection; rather they first run an A/B test and then use this formulation for off-policy design.

The case where α, β, γ are constant functions of the context $\mathbf{c}$ recovers the demand model considered in Section 5. As discussed in Section 2, prior work in the dynamic pricing case (i.e. without promotion) has considered the setting where α, β are linear functions of the context and that the contexts come from a fixed known distribution.[14] As far as we know, we are the first to consider setting optimal prices and promotions, when α, β, γ are unknown and potentially non-linear functions of the context. Furthermore, we make no distributional assumptions on the context or its stationarity.

The main advantage of our formulation is that we allow α, β, γ to be arbitrary and as a result, we can handle more expressive models and are more robust to model misspecification. However, allowing for more general classes of functions comes with its own challenge. Firstly, we need a way to estimate the underlying model. To do so, we express α, β, γ as arbitrary neural networks that are functions of the underlying customer features $\mathbf{c}_t$. To learn these neural networks, we can use standard packages such as PyTorch. The second, and more interesting question is how to conduct exploration in the price and promotion space. Given the theoretical and experimental success of our Thompson sampling approach in non-contextual settings, ideally, we would be able to extend it to the contextual setting. However, Laplace approximation cannot be easily implemented for arbitrary neural networks. Indeed, computing and sampling from the Hessian of the MLE (i.e. Fisher information) can be computationally difficult if the number of parameters of the neural network is high. Fortunately, our Langevin dynamics approach (see Section 4.2) is still computationally feasible

[14] Ban and Keskin (2021) considers the single product setting with linear demand and Javanmard and Nazerzadeh (2016) considers the multiproduct multinomial setting. We remark that Javanmard and Nazerzadeh (2016) considers a setting where the firm has additional demand information at each round for each product (which is interpreted as product or market features) which enters the linear utility function – however, they do not see customer features. Though these problems may seem different at first glance, mathematically our framework (assuming no promotions) can capture their model and we have modified M3P for our setting (see Section F for details). As we will demonstrate, we perform empirically better than M3P consistently.

Algorithm 3 Langevin-based Contextual Thompson Sampling with Multinomial Demand Model

Input: $\mathcal{P} \times \mathcal{X}$, step sizes $\{\eta_t\}_{t\geq 1}$, inverse temperature parameters $\{\psi_t\}_{t\geq 1}$, Langevin steps N, regularization factor λ

$\theta_{1,0} \leftarrow [0, \ldots, 0] \in \mathbb{R}^K$

for $t = 1, 2, \cdots,$ **do**

$\theta_{t,0} = \theta_{t-1,N}$

Observe context $\mathbf{c}_t$.

for $n = 1, \ldots, N_t$ **do**

Sample $\epsilon_{t,n} \sim \mathcal{N}(0, \mathbf{I})$

Define $\mathcal{L}_t(\{(\mathbf{p}_s, \mathbf{x}_s, \mathbf{c}_s, I_s)\}_{s=1}^t, \theta) = \sum_{s=1}^t \log(\mathbb{P}_\theta(I_s|\mathbf{p}_s, \mathbf{x}_s, \mathbf{c}_s)) + \frac{\lambda}{2}\|\theta\|_2^2$

$\theta_{t,n} \leftarrow \theta_{t,n-1} - \eta_t \nabla \mathcal{L}_t(\{(\mathbf{p}_s, \mathbf{x}_s, \mathbf{x}_s I_s)\}_{s=1}^t, \theta_{t,n-1}) + \sqrt{2\eta_t \psi_t^{-1}}\epsilon_{t,n}$

end for

Set $\mathbf{p}_t, \mathbf{x}_t = \arg\max_{\mathbf{p}\in\mathcal{P}, \mathbf{x}\in\mathcal{X}} R_{\theta_{t,N_t}}(\mathbf{p}, \mathbf{x}, \mathbf{c}_t)$

Observe I_t and $r_t := p_{I_t} - m_{I_t}$

end for

and does not require difficult computation beyond the ability to conduct a gradient step, easily implemented with automatic differentiation. Such an approach has recently been used in the contextual bandit literature, and extends the Thompson sampling approach to a non-parametric setting. Though there is still a lack of theoretical results for this approach, it has shown empirical promise (Xu et al., 2022; Mazumdar et al., 2020).

Our precise formulation of Langevin dynamics is similar to that of Algorithm 2 with two changes. Firstly, we observe the context $\mathbf{c}_t$ in each round, and secondly, θ is a set of parameters representing our parametrization of α, β, γ. Define the regularized log-likelihood function,

$$\mathcal{L}_t(\{(\mathbf{p}_s, \mathbf{x}_s, \mathbf{c}_s, I_s)\}_{s=1}^t, \theta) = \sum_{s=1}^t \log(\mathbb{P}_\theta(I_s|\mathbf{p}_s, \mathbf{x}_s, \mathbf{c}_s)) + \frac{\lambda}{2}\|\theta\|_2^2.$$

Then our extension of Langevin dynamics is given in Algorithm 3.

6.2 Theoretical Guarantees for Linear Contextual Setting

In this section, we focus on the case where $\alpha_i(\mathbf{c}), \beta_i(\mathbf{c}), \gamma_i(\mathbf{c})$ are linear functions of $\mathbf{c}$. That is, there exist vectors $\alpha_i, \beta_i, \gamma_i \in \mathbb{R}^d, 1 \leq i \leq K$ (abusing notation slightly) such that

$$\alpha_i(\mathbf{c}) = \langle \alpha_i, \mathbf{c}\rangle, \beta_i(\mathbf{c}) = \langle \beta_i, \mathbf{c}\rangle, \gamma_i(\mathbf{c}) = \langle \gamma_i, \mathbf{c}\rangle, \tag{12}$$

and so α, β, γ have $K \times d$ matrix representations as linear functions of the context $\mathbf{c}$,

$$\alpha(\mathbf{c}) = [\alpha_1, \alpha_2, \cdots, \alpha_K]^\top \mathbf{c}, \beta(\mathbf{c}) = [\beta_1, \beta_2, \cdots, \beta_K]^\top \mathbf{c}, \gamma(\mathbf{c}) = [\gamma_1, \gamma_2, \cdots, \gamma_K]^\top \mathbf{c}$$

Define $\theta = (\alpha_1, \beta_1, \gamma_1, \cdots, \alpha_K, \beta_K, \gamma_K) \in \mathbb{R}^{3dK}$.

Theorem 3. *Assume that* $\max_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x}\|_\infty \leq B$*, the largest price is bounded by* u*,* $\|\theta\| \leq S$ *with probability 1. With probability at least* $1 - \delta$*, the regret of our Thompson Sampling algorithm in the contextual setting is bounded by* $\widetilde{O}\left(SudK\sqrt{\kappa T}\right)$ *where* $\widetilde{O}(\cdot)$ *hides constants and doubly logarithmic factors.*

Proof: See Appendix D.

We make a few points about this theorem. Firstly, in the case where $d = 1$ and $\mathbf{c}_t = 1$ for all $t \geq 1$, this result generalizes Theorem 2. In general, the regret scales linearly in d and K. This matches existing optimal minimax regret rates in the contextual linear bandit setting Lattimore and Szepesvári (2020). Secondly, we do not make the assumption that the contexts are drawn i.i.d. from a fixed distribution at each time. Or in other words, we are not assuming that the context distribution is stochastic. Indeed, the context distribution could be changing at each time, or even potentially adapt to past actions of the firm. In the bandits literature, a setting with arbitrary contexts is referred to as an adversarial setting (Lattimore and Szepesvári, 2020). Such *adversarial* settings are common in marketing. For example, the company could use a promotion strategy that causes the underlying population of customers to change over time. Alternatively, depending on the season we could have different types of customers with varying preferences. We believe that this is the first result in this setting that allows for arbitrary, and even potentially adversarial, contexts. Past works in the contextual setting such as Javanmard et al. (2020); Ban and Keskin (2021) relied on stochastic contexts. As a result, our theory is strictly more general than past guarantees.

7 Numerical Experiments in Contextual Settings

We now present three sets of numerical simulations for the contextual setting. The goal of these simulations is to demonstrate the flexibility of our methodology in different but natural marketing settings. Throughout the following, we use Langevin dynamics to approximate posterior sampling for our Thompson Sampling implementation. First, in Section 7.1, we construct a series of synthetic experiments under the linear contextual setting where we see the impact of different context distributions. In Section 7.2, we apply our algorithm to the NielsenIQ dataset in a setting where the context distribution shifts over time. Finally, in Section 7.3, we extend our model to choice-based demand models with non-linear utility functions.

7.1 Synthetic Experiments for the Linear Contextual Setting

In this section, we consider a set of synthetic experiments that demonstrate that our algorithm leads to a lower regret compared to non-contextual baselines, and other natural baselines in the contextual setting. We consider a setting with 9 products and four-dimensional context vectors, i.e. each $\mathbf{c}_t \in \mathbb{R}^4$ for all $t \geq 1$. As a result, the parameters are four-dimensional as well, i.e. $\alpha_i, \beta_i, \gamma_i \in \mathbb{R}^4$.

7.1.1 Context Vectors

We consider two different context distributions that capture two different types of customer populations.

- OrthogonalGroups: At each time the context $\mathbf{c}_t \in \mathbb{R}^4$ is uniformly drawn from the set of standard basis vectors in $\mathbb{R}^4$, that is $\mathbf{c}_t \in \{\mathbf{e}_1 = (1,0,0,0), \mathbf{e}_2 = (0,1,0,0), \mathbf{e}_3 = (0,0,1,0), \mathbf{e}_4 = (0,0,0,1)\}$. Intuitively, this corresponds to a setting where we have four segments of consumers, where the population within each segment is homogeneous in its preferences. For each segment $1 \leq j \leq 4$, and product $1 \leq i \leq 9$ there is an unknown set of parameters $\alpha_{ij} := \langle \alpha_i, \mathbf{e}_j \rangle$, $\beta_{ij} := \langle \beta_i, \mathbf{e}_j \rangle$ and $\gamma_{ij} := \langle \gamma_i, \mathbf{e}_j \rangle$. The underlying demand model for segment j is given by Equation (2) in Section 3.1. Notice that since the contexts are orthogonal, there is no information sharing between the segments - namely, in each context, our goal is to find the set of prices and promotions that optimize the revenue for the corresponding segment.

If the firm, a priori, knew the context distribution was orthogonal, it could run a different instance of our Thompson Sampling algorithm designed for the non-contextual case (Algorithm 1) for each group. However, since the firm may not have this information it needs an algorithm that is general-purpose and can naturally adapt to this specific setting. Furthermore, as we will see if the firm incurs significantly more regret if it ignores contexts altogether.

- WeightedAverages: This setting generalizes the one above. In this setting we select context vectors uniformly from the unit simplex of $\mathbb{R}^4$ space, that is $\mathbf{c}_t \sim \text{Unif}(\Delta_4)$, with $\Delta_4 = \{x \in \mathbb{R}^4_{\geq 0} : \sum_{i=1}^4 x_i = 1\}$. Intuitively, this extends the above model where we now allow for users to be weighted combinations of the four contexts above. In particular, unlike the previous example, this allows for a population with very heterogeneous preferences. Since the number of possible contexts is infinite, it's not possible to run a separate algorithm for each context.

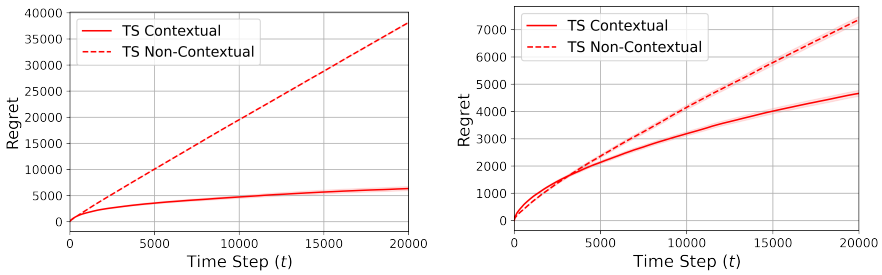
7.1.2 Model

The values of the parameters $\alpha_i, \beta_i, \gamma_i \in \mathbb{R}^4, 1 \leq i \leq 9$ were chosen to extend the example in Section 5.1 and are given in Table A6 in Web Appendix I.1. We omit them here for brevity. In addition to these parameters, the implementation details (such as the batch size or range of prices considered) are detailed in Web Appendix I.2. Finally, for our baseline algorithms, we considered Greedy and M3P, which are similar to those described in Section 5.2.2. The Greedy method plays the optimal prices and promotions for each product based on the current estimated model. For M3P, we follow the algorithm of Javanmard et al. (2020).

7.1.3 Results

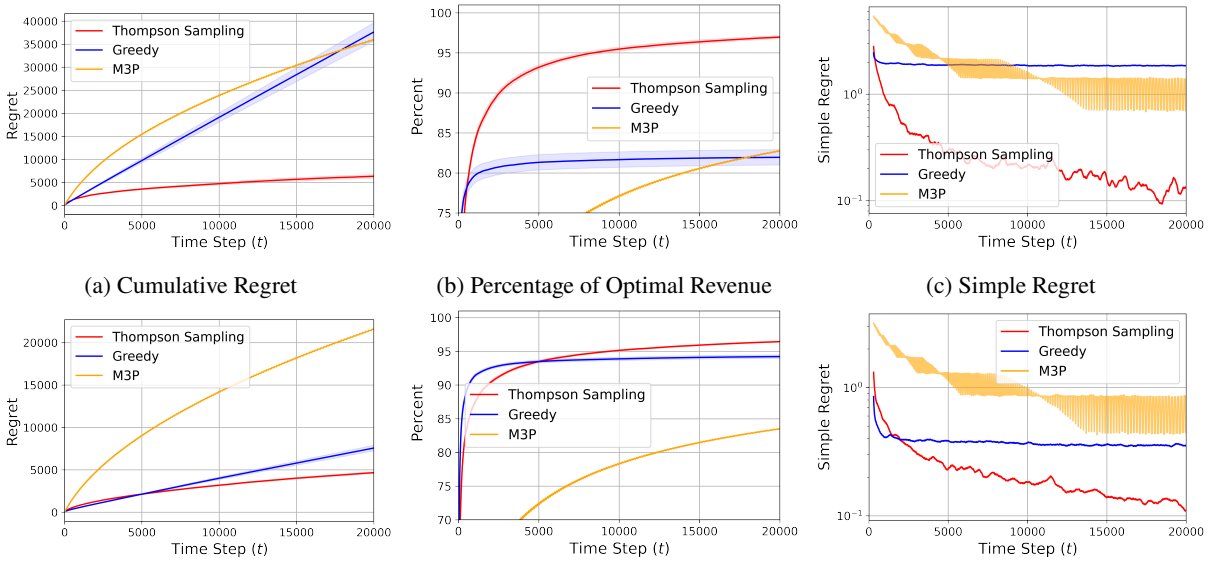
We now discuss the experiment results for the settings described above. Our first set of plots provides insight into the impact of including contextual information. Figure 5 compares the performance of Algorithm 3, Thompson Sampling in the linear contextual setting, (i.e. using the demand model from Section 6.2) to Algorithm 2, Thompson Sampling in the non-contextual setting. Note that the non-contextual variants learn models that are misspecified. Namely, by not considering the context distribution, the non-contextual model assumes that consumers are homogeneous, and as a result, can only optimize at a population level. In contrast, contextual-TS is able to learn how demand changes with contexts and customize prices and promotions accordingly. On the left (Figure 5a), we show the regret plot for the OrthogonalGroups context distribution, and on the right we have the regret plot for the WeightedAverages contexts (Figure 5b). We can see that in both settings, ignoring the contexts results in significantly worse regret.

Now we compare Algorithm 3, the contextual version of Thompson Sampling, to other contextual baselines, Greedy and M3P (Figure 6); see Web Appendix F for details. The first row shows the results for OrthogonalGroups, and the second row shows the results for WeightedAverages. Both Greedy and Thompson Sampling perform better than M3P in this time horizon. After 20,000 steps, when the context distribution is OrthogonalGroups, Thompson Sampling recovers 15% more of the optimal revenue compared to Greedy, whereas in the WeightedAverages case, it recovers 2.23% more than Greedy. This behavior is also captured in the simple regret plot where we can see that the simple regret for Greedy does not improve, and by time 20,000, the simple regret of Greedy is more than three times larger than the TS simple regret. We remark that Greedy is more competitive in early rounds in the WeightedAverages setting than in



(a) OrthogonalGroups (b) WeightedAverages

Figure 5: We compare the cumulative regret of the contextual versions of Thompson Sampling (TS Contextual) with its non-contextual counterparts (TS Non-Contextual) for two context distributions –(1) OrthogonalGroups and (2) WeightedAverages. In this example, the 95% confidence intervals are shown, but they are very small and so can't be seen easily.



(a) Cumulative Regret (b) Percentage of Optimal Revenue (c) Simple Regret

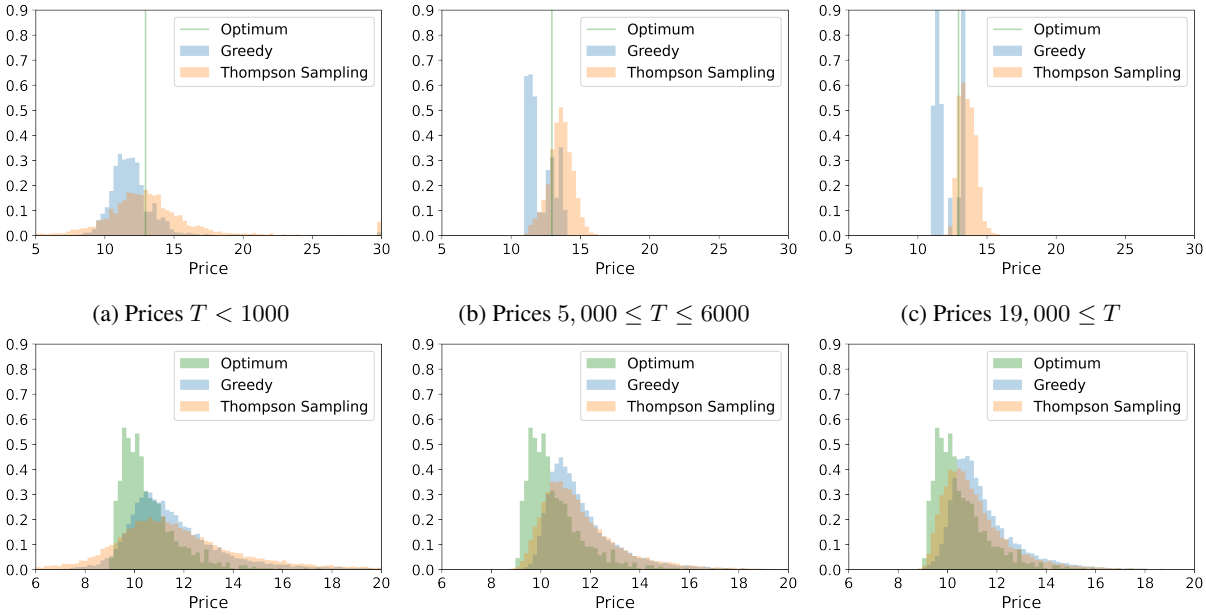
(d) Cumulative Regret (e) Percentage of Optimal Revenue (f) Simple Regret

Figure 6: Comparison of the performances of three contextual methods – Thompson Sampling, M3P, and Greedy. The first row shows the results for OrthogonalGroups, and the second row shows the results for WeightedAverages.

OrthogonalGroups. As described above, the latter setting is basically a different bandit for each of the four groups with no information sharing. In the former setting, the information is shared since each customer is a mixture of the four groups. As a result, we believe that this allows Greedy to gain more information about the parameters in the early time steps. However, eventually, we see that Greedy has worse regret due to its inability to effectively explore. For the interested reader, we introduce another context distribution, namely Box context distribution, in Web Appendix I.3 and compare the methods' performance in this setting.

In Figure 7, we investigate the methods further by comparing the distribution of prices played by Thompson Sampling and Greedy.[15] The first row shows the histogram of the prices played for a randomly chosen product from the set of products (product seven) for the OrthogonalGroups setting in different time frames for time periods where the context vector $\mathbf{c}_t$ is $\mathbf{e}_1$. The green "Optimum" vertical line shows the optimum price

[15]We do not include M3P in this comparison, since its performance is weaker than Greedy's. We show the distribution of the prices played by the M3P method and compare it to the distributions of Thompson Sampling and Greedy methods in Web Appendix I.4.



(a) Prices $T < 1000$ (b) Prices $5,000 \leq T \leq 6000$ (c) Prices $19,000 \leq T$

(d) Prices $T < 1000$ (e) Prices $5,000 \leq T \leq 6000$ (f) Prices $19,000 \leq T$

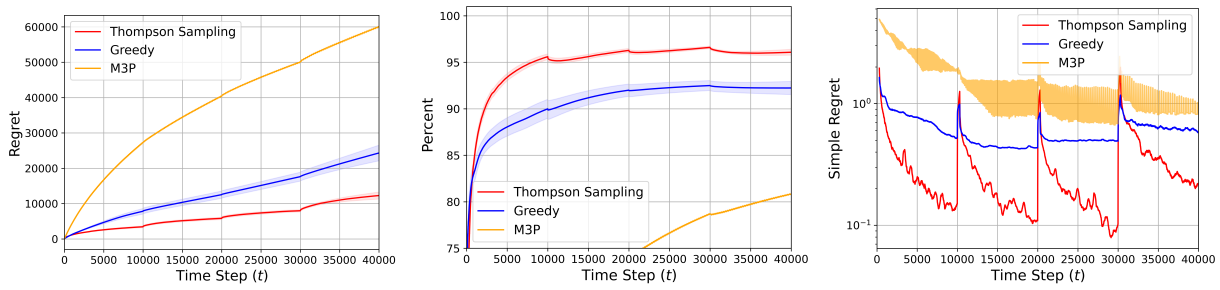
Figure 7: We compare the distribution of prices played for a randomly chosen product (product seven) by Thompson Sampling and Greedy. In the first row, we show the distribution of the prices for the OrthogonalGroups when the context vector is the first basis vector ($\mathbf{c} = e_1$). The green line shows the optimum price point for this context. The second row shows the marginal distribution over different contexts for the WeightedAverages case, and the green histogram shows the distribution of optimum prices. Each of the three columns captures different timeframes (early stage, middle stage, final stage).

for the product in this context, and the histograms show the distribution of prices played by Greedy and TS s. While initially ($T < 1000$), TS is exploring the feasible price range more, by the end ($19,000 \leq T$), the distribution of prices played is centered at the optimum price of \$12.94. In comparison, the distribution of prices played by Greedy at the final stage has a peak near \$11, explaining the high regret incurred.

The second row shows the distribution of prices played for the WeightedAverage contexts. In this case, "Optimum" denotes the histogram of optimal prices for the sequence of contexts observed in all 20 replicates of the algorithm. We have also plotted histograms of the prices played by TS and Greedy. Again, note TS has more exploration at the beginning stages and in later time steps, the peak of the distribution of prices played by TS is farther to the left, approaching that of the optimum distribution. To conclude, Thompson Sampling with Langevin Dynamics is a competitive method in the linear contextual setting, regardless of any knowledge of the context distribution or the specific distribution itself.

7.2 Experiments based on NielsenIQ Data in a Non-Stationary Linear Contextual Setting

In this section, we present a set of linear contextual experiments based on parameters estimated from the NielsenIQ dataset. Though there are no specific user-level contextual variables in the dataset, we construct a contextual setting that assumes demand parameters depend on which quarter of the year and the store where the purchase was made in. In particular, due to the quarter variable, unlike the settings in the synthetic experiments above, our customer context distribution will be allowed to shift over time. As we will see, our algorithm effectively adapts to the shifts without prior knowledge of the distribution and achieves a



(a) Cumulative Regret (b) Percentage of Optimal Revenue (c) Simple Regret

Figure 8: We compare the performance of Thompson Sampling, M3P, and Greedy baseline in a setting with shifting context distribution. The model and contexts are based on the NielsenIQ data in the ground coffee category. The context vectors represent the quarter of the year and store. The jumps in simple regret are due to the model learning the demand parameters in each quarter.

significantly lower regret than other benchmarks.

We use the NielsenIQ weekly retail scanner dataset from 2019 as described in Section 5.2. While Section 5.2.1 used data from only one store, in this section, we select the top two stores (by ground coffee revenue) in King County, WA that have data on weekly prices, features, and displays. This allows us to capture the variation in demand across the two stores and use the store identity as a contextual feature. Similarly to Section 5.2, we considered the 9 ground coffee brands with the highest revenue (after combining data from both stores) and grouped all the remaining brands under an *Other* category.

We consider a demand model where the utility parameters are assumed to depend on the store where the purchase happens and the quarter of the year. Notice that we are implicitly assuming consumers are not choosing the store where they make the purchase or the quarter when they make a purchase. Instead, the store and quarter variables are treated as contextual variables. To that end, we specify a 6-dimensional context vector, with four dimensions representing quarter dummies ($quarter \in \{Q_1, Q_2, Q_3, Q_4\}$) and two dimensions representing store dummies ($store \in \{1, 2\}$). Thus, the demand parameters for brand $i \in \{1, \cdots, 10\}$ are given by:

$$\begin{aligned}
\alpha_i(\mathbf{c}) =& \alpha_{iQ_1}\mathbb{I}(Q_1) + \alpha_{iQ_2}\mathbb{I}(Q_2) + \alpha_{iQ_3}\mathbb{I}(Q_3) + \alpha_{iQ_4}\mathbb{I}(Q_4) + \alpha_{iS_1}\mathbb{I}(store = 1) + \alpha_{iS_2}\mathbb{I}(store = 2) \\
\beta_i(\mathbf{c}) =& \beta_{iQ_1}\mathbb{I}(Q_1) + \beta_{iQ_2}\mathbb{I}(Q_2) + \beta_{iQ_3}\mathbb{I}(Q_3) + \beta_{iQ_4}\mathbb{I}(Q_4) + \beta_{iS_1}\mathbb{I}(store = 1) + \beta_{iS_2}\mathbb{I}(store = 2) \\
\gamma_i(\mathbf{c}) =& \gamma_{iQ_1}\mathbb{I}(Q_1) + \gamma_{iQ_2}\mathbb{I}(Q_2) + \gamma_{iQ_3}\mathbb{I}(Q_3) + \gamma_{iQ_4}\mathbb{I}(Q_4) + \gamma_{iS_1}\mathbb{I}(store = 1) + \gamma_{iS_2}\mathbb{I}(store = 2).
\end{aligned}$$

The utility for a customer at time t with context vector $\mathbf{c}_t$ with a price vector $\mathbf{p}_t$ and promotion $\mathbf{x}_t$ for product i is given by

$$U_i(\mathbf{c}_t, \mathbf{p}_t, \mathbf{x}_t) = \alpha(\mathbf{c}_t) - \beta(\mathbf{c}_t)p_{it} + \gamma_i(\mathbf{c}_t)x_{it} + \epsilon_{it}. \tag{13}$$

We estimate a demand model based on the utility specification above on weekly purchase data using a procedure similar to Section 5.2.1. We then lightly modify the estimated parameters to handle some issues arising from missing/insufficient data for certain store-brand combinations. The details of the estimation, the modifications, and the parameters used in the experiments are described in detail in Web Appendix J.

Next, we describe how the context variable is generated at each time in the simulation exercise. For

quarter dummies, we divide the total time horizon $T = 40000$ into four equal time intervals with each interval corresponding to one of the quarters. In each interval, the associated quarter dummy variable is 1, and all other quarter dummies are zero. For the store dummies, we make a random draw at each time step, with probabilities that mimic the ratio of store market shares in the data. In the estimation data, the total market sizes are 669,193, and 355,307 ounces respectively for each of the stores. Based on these, we assign a probability of 0.65 for the first store and 0.35 for the second store.

Figure 8 shows the simulation results. In the left panel, we see that our TS approach has the lowest regret, followed by Greedy, while M3P performs significantly worst. The regret plot demonstrates that TS has 49.49% less regret at the end of the time horizon compared to Greedy. In terms of the percentage of optimal revenue recovered that is shown in the middle panel, Thompson Sampling can recover 4.20% more of the optimal revenue than Greedy. Finally, we see that whenever we switch quarters (namely at times $t \in \{10000, 20000, 30000\}$), there is a jump in the regret. This is especially prominent in the last panel, where we show simple regret, i.e., per-period regret. These jumps happen because the model needs some time at the start of a quarter to learn the parameters corresponding to that quarter. Overall, even when contexts are changing, our contextual-TS algorithm is able to learn the demand model in order to personalize prices and promotions for each context.

7.3 Experiments in Non-Linear Contextual Settings: Clustered Customer Groups

We now extend our approach to settings where the demand parameters are non-linear functions of the context. More precisely, in the notation of Section 6.1, we assume that α, β, γ are unknown and non-linear functions of the context vector $\mathbf{c}_t$ and model them using a neural network.

We consider a setting where our customer population can be clustered into a set of groups with similar characteristics. However, the firm does not know that the context vectors have this underlying group structure and instead only observes a four-dimensional context vector representing each customer. We furthermore assume that the underlying demand model for a set of products in each group is constant for all customers in that group, and is given by demand formulation of Equation (2) for group-specific demand parameters. We now describe the context distribution arising from these groups and the parameters for each group.

To define our context distribution, we consider a clustered setting where each user is represented by a four-dimensional context vector $\mathbf{c} \in \mathbb{R}^4$. The underlying distribution is a Gaussian Mixture Model (GMM) over eight centers in four dimensions. More precisely, $\mathbf{c}_t \sim \frac{1}{8}\sum_{g=1}^{8} \mathcal{N}(C_g, 0.12\mathbf{I}_4)$, for all $t \geq 1$ where the cluster centers $\{C_g\}_{g=1}^{8} \subset \mathbb{R}^4$ are given in the Web Appendix in Table A10. We chose the covariance matrix to ensure that with high probability the clusters are non-overlapping. A sample of 1,000 contexts drawn from this distribution is shown in Figure 9. As we can see, this setting can be interpreted as one where there are eight segments of well-separated consumers.

We consider a setting with nine products. We use the notation from Section 6.1 and define the functions $\alpha : \mathbb{R}^4 \to \mathbb{R}^9, \beta : \mathbb{R}^4 \to \mathbb{R}^9, \gamma : \mathbb{R}^4 \to \mathbb{R}^9$ as non-linear functions that map the contexts to demand parameters as defined in Equation (11). Our specific choice of the true α, β, γ exploits the cluster structure discussed above. To achieve this, we define eight regions and utilize a piecewise constant function that outputs a set of parameters for contexts coming from each region. Specifically, we divide $\mathbb{R}^4$ into eight regions $\mathcal{S}_1, \cdots, \mathcal{S}_8$, where $\mathcal{S}_g = \{\mathbf{c} \in \mathbb{R}^4 : \|\mathbf{c} - C_g\| \leq \min_{\mathbf{c}' \in \{1,\cdots,8\}} \|\mathbf{c} - C_{g'}\|\}$, i.e. $\mathcal{S}_g$ is the set of

points whose closest cluster center is C_g. For each region $1 \leq g \leq 8$ we define a set of parameters, namely $\bar{\alpha}_g, \bar{\beta}_g, \bar{\gamma}_g \in \mathbb{R}^9$ (the precise values of these are given in Table A11). Using these regions $\mathcal{S}_1, \cdots, \mathcal{S}_8$ and parameters $\bar{\alpha}_g, \bar{\beta}_g, \bar{\gamma}_g \in \mathbb{R}^9$, we define our demand parameter functions α, β, γ as a piecewise constant functions in each region:

$$\alpha(\mathbf{c}), \beta(\mathbf{c}), \gamma(\mathbf{c}) = \begin{cases} \bar{\alpha}_1, \bar{\beta}_1, \bar{\gamma}_1 & \mathbf{c} \in S_1 \\ \vdots & \\ \bar{\alpha}_8, \bar{\beta}_8, \bar{\gamma}_8 & \mathbf{c} \in S_8 \end{cases} \tag{14}$$

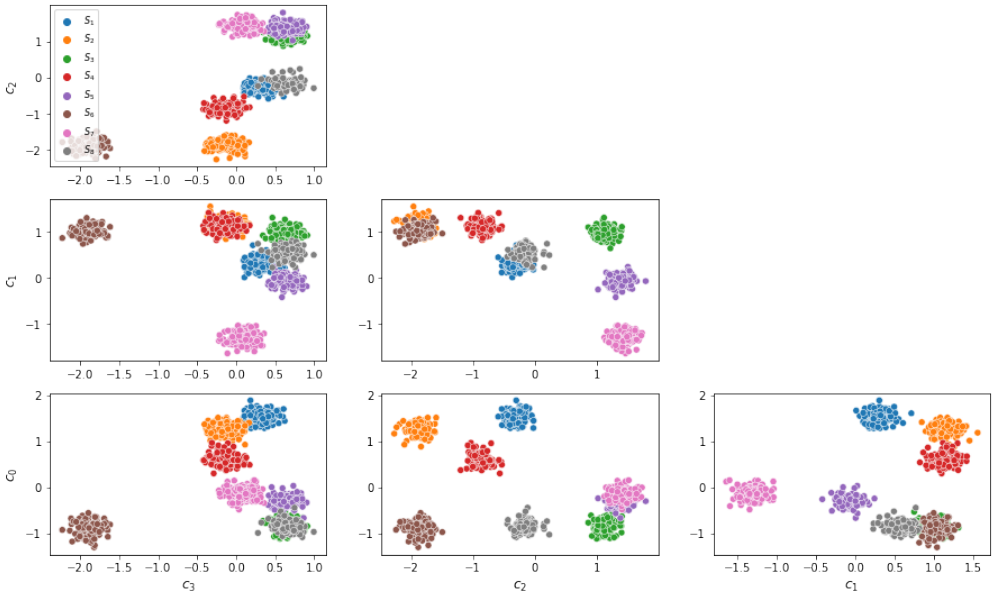
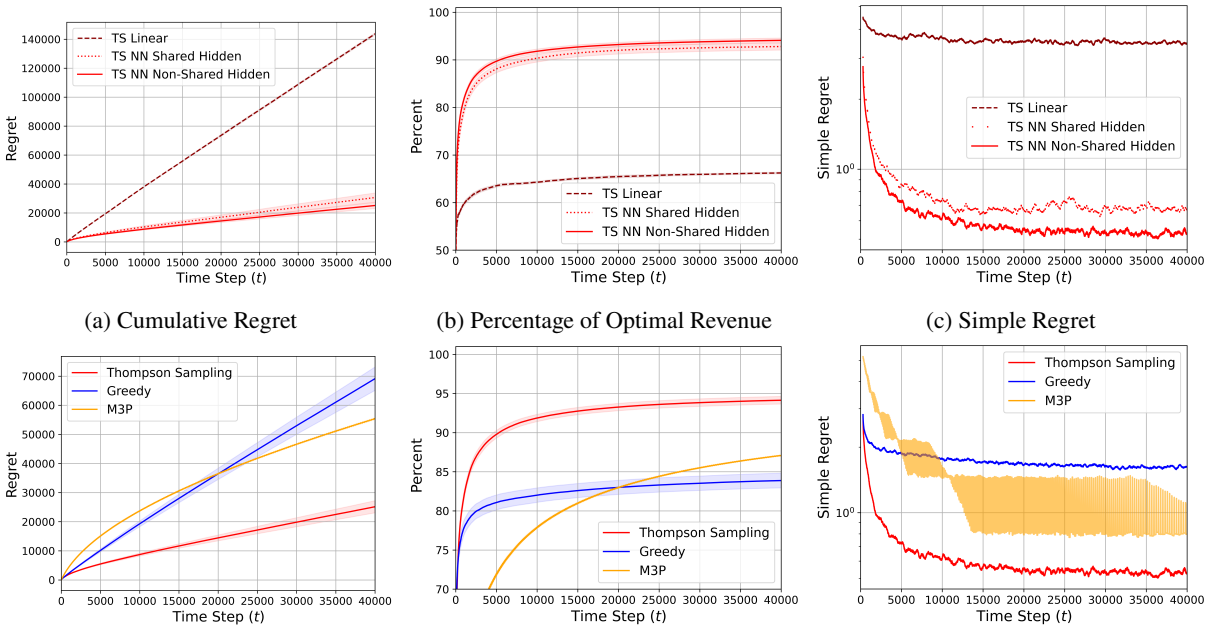


Figure 9: Illustration of the context distribution in the GMM contextual setup. We generated 1000 points from this distribution and plotted them above. Each plot shows a pair of selected dimensions from the 4-dimensional points. The color of the points indicates the region to which each point belongs. Note that in $\mathbb{R}^4$ the regions are non-overlapping.

Of course, in practice, the firm certainly does not know the parameters $\bar{\alpha}_i, \bar{\beta}_i, \bar{\gamma}_i, 1 \leq i \leq 8$ and they may not even know that the underlying demand is driven by the clusters. Furthermore, if the firm has just entered a market or has not collected extensive customer data, it may not even be aware of the underlying clustering of customer contexts. Thus, it needs to model α, β, γ in a sufficiently general way to capture arbitrary functional forms. In our simulation, we compare the regret incurred if the firm were to try two different neural network models against a linear baseline. For our simulation, we estimate α, β, γ using three different models:

- *Linear model:* This model follows the formulation of Equation (12) using a linear mapping from context vector $\mathbf{c} \in \mathbb{R}^4$ to the parameter space α, β, γ.
- *Two-Layer Neural Network with Separate Hidden Layers:* In this model, we have three separate two-layer neural networks for each of the parameters α, β, γ. Each network maps from a four-dimensional input to a nine-dimensional output and has a hidden layer of size four along with sigmoid activation functions.
- *Two-Layer Neural Network with Shared Hidden Layer:* This is similar to the previous model, except that the neural network representations use a shared hidden layer. This network has fewer parameters to learn, and therefore less expressive; however, it may also require less data to train.



(a) Cumulative Regret (b) Percentage of Optimal Revenue (c) Simple Regret

(d) Cumulative Regret (e) Percentage of Optimal Revenue (f) Simple Regret

Figure 10: We compare the performances of different methods in the non-linear contextual setting. The first row compares the linear versions of Thompson Sampling (TS Linear) to variants that use a two-layer neural network; namely, TS NN Shared Hidden (Figure A7a) and TS NN Non-Shared Hidden (Figure A7b). In the second row, we compare Thompson Sampling to Greedy and M3P baselines, where all three methods use the architecture of Figure A7b for estimating demand parameters.

We describe further implementation details of this experiment and the neural network in Web Appendix K. The results are shown in Figure 10. We focus on the first row, which demonstrates that using a neural network model leads to significantly less regret compared to a linear model. This is perhaps not surprising since the true demand parameters α, β, γ are non-linear functions of the context vectors. Thus a linear model is misspecified, and hence our estimates of the demand model will poorly approximate the true demand model. In contrast, both neural network models can better learn the demand function and incur significantly less regret. Furthermore, the model without parameter-sharing performs better and achieves a slightly smaller regret. This is due to the additional level of flexibility of this model over the model with a shared hidden layer. In the second row, we compare the neural network with separate hidden layers version of our method with the neural network versions of the Greedy and M3P baselines with shared hidden layers. Consistent with our previous results, we can see that Greedy has a simple regret which is almost twice the simple regret of TS. By the end of the time horizon, TS has a 7.83% increase in the percentage of optimal revenue recovered compared to Greedy and 4.31% compared to M3P.

Overall, our approach is robust and can easily be applied to a variety of contextual settings that are commonly faced by marketers. Specifically, we show that the approach works irrespective of – (1) whether the underlying demand parameters are linear in contexts or non-linear, (2) whether the context distribution is known ahead of time or not, (3) whether the context distribution remains constant or changes over time.

8 Conclusion

In conclusion, this study provides an effective approach to adaptive pricing and promotions with discrete choice models. Using a Thompson Sampling approach, we develop a regret minimizing, alternatively profit maximizing, algorithm for the retailer. Using simulations based on real-life grocery store data, we show that our method significantly outperforms existing approaches. We also extend our methodology to settings where the demand parameters are functions of customer features and show that our approach can be used to customize prices and promotions in real-time while simultaneously learning the demand model.

Further, our method allows managers to optimize promotions and marketing mix variables, even when price alterations are not feasible. We identify a finite set of optimal promotions that managers can experiment with, utilizing well-established multi-armed bandit techniques such as UCB (Lattimore and Szepesvári, 2020). Future research can build on this foundation by focusing on pure exploration strategies or by incorporating market dynamics into the model, further enhancing the understanding and optimization of adaptive pricing. Extending this work to product categories with a large number of items is also relevant – firms could utilize existing product features to cluster items as part of the underlying parametrization.

From a practical perspective, we give extensive empirical proof that a $O(\sqrt{T})$ regret algorithm (e.g., Javanmard et al. (2020)) does not necessarily translate to empirical success. We illustrate the advantages of using the model-based approach of Thompson sampling in optimizing prices and promotions. By moving away from forced exploration—which proves to be counterproductive in terms of regret minimization and managerial implications, we successfully extend model-based exploration to adaptive pricing and promotion.

Finally, we primarily considered a setting where the context represents customer and market features at any time. However, it is straightforward to also incorporate information about product features at each time, e.g., manufacturers may engage in national advertising or provide seasonal promotions (which would be constant across consumers for a given product in a given time period). An almost identical Thompson algorithm with a similar theoretical guarantee would effectively optimize prices and promotions in such cases.

In general, we expect the model to be useful in cases where the retailer can change prices automatically regularly (e.g., in batches every few hours, daily, or weekly), and for well-defined categories where products are reasonable substitutes for each other (e.g., Consumer Packaged Goods in digital retail settings). Further, it can be particularly useful in settings where there are contextual or time-varying factors that affect demand (e.g., national/brand-level advertising and holiday, seasonality, competitors' pricing, and promotion decisions). It would be harder to use in categories like furniture, high-end fashion, housing, etc., where products are very differentiated and unlikely to be substituted for each other, and each product is unique.

Competing Interests Declaration

Author(s) have no competing interests to declare.

References

M. Abeille and A. Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184. PMLR, 2017.

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

G. M. Allenby and P. E. Rossi. Marketing models of consumer heterogeneity. *Journal of econometrics*, 89 (1-2):57–78, 1998.

N. Aramayo, M. Schiappacasse, and M. Goic. A multi-armed bandit approach for house ads recommendations. *Available at SSRN 4107976*, 2022.

G. Aydin and J. K. Ryan. Product line selection and pricing under the multinomial logit choice model. In *Proceedings of the 2000 MSOM conference*. Citeseer, 2000.

D. Bakry, I. Gentil, M. Ledoux, et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

G.-Y. Ban and N. B. Keskin. Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science*, 2021.

H. Bastani, D. Simchi-Levi, and R. Zhu. Meta dynamic pricing: Transfer learning across experiments. *Management Science*, 2021.

S. Berry, J. Levinsohn, and A. Pakes. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890, 1995.

S. T. Berry. Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, pages 242–262, 1994.

S. T. Berry and P. A. Haile. Foundations of demand estimation. In *Handbook of Industrial Organization*, volume 4, pages 1–62. Elsevier, 2021.

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

O. Besbes and A. Zeevi. On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Science*, 61(4):723–739, 2015.

E. T. Bradlow, M. Gangwar, P. Kopalle, and S. Voleti. The role of big data and predictive analytics in retailing. *Journal of retailing*, 93(1):79–95, 2017.

J. Broder and P. Rusmevichientong. Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980, 2012.

G. Casella and R. L. Berger. *Statistical inference*. Cengage Learning, 2021.

O. Chapelle and L. Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

W. C. Cheung, D. Simchi-Levi, and H. Wang. Dynamic pricing and demand learning with limited price experimentation. *Operations Research*, 65(6):1722–1731, 2017.

S. DellaVigna and M. Gentzkow. Uniform pricing in us retail chains. *The Quarterly Journal of Economics*, 134(4):2011–2084, 2019.

A. V. den Boer and B. Zwart. Simultaneously learning and optimizing using controlled variance pricing. *Management science*, 60(3):770–783, 2014.

L. Dong, P. Kouvelis, and Z. Tian. Dynamic pricing and inventory control of substitute products. *Manufacturing & Service Operations Management*, 11(2):317–339, 2009.

J.-P. Dubé and S. Misra. Personalized pricing and consumer welfare. *Journal of Political Economy*, 131(1): 131–189, 2023.

L. Faury, M. Abeille, C. Calauzènes, and O. Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

J. Feldman, D. J. Zhang, X. Liu, and N. Zhang. Customer choice models vs. machine learning: Finding optimal product displays on alibaba. *Operations Research*, 70(1):309–328, 2022.

K. J. Ferreira, D. Simchi-Levi, and H. Wang. Online network revenue management using thompson sampling. *Operations research*, 66(6):1586–1602, 2018.

S. Gabel and A. Timoshenko. Product choice with large assortments: A scalable deep-learning model. *Management Science*, 68(3):1808–1827, 2022.

R. Ganti, M. Sustik, Q. Tran, and B. Seaman. Thompson sampling for dynamic pricing. *arXiv preprint arXiv:1802.03050*, 2018.

M. S. Goeree. Limited information and advertising in the us personal computer industry. *Econometrica*, 76 (5):1017–1074, 2008.

A. Goli, D. Reiley, and H. Zhang. Personalized versioning: Product strategies constructed from experiments on pandora. *Available at SSRN*, 2021.

P. M. Guadagni and J. D. Little. A logit model of brand choice calibrated on scanner data. *Marketing science*, 2(3):203–238, 1983.

N. Hamidi and M. Bayati. On worst-case regret of linear thompson sampling. *arXiv preprint arXiv:2006.06790*, 2020.

W. Hanson and K. Martin. Optimizing multinomial logit profit functions. *Management Science*, 42(7): 992–1003, 1996.

J. R. Hauser, G. L. Urban, G. Liberali, and M. Braun. Website morphing. *Marketing Science*, 28(2):202–223, 2009.

G. J. Hitsch, A. Hortaçsu, and X. Lin. Prices and promotions in us retail markets. *Quantitative Marketing and Economics*, 19(3):289–368, 2021.

S. J. Hoch, B.-D. Kim, A. L. Montgomery, and P. E. Rossi. Determinants of store-level price elasticity. *Journal of marketing Research*, 32(1):17–29, 1995.

D. Horsky, S. Misra, and P. Nelson. Observed and unobserved preference heterogeneity in brand-choice models. *Marketing Science*, 25(4):322–335, 2006.

J. R. Howell, S. Lee, and G. M. Allenby. Price promotions in choice models. *Marketing Science*, 35(2): 319–334, 2016.

Y. Huang, P. B. Ellickson, and M. J. Lovett. Learning to set prices. *Journal of Marketing Research*, 59(2): 411–434, 2022.

A. Javanmard. Perishability of data: dynamic pricing under varying-coefficient models. *The Journal of Machine Learning Research*, 18(1):1714–1744, 2017.

A. Javanmard and H. Nazerzadeh. Dynamic pricing in high-dimensions. *arXiv preprint arXiv:1609.07574*, 2016.

A. Javanmard, H. Nazerzadeh, and S. Shao. Multi-product dynamic pricing in high-dimensions with heterogeneous price sensitivity. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2652–2657. IEEE, 2020.

J. Johnson, G. J. Tellis, and E. H. Ip. To whom, when, and how much to discount? a constrained optimization of customized temporal discounts. *Journal of Retailing*, 89(4):361–373, 2013.

V. Kadiyali, K. Sudhir, and V. R. Rao. Structural analysis of competitive behavior: New empirical industrial organization methods in marketing. *International Journal of Research in Marketing*, 18(1-2):161–186, 2001.

N. Kallus and A. Zhou. Fairness, welfare, and equity in personalized pricing. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 296–314, 2021.

N. B. Keskin and A. Zeevi. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations research*, 62(5):1142–1167, 2014.

R. Kleinberg and T. Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 594–605. IEEE, 2003.

B. Kveton, M. Zaheer, C. Szepesvari, L. Li, M. Ghavamzadeh, and C. Boutilier. Randomized exploration in generalized linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2066–2076. PMLR, 2020.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

H. Li and W. T. Huh. Pricing multiple products with the multinomial logit and nested logit models: Concavity and implications. *Manufacturing & Service Operations Management*, 13(4):549–563, 2011.

G. Liberali and A. Ferecatu. Morphing for consumer dynamics: Bandits meet hidden markov models. *Marketing Science*, 41(4):769–794, 2022.

L. Luo, P. Kannan, and B. T. Ratchford. New product development under channel acceptance. *Marketing Science*, 26(2):149–163, 2007.

E. Mazumdar, A. Pacchiano, Y.-a. Ma, P. L. Bartlett, and M. I. Jordan. On thompson sampling with langevin algorithms. *arXiv preprint arXiv:2002.10002*, 2020.

D. McFadden. Econometric models for probabilistic choice among products. *Journal of Business*, pages S13–S29, 1980.

D. McFadden, A. Talvitie, S. Cosslett, I. Hasan, M. Johnson, F. Reid, and K. Train. Demand model estimation and validation. *Urban Travel Demand Forecasting Project, Phase*, 1, 1977.

N. Mehta, P. Detroja, and A. Agashe. "amazon changes prices on its products about every 10 minutes — here's how and why they do it". *Business Insider*, Aug 2018. URL https://www.businessinsider.com/amazon-price-changes-2018-8?international=true&r=US&IR=T.

S. Miao and X. Chao. Dynamic joint assortment and pricing optimization with demand learning. *Manufacturing & Service Operations Management*, 23(2):525–545, 2021.

K. Misra, E. M. Schwartz, and J. Abernethy. Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, 38(2):226–252, 2019.

J. Mueller, V. Syrgkanis, and M. Taddy. Low-rank bandit methods for high-dimensional dynamic pricing. *arXiv preprint arXiv:1801.10242*, 2018.

J. Pancras and K. Sudhir. Optimal marketing strategies for a customer data intermediary. *Journal of Marketing research*, 44(4):560–578, 2007.

A. Priester, T. Robbert, and S. Roth. A special price just for you: Effects of personalized dynamic pricing on consumer fairness perceptions. *Journal of Revenue and Pricing Management*, 19:99–112, 2020.

S. Qiang and M. Bayati. Dynamic pricing with demand covariates. *Available at SSRN 2765257*, 2016.

O. Rafieian and H. Yoganarasimhan. Targeting and privacy in mobile advertising. *Marketing Science*, 40(2): 193–218, 2021.

G. O. Roberts and R. L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

P. E. Rossi, R. E. McCulloch, and G. M. Allenby. The value of purchase history data in target marketing. *Marketing Science*, 15(4):321–340, 1996.

D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

E. M. Schwartz, E. T. Bradlow, and P. S. Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.

A. N. Smith, S. Seiler, and I. Aggarwal. Optimal price targeting. *Marketing Science*, 2022.

K. Sudhir. Structural analysis of manufacturer pricing in the presence of a strategic retailer. *Marketing Science*, 20(3):244–264, 2001.

K. E. Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

F. Trovo, S. Paladino, M. Restelli, and N. Gatti. Improving multi-armed bandit algorithms for pricing. In *16th European Conference on Multi-Agent Systems*, pages 1–15, 2018.

Y. Wang, B. Chen, and D. Simchi-Levi. Multimodal dynamic pricing. *Management Science*, 2021.

I. Weaver and V. Kumar. Nonparametric bandits leveraging informational externalities to learn the demand curve. Working Paper, 2022.

Y. Wei and Z. Jiang. Estimating parameters of structural models using neural networks. Working Paper, 2022.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

J. Xu and Y.-x. Wang. Logarithmic regret in feature-based dynamic pricing. *arXiv preprint arXiv:2102.10221*, 2021.

P. Xu, H. Zheng, E. V. Mazumdar, K. Azizzadenesheli, and A. Anandkumar. Langevin monte carlo for contextual bandits. In *International Conference on Machine Learning*, pages 24830–24850. PMLR, 2022.

H. Yoganarasimhan. Search personalization using machine learning. *Management Science*, 66(3):1045–1070, 2020.

J. Zhang and L. Krishnamurthi. Customizing promotions in online stores. *Marketing science*, 23(4):561–578, 2004.

J. Zhang and M. Wedel. The effectiveness of customized promotions in online and offline stores. *Journal of marketing research*, 46(2):190–206, 2009.

Web Appendix

A Additional Details for Related Work

We now present a more detailed discussion of the work on adaptive pricing. The related works on this topic are extensive and the papers in this area consider the following common setup. In the most common case, there is one retailer/seller who sells a single product for a fixed number of rounds. In each round, the seller first chooses a price for the product and then observes an associated demand. This demand is either a discrete variable representing a single purchase by a single customer, or a continuous variable representing aggregate expected demand over a population. Finally, the seller receives a revenue, which is simply the price times the observed demand. In general, the seller must learn the demand function through price exploration and the goal is to bound the *regret* of the pricing policy, namely the total loss in profit due to playing sub-optimal prices. The hope (in stochastic settings) is to establish pricing policies that suffer no more than $O(\sqrt{T})$ regret where T is the time horizon over which the game is played

Beyond this stylized model, many variations arising from different assumptions on the number of products, the nature of the price exploration and demand function, and the type of customers have been considered. A majority of papers in this area consider single-product non-parametric demand models (Wang et al., 2021; Besbes and Zeevi, 2015, 2009; Kleinberg and Leighton, 2003; Weaver and Kumar, 2022). To make the analysis tractable, these papers typically include assumptions on the concavity, Lipschitzness, smoothness, modality of the profit function, or that the demand function is drawn from a Gaussian process. The non-parametric setting also includes works which apply multi-armed bandit techniques to setting with just a finite set of prices or potential demand functions as in Cheung et al. (2017); Misra et al. (2019) or Kleinberg and Leighton (2003). A potential downside of naively pursuing this approach is that it leads to regret which can scale with the size of discretization which could be potentially large (Misra et al., 2019).[16]

A separate line of work (also with a single product), closer to this paper, considers settings where an explicit parametric form on the demand is assumed. Most often, the demand observed follows a generalized linear model of the price (den Boer and Zwart, 2014; Besbes and Zeevi, 2009). This level of generality allows the observed demand to be either a binary random variable which represents whether the customer purchased the product or not (Broder and Rusmevichientong, 2012; Kleinberg and Leighton, 2003), or a continuous random variable representing aggregate expected demand over a population (Keskin and Zeevi, 2014; Mueller et al., 2019; den Boer and Zwart, 2014). We also refer the reader to the survey (Den Boer, 2015) which discusses several older works.

Fewer works have considered the setting of multiple possible products available to a given user at any time. Notable among them are Keskin and Zeevi (2014); Mueller et al. (2018); Javanmard et al. (2020); Miao and Chao (2021) and Goyal and Perivier (2021). Even fewer works have considered the problem of adaptively setting pricing under a multinomial model. In particular, Javanmard et al. (2020); Goyal and Perivier (2021); Miao and Chao (2021) are the closest to our work. Firstly Javanmard et al. (2020), also assumes a multinomial response model, but unlike our setting where we assumed a fixed set of items each round, they assume that the set of items changes each round and that the seller receives covariate information for each item. Their proposed M3P algorithm for pricing in this setting alternates between rounds of playing random prices and then greedily playing the optimal price according to the estimated demand model at that time. We quickly remark that the algorithm of Goyal and Perivier (2021) parallels MLE-Cycle algorithm of Broder and Rusmevichientong (2012). The second work, Miao and Chao (2021) proposes an algorithm that runs in cycles, with each cycle restarting when the outside option is selected. During a cycle, a hybrid Thompson Sampling and price shock method is to choose the fixed prices played for the whole round. It is

[16] A notable exception to this is Kleinberg and Leighton (2003) which uses a discretization adapted to the time horizon and obtains a regret bound independent of the size of the discretization.

not clear how to extend their method for the case of promotion, or for the batch settings we consider. As we discuss next, both of these works rely on a degree of *forced exploration*.

Forced exploration is often used in the pricing literature to guarantee sufficient exploration to learn the underlying parameters of the pricing problem. In general, forced exploration algorithms come in two forms. In the first method the exploration is done through totally uniform prices or a fixed set of prices (Broder and Rusmevichientong, 2012; Javanmard et al., 2020). In the second approach, the Hessian of the log-likelihood of the pricing problem is computed and then, price exploration is designed to ensure that the minimum eigenvalue of the Hessian grows at a rate of $\Omega(\sqrt{T})$ (Keskin and Zeevi, 2014; den Boer and Zwart, 2014; Miao and Chao, 2021). From a computational perspective, computing the minimum eigenvalue of the Hessian can be computationally expensive. In addition, from an algorithmic design perspective, changes in the underlying model can lead to a change in the computation of the Hessian, which then leads to a change in the choice of prices. As a result, these methods are less flexible and may require onerous computation to modify. In contrast to these works, our Thompson Sampling algorithm relies on *model-based exploration*. As we will see, our approach avoids explicit Hessian computation and as a result, is far more plug-and-play.

Finally, a recent line of work considers the linear parametric contextual setting, where at each time additional covariate information about products, customers, or market information is revealed (Javanmard and Nazerzadeh, 2016; Qiang and Bayati, 2016; Ban and Keskin, 2021; Xu and Wang, 2021; Javanmard, 2017; Javanmard et al., 2020; Shah et al., 2019). This additional information is allowed to affect the underlying utility obtained through an additive shift or multiplicatively on the price sensitivity by a linear function of the context. One common feature of all these works is that (with the exception of Javanmard et al. (2020), discussed below) they consider a setting with a single product. In contrast, in this paper, we consider a more general case, where we allow the utility function to potentially be an unknown non-linear function of the context. We adapt our Thompson sampling approach to this setting and show promising empirical performance modeling the unknown utility as a neural network.

We quickly point out three other works in the multi-product setting that use Thompson sampling for price exploration, namely Ganti et al. (2018); Ferreira et al. (2018); Bastani et al. (2021). The former considers a stylized demand model that does not take product cross-elasticities into account. The second considers a discrete set of prices and focuses on the problem of effective inventory management. The final paper considers the case of a linear demand model and focuses on the problem of prior misspecification where the definition of regret is with respect to an algorithm that has a correctly specified prior. None of these works consider the setting of multiple products with promotions and none of them extend to a setting with non-linear utility.

B Appendix for Section 4.1

B.1 Proof of Lemma 1

The proof here follows the same procedure as Proposition 3.1 in Javanmard et al. (2020). Let's denote $e_i = \exp(\alpha_i - \beta_i p_i + \gamma_i x_i)$ and $G(e) = \sum_{i\in[K]} e_i$. In this notation we could rewrite Equation (3) as

$$R_\theta(\mathbf{p}, \mathbf{x}) = \sum_{\ell=1}^{K} (p_\ell - m_\ell) \frac{e_\ell}{1 + G(e)},$$

taking derivatives with respect to p_i we would get

$$\begin{aligned}
\frac{\partial R_\theta(\mathbf{p}_\star, \mathbf{x})}{\partial p_i} &= \frac{e_i - \beta_i e_i (p_{\star,i} - m_i)}{1 + G(e)} + \frac{(\sum_{\ell=1}^{K} (p_{\star,\ell} - m_\ell) e_\ell) e_i \beta_i}{(1 + G(e))^2} \\
&= \frac{\beta_i e_i}{1 + G(e)} \left(\frac{1}{\beta_i} + (p_{\star,i} - m_i) + \frac{\sum_{\ell=1}^{K} (p_{\star,\ell} - m_\ell) e_\ell}{1 + G(e)} \right) = 0
\end{aligned}$$

Because e_i is non-zero, the expression inside the brace should be zero. We could also use the fact that $\frac{\sum_{\ell=1}^{K}(p_{\star,\ell}-m_\ell)e_\ell}{1+G(e)} = R_\theta(\mathbf{p}_\star, \mathbf{x})$. Together we get

$$p_{\star,i} - m_i = \frac{1}{\beta_i} + R_\theta(\mathbf{p}_\star, \mathbf{x}).$$

For simplicity we use $R = R_\theta(\mathbf{p}_\star, \mathbf{x})$, then multiply the previous equation by e_i and sum over all $i \in [K]$. We will get

$$\begin{aligned}
\sum_{i=1}^{K}(p_{\star,i} - m_i)e_i = \sum_{i=1}^{K}\frac{e_i}{\beta_i} + R\sum_{i=1}^{K}e_i \implies \\
R(1+G(e)) = \sum_{i=1}^{K}\frac{e_i}{\beta_i} + RG(e) \implies \\
R = \sum_{i=1}^{K}\frac{e_i}{\beta_i} = \sum_{i=1}^{K}\frac{\exp(\alpha_i - \beta_i p_i + \gamma_i x_i)}{\beta_i} \\
= \sum_{i=1}^{K}\frac{1}{\beta_i}e^{\alpha_i - \beta_i(m_i + \frac{1}{\beta_i} + R) + \gamma_i x_i} \\
= \sum_{i=1}^{K}\frac{1}{\beta_i}e^{-(1+R\beta_i + m_i\beta_i)}e^{\alpha_i + \gamma_i x_i},
\end{aligned}$$

which is the fixed point (Equation (8)). This equation has a unique solution because the left-hand side is increasing and starts from zero, and the right-hand side is positive at $R = 0$ and decreasing.

B.2 Algorithm for Fixed Point

We now provide the algorithm used to find the fixed point of Equation (8).

Algorithm 4 Method for finding the fixed point of Equation (8)

Input: ϵ the accepted error in the result

Find an R_m where $R_m > \sum_{i=1}^{K}\frac{1}{\beta_i}e^{-(1+\beta_i R_m + \beta_i m_i)}e^{\alpha_i + \gamma_i x_i}$

  Set $R_m = 1$

  while $R_m < \sum_{i=1}^{K}\frac{1}{\beta_i}e^{-(1+R_m\beta_i + m_i\beta_i)}e^{\alpha_i + \gamma_i x_i}$ **do**

    $R_m = R_m * 2$

  end while

Find the solution between 0 and R_m with maximum ϵ error using binary search

  Set $l = 0, r = R_m$

  while $r - l > \epsilon$ **do**

    $m = \frac{r+l}{2}$

    if $m < \sum_{i=1}^{K}\frac{1}{\beta_i}e^{-(1+m\beta_i + m_i\beta_i)}e^{\alpha_i + \gamma_i x_i}$ **then**

      $l = m$

    else

      $r = m$

    end if

  end while

return l

B.3 Proof of Theorem 1

We first prove that there is no critical point in the interior. For a positive constant $c > 1$ define

$$R_c(\mathbf{p}, \mathbf{x}) = \sum_{i=1}^{K} \frac{(p_{it} - m_i) \exp(\alpha_i - \beta_i p_{it} + \gamma_i x_{it})}{c + \sum_{k=1}^{K} \exp(\alpha_k - \beta_k p_{kt} + \gamma_k x_{kt})} \tag{A1}$$

In particular, $R_1(\mathbf{p}, \mathbf{x}) = R(\mathbf{p}, \mathbf{x})$. Again, let's denote $e_i = \exp(\alpha_i - \beta_i p_i + \gamma_i x_i)$ and $G(e) = \sum_{i \in [K]} e_i$. The partial derivative of $R_c(\mathbf{p}, \mathbf{x})$ with respect to x_i is

$$\begin{aligned}
\frac{\partial R_c(\mathbf{p}, \mathbf{x})}{\partial x_i} &= \frac{(p_i - m_i)\gamma_i e_i}{c + G(e)} - \frac{(\sum_{\ell=1}^{K}(p_\ell - m_\ell)e_\ell)\gamma_i e_i}{(c + G(e))^2} \\
&= \frac{\gamma_i e_i}{c + G(e)} \left(p_i - m_i - \frac{\sum_{\ell=1}^{K}(p_\ell - m_\ell)e_\ell}{c + G(e)} \right) \\
&= \frac{\gamma_i e_i}{c + G(e)} \left(p_i - m_i - R_c(\mathbf{p}, \mathbf{x}) \right)
\end{aligned}$$

Now firstly we claim that $R_c(\mathbf{p}, \mathbf{x})$ has no critical points. For there to be a critical point, by the assumption we have $\gamma_i > 0$ that gives $p_i - m_i = R_c(\mathbf{p}, \mathbf{x})$ for each i. In particular this implies,

$$\begin{aligned}
R_c(\mathbf{p}, \mathbf{x}) &= \sum_{i=1}^{K} \frac{(p_i - m_i)e_i}{c + G(e)} \\
&= \sum_{i=1}^{K} \frac{(p_i - m_i)e_i}{c + G(e)} \\
&= R_c(\mathbf{p}, \mathbf{x}) \sum_{i=1}^{K} \frac{e_i}{c + G(e)} \\
&= R_c(\mathbf{p}, \mathbf{x}) \frac{G(e)}{c + G(e)} \\
&< R_c(\mathbf{p}, \mathbf{x})
\end{aligned}$$

which is a contradiction.

Now we show the box constraint. Denoting the extremum as $\mathbf{x}_* := \mathbf{x}_*(\mathbf{p})$ and let $\mathbf{x}_* = (x_1^*, \cdots, x_K^*)$. We use proof by contradiction. Suppose at a maximum there exists some $i \in [K]$ such that $x_i^* \in (0, B)$. Then,

$$\begin{aligned}
R_c(\mathbf{p}, \mathbf{x}) &= \sum_{k=1}^{K} \frac{(p_k - m_k)e_k}{c + G(e)} \\
&= \frac{\sum_{k \neq i}^{K}(p_k - m_k)e_k + (p_i - m_i)e_i}{c + \sum_{k \neq i}^{K} e_k + e_i}.
\end{aligned}$$

Let $\tilde{a} = \sum_{k \neq i}^{K}(p_k - m_k)e_k^*$ and $\tilde{c} = c + \sum_{k \neq i}^{K} e_k^*$. Consider

$$f(x_i) = R_c(\mathbf{p}, (x_1^*, \cdots, x_{i-1}^*, x_i, x_{i+1}^*, \cdots, x_K^*)) = \frac{\tilde{a} + (p_i - m_i)e_i}{\tilde{c} + e_i} = p_i - m_i + \frac{\tilde{a} - (p_i - m_i)\tilde{c}}{\tilde{c} + e_i}.$$

This is a monotonic function of e_i and thus x_i. Therefore, $\tilde{x}_i := \arg\max_{x_i \in [0,B]} f(x_i) \in \{0, B\}$. Then,

let $\tilde{\mathbf{x}} := (x_1^*, \cdots, x_{i-1}^*, x_i, x_{i+1}^*, \cdots, x_K^*)$, we have $f(\tilde{x}_i) > f(x_i^*)$ and so $R_c(\mathbf{p}, \tilde{\mathbf{x}}) > R_c(\mathbf{p}, \mathbf{x}_*)$, which contradicts with the fact that $\mathbf{x}_*$ is the maximum.

Now we would like to show the simplex case. Denoting the extremum as $\mathbf{x}_* \in \triangle_K$ and let $\mathbf{x}_* = (x_1^*, \cdots, x_K^*)$. Without loss of generality, we could assume $m_i = 0$ for all i, since we could always shift the prices p_{it} and α_i to incorporate m_i, and we omit the dependence on t. Then, let $\tilde{x}_i = x_i + \frac{\alpha_i - \beta_i p_i}{\gamma_i}$. Since $\{x_i\}_{i=1}^K \in \Delta_K$, we have that $\sum_i x_i \leq 1$, and so $\sum_i \tilde{x}_i \leq 1 + \sum_i \frac{\alpha_i - \beta_i p_i}{\gamma_i}$. Then, we write down the Lagrangian of $R_c(\mathbf{p}, \mathbf{x})$ as

$$\mathcal{L}_c(\mathbf{p}, \mathbf{x}) = \sum_{i=1}^K \frac{p_i \exp(\gamma_i \tilde{x}_i)}{c + \sum_{k=1}^K \exp(\gamma_k \tilde{x}_k)} + \lambda\Big(\sum_{i=1}^K \tilde{x}_i - \Big(1 + \sum_i \frac{\alpha_i - \beta_i p_i}{\gamma_i}\Big)\Big). \tag{A2}$$

Let $g(\mathbf{x}) = \sum_{i=1}^K \mathbf{x}_i - \Big(1 + \sum_i \frac{\alpha_i - \beta_i p_i}{\gamma_i}\Big)$. By primal feasibility, a critical point must satisfiy that $\nabla_{\mathbf{x}} \mathcal{L}_c(\mathbf{p}, \mathbf{x}) = 0$. Let e_i and $G(e)$ be defined as before, a computation shows

$$\frac{\partial \mathcal{L}_c(\mathbf{p}, \mathbf{x})}{\partial x_i} = \frac{\gamma_i e_i}{c + G(e)}(p_i - R_c(\mathbf{p}, \mathbf{x})) + \lambda.$$

Setting all of them equals zero implies that $\gamma_i e_i (p_i - R_c(\mathbf{p}, \mathbf{x})) = -\lambda$ for all i. We first show that $\lambda < 0$ and so $-\lambda > 0$. If $-\lambda < 0$, we have $p_i < R_c(\mathbf{p}, \mathbf{x})$ for all i, so

$$\begin{aligned} R_c(\mathbf{p}, \mathbf{x}) &= \sum_{i=1}^K \frac{p_i e_i}{c + G(e)} \\ &< R_c(\mathbf{p}, \mathbf{x}) \sum_{i=1}^K \frac{e_i}{c + G(e)} \\ &= R_c(\mathbf{p}, \mathbf{x}) \frac{G(e)}{c + G(e)} \\ &< R_c(\mathbf{p}, \mathbf{x}) \end{aligned}$$

which is a contradiction. Then we consider the Hessian at a critical point. Note that

$$\frac{\partial^2 R_c(\mathbf{p}, \mathbf{x}_*)}{\partial x_i^2} = \frac{\gamma_i^2 e_i (1 + G(e)) - \gamma_i e_i \cdot \gamma_i e_i}{(1 + G(e))^2}(p_i - R_c(\mathbf{p}, \mathbf{x}_*)) + \frac{\gamma_i e_i}{1 + G(e)} \cdot \frac{\gamma_i e_i}{1 + G(e)}(R_c(\mathbf{p}, \mathbf{x}_*) - p_i)$$

and

$$\frac{\partial^2 R_c(\mathbf{p}, \mathbf{x}_*)}{\partial x_i \partial x_j} = \frac{-\gamma_i e_i \cdot \gamma_j e_j}{(1 + G(e))^2}(p_i - R_c(\mathbf{p}, \mathbf{x}_*)) + \frac{\gamma_i e_i}{1 + G(e)} \cdot \frac{\gamma_j e_j}{1 + G(e)}(R_c(\mathbf{p}, \mathbf{x}_*) - p_j).$$

Let the Hessian be H. Plugging in the critical point condition gives us

$$\begin{aligned} H_{ii} &= \frac{-\lambda}{(1 + G(e))^2}\Big(\gamma_i(1 + G(e)) - 2\gamma_i e_i\Big), \\ H_{ij} &= \frac{-\lambda}{(1 + G(e))^2}\Big(-\gamma_i e_i - \gamma_j e_j\Big). \end{aligned}$$

Therefore, for some direction $a \in \mathbb{R}^K$,

$$\begin{aligned} a^\top H a &= \sum_i \sum_j a_i H_{ij} a_j \\ &= \sum_i a_i^2 \gamma_i (1 + G(e) - 2e_i) - \sum_i \sum_{j \neq i} a_i a_j (\gamma_i e_i + \gamma_j e_j) \\ &= \sum_i a_i^2 \gamma_i (1 + \sum_{k \neq i} e_k) - \sum_i \sum_j a_i a_j (\gamma_i e_i + \gamma_j e_j). \end{aligned}$$

Assume that $K \geq 2$ and consider $\boldsymbol{a} = (a_1, a_2, 0, \cdots, 0)$. Then

$$a^\top H a = a_1^2 \gamma_1 \Big(1 + \sum_{k=2}^K e_k\Big) + a_2^2 \gamma_2 \Big(1 + e_1 + \sum_{k=3}^K e_k\Big) - a_1^2 \gamma_1 e_1 - a_2^2 \gamma_2 e_2 - 2 a_1 a_2 (\gamma_1 e_1 + \gamma_2 e_2).$$

If $a_1 = 1$, $a_2 = -1$, then the above becomes

$$\gamma_1 \Big(1 + \sum_{k=2}^K e_k\Big) + \gamma_2 \Big(1 + e_1 + \sum_{k=3}^K e_k\Big) + \gamma_1 e_1 + \gamma_2 e_2 > 0.$$

Then, by Taylor expansion,

$$\mathcal{L}_c(\mathbf{p}, \mathbf{x}_* + \boldsymbol{a}) = \mathcal{L}_c(\mathbf{p}, \mathbf{x}_*) + \boldsymbol{a}^\top \nabla_x \mathcal{L}_c(\mathbf{p}, \mathbf{x}_*) + \boldsymbol{a}^\top \nabla_x^2 \mathcal{L}_c(\mathbf{p}, \mathbf{x}_*) \boldsymbol{a} + O(\|\boldsymbol{a}\|^3).$$

Let $\boldsymbol{a} = (1, -1, 0, \cdots, 0)$. Note that $\nabla_x \mathcal{L}_c(\mathbf{p}, \mathbf{x}_*) = \mathbf{0}$ and $\nabla_x^2 \mathcal{L}_c(\mathbf{p}, \mathbf{x}_*) = \nabla_x^2 R_c(\mathbf{p}, \mathbf{x}_*)$, we have that $\mathcal{L}_c(\mathbf{p}, \mathbf{x}_* + \boldsymbol{a}) > \mathcal{L}_c(\mathbf{p}, \mathbf{x}_*)$. Also, by complementary slackness, $\lambda g(\mathbf{x}_*) = 0$ and so $\lambda g(\mathbf{x}_* + \boldsymbol{a}) = 0$. Therefore, $R_c(\mathbf{p}, \mathbf{x}_* + \boldsymbol{a}) > R_c(\mathbf{p}, \mathbf{x}_*)$. Also, by definition of $\boldsymbol{a}$, $(\mathbf{x}_* + \boldsymbol{a}) \in \triangle_K$. Therefore, any critical point cannot be the global maximum.

Then we prove the statement by contradiction. Without loss of generality, assume that $x_1^* \geq x_2^* \geq \cdots \geq x_K^* \geq 0$. Assume $x_1^* \geq x_2^* \geq \cdots \geq x_m^* > 0 = x_{m+1}^* = \cdots = x_K^*$ for $m \geq 2$. Let $\mathbf{x}_m^* = (x_1, \cdots, x_m)$. Then, $\mathbf{x}_m^* = \arg\max_{\mathbf{x}_m \in \triangle_m} R_c(\mathbf{p}, (\mathbf{x}_m^*, \mathbf{0}))$. Let $G_m(e) = \sum_{i=1}^m e_i$ and $R_{c,s}(\mathbf{p}, \mathbf{x}_m) = \frac{s + \sum_{i=1}^m p_i e_i}{c + G_m(e)}$. Then,

$$\begin{aligned} \frac{\partial R_{c,s}(\mathbf{p}, \mathbf{x}_m)}{\partial x_i} &= \frac{-\left(s + \sum_{i=1}^m p_i e_i\right) \gamma_i e_i + p_i \gamma_i e_i \cdot (c + G_m(e))}{(c + G_m(e))^2} \\ &= \frac{\gamma_i e_i}{c + G_m(e)} \left(p_i - \frac{s + \sum_{i=1}^m p_i e_i}{c + G_m(e)}\right) = \frac{\gamma_i e_i}{c + G_m(e)} \left(p_i - R_{c,s}(\mathbf{p}, \mathbf{x}_m)\right). \end{aligned}$$

It can be shown that the Hessian of $R_{c,s}(\mathbf{p}, \mathbf{x}_m)$ is of the same form as $R_c(\mathbf{p}, \mathbf{x}_m)$. In particular, let $\tilde{H}$ be the Hessian of $R_{c,s}(\mathbf{p}, \mathbf{x}_m)$, we have

$$\begin{aligned} \tilde{H}_{ii} &= \frac{-\lambda}{(1 + G_m(e))^2} \Big(\gamma_i (1 + G_m(e)) - 2\gamma_i e_i\Big), \\ \tilde{H}_{ij} &= \frac{-\lambda}{(1 + G_m(e))^2} \Big(-\gamma_i e_i - \gamma_j e_j\Big). \end{aligned}$$

Then, by above argument, there exists some direction that increases the function value, which contradicts with the fact that $\mathbf{x}_*$ is the maximum. Therefore, the maximum must be attained at the boundary.

Algorithm 5 Procedure for finding the best price and promotion given a parameters θ

Input: $\theta, \mathcal{X}$
  Set $R_{max} = -1$
  for all $x \in \mathcal{X}$ **do**
    Find the revenue maximizing price p using Lemma 1 and get revenue R_x for this x
    if $R_{max} < R_x$ **then**
      $x_\star = x$
      $p_\star = p$
    end if
  end for
  return $p_\star, x_\star$

B.4 Example of Optimal Promotion

For the case of the simplex constraint, Figure A1 illustrates Theorem 1 using a three-product example at three different price vectors $\mathbf{p}$. The parameters for this demand setting and the price vectors considered are shown in Table A1. The Figure shows the heatmap of the revenue function for different possibilities of the promotion vector $\mathbf{x}$ for three price configurations. As we can see, the Figure shows that the firm should completely use the promotion on one of the products, where the choice of the product depends on the price vector of the products. In particular, the promotion does not need to be on the product with the highest price; e.g., for the price vector $\mathbf{p} = (3, 9, 20)$ the optimal promotion is $\mathbf{x} = (0, 1, 0)$.

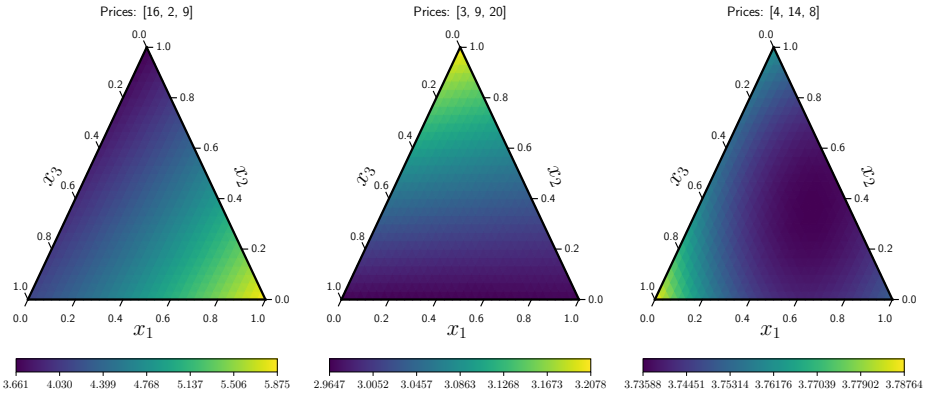


Figure A1: Heatmap of revenue for each possible promotion vector in a simplex over three products ($B = 1$). Parameters of the model are given in Table A1.

	Parameters	Product 1	Product 2	Product 3
	α	1	1	1
	β	.1	.2	.3
	γ	.8	.3	.5
Case 1	$\mathbf{p}$	\$16	\$2	\$9
	$\mathbf{x}_*$	1	0	0
Case 2	$\mathbf{p}$	\$3	\$9	\$20
	$\mathbf{x}_*$	0	1	0
Case 3	$\mathbf{p}$	\$4	\$14	\$8
	$\mathbf{x}_*$	0	0	1

Table A1: Model parameters and optimal promotions at three different choices of prices. All the marginal costs are set to zero.

C Proof of Theorem 2

We state a full version of Theorem 2 that includes specific constants.

Theorem 4. *Assume that* $\max_{x\in\mathcal{X}} \|x\|_\infty \leq B$*, the largest price is bounded by* u*,* $\|\theta_*\| \leq S$ *with probability 1, and the regularization in Algorithm 1 is* λ*. With probability at least* $1-\delta$*, the regret of Algorithm 1 is bounded by*

$$Reg_T^B \lesssim Su\left(\sqrt{\lambda} + \frac{1}{\sqrt{\lambda}}\log\left(\frac{T(\lambda+T/3)^{1.5K}\lambda^{1.5K}}{\delta}\right) + \frac{K}{\sqrt{\lambda}}\right)\sqrt{\kappa KT\log\left(\frac{\lambda+T(u^2+B^2+1)}{\lambda^{1/K}}\right)}.$$

Corollary 1. *In the same setting as Theorem 4, if Algorithm 1 is run with regularization* $\lambda = O\left(K\log\left(KT(u^2+B^2+1)\right)\right)$ *and* $\delta = 1/T$*. Then the regret of Algorithm 1 is bounded in expectation by*

$$\widetilde{O}\left(SuK\sqrt{\kappa T}\log\left(KT(u^2+B^2+1)\right)\right)$$

where $\widetilde{O}(\cdot)$ *hides constants and doubly logarithmic factors.*

Confidence Interval: We begin by formally defining the λ-regularized maximum likelihood estimator at time t, $\hat{\theta}_t^\lambda$. Given a dataset $\{\mathbf{p}_s, \mathbf{x}_s, I_s\}_{s=1}^t$ where $\mathbf{p}_s$ denotes the price vector played at time s, $\mathbf{x}_s$ the marketing mix, and $I_s \in \{0, 1, \ldots, K\}$ is the item selected,

$$\hat{\theta}_t{}^\lambda := \arg\min_{\theta\in\mathbb{R}^{3K}} \sum_{s=0}^{t}\sum_{i=0}^{K} \mathbf{1}\{I_s = i\}\log\left(\mu_i\left(\theta, \overline{\mathbf{p}}_s\right)\right) - \frac{\lambda}{2}\|\theta\|_2^2 \tag{A3}$$

where $\overline{\mathbf{p}}_s = [1, \mathbf{p}_s, \mathbf{x}_s] \in \mathbb{R}^{3K}$, and μ_i is the multinomial probability of item i

$$\mu_i(\theta, \overline{\mathbf{p}}) = \mathbb{P}_\theta(I = i|\mathbf{p}, \mathbf{x}) = \frac{\exp(\alpha_i - \beta_i p_i + \gamma_i x_i)}{1 + \sum_{k=1}^K \exp(\alpha_k - \beta_k p_k + \gamma_k x_k)}. \tag{A4}$$

We have the following concentration result.

Lemma 2 (After Lemmas 5, 12 from Agrawal et al. (2020))**.** *Let* $\hat{\theta}_t^\lambda$ *solve Equation* (A3)*. Fix* $\delta > 0$*,* $\lambda > 0$ *and* $T < \infty$*. Let* $\|\theta\|_2 \leq S$*. Define the event* $\mathcal{E}$ *as*

$$\mathcal{E} := \bigcap_{t=1}^{T}\{\|\hat{\theta}_t^\lambda - \theta_*\|_{V_t(\hat{\theta}_t^\lambda)} \leq \psi_t(\delta)\}.$$

$\mathcal{E}$ holds with probability at least $1-\delta$, where

$$V_t(\theta) := \sum_{s=1}^{t-1} \dot{\mu}_i(\theta, \overline{\mathbf{p}}_s)\overline{\mathbf{p}}_s\overline{\mathbf{p}}_s^\top + \lambda \mathbf{I}$$

for $\dot{\mu}_i := \nabla_\theta \mu_i(\theta, \overline{\mathbf{p}})$ and

$$\psi_t(\delta) := 2(1+2S)\left(\frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}}\log\left(\frac{T(\lambda + t/3)^{1.5K}\lambda^{1.5K}}{\delta}\right) + \frac{6K}{\sqrt{\lambda}}\log(2)\right).$$

Remark 1. *While Agrawal et al. (2019) assume that $\mathbf{x}_t$ is i.i.d. from a context distribution, we note that the same bound holds for adaptively chosen sequences as they do in Faury et al. (2020).*

Upper bound using the confidence interval:

For the following, we fix some time t and suppress the dependence on t to simplify notation. Define the following upper bound on the revenue at time t as

$$U(\hat{\theta}_t^\lambda, \mathbf{p}, \mathbf{x}) = \sum_{i=1}^K p_i \frac{\exp\left((\hat{\theta}_t^\lambda)^\top \overline{\mathbf{p}}_i + \psi_t(\delta)\|\overline{\mathbf{p}}_i\|_{V_t(\hat{\theta}_t^\lambda)^{-1}}\right)}{1 + \sum_{j=1}^K \exp\left((\hat{\theta}_t^\lambda)^\top \overline{\mathbf{p}}_j + \psi_t(\delta)\|\overline{\mathbf{p}}_i\|_{V_t(\hat{\theta}_t^\lambda)^{-1}}\right)}$$

$$U_t(\mathbf{p}, \mathbf{x}) = \max_{\theta \in C_t} \sum_{i=1}^K p_i \frac{\exp\left(\theta^\top \overline{\mathbf{p}}_i\right)}{1 + \sum_{j=1}^K \exp\left(\theta^\top \overline{\mathbf{p}}_j\right)}$$

where we recall $\overline{\mathbf{p}}$ is a function of both $\mathbf{p}$ and $\mathbf{x}$, and we define $\overline{\mathbf{p}}_i$ to be $(1, p_{si}, x_{si})$ for the i^{th} set of 3 entries and zero elsewhere. We may decompose the regret at time $t \leq T$ conditioned on the filtration $\mathcal{F}_{t-1}$ as

$$\begin{aligned}
&\mathbb{E}[\text{Regret}(t) - \text{Regret}(t-1)|\mathcal{F}_{t-1}] \\
&\quad = \mathbb{E}[R(\mathbf{p}_*, \mathbf{x}_*, \theta_*) - R(\mathbf{p}_t, \mathbf{x}_t, \theta_*)|\mathcal{F}_{t-1}] \\
&\quad = \mathbb{E}[R(\mathbf{p}_*, \mathbf{x}_*, \theta_*) - U(\mathbf{p}_*, \mathbf{x}_*, \hat{\theta}_t^\lambda)|\mathcal{F}_{t-1}] + \mathbb{E}[U(\mathbf{p}_*, \mathbf{x}_*, \hat{\theta}_t^\lambda) - U(\mathbf{p}_t, \mathbf{x}_t, \hat{\theta}_t^\lambda)|\mathcal{F}_{t-1}] \\
&\qquad + \mathbb{E}[U(\mathbf{p}_t, \mathbf{x}_t, \hat{\theta}_t^\lambda) - R(\mathbf{p}_t, \mathbf{x}_t, \theta_*)|\mathcal{F}_{t-1}] \\
&\quad = \mathbb{E}[R(\mathbf{p}_*, \mathbf{x}_*, \theta_*) - U(\mathbf{p}_t, \mathbf{x}_t, \hat{\theta}_t^\lambda)|\mathcal{F}_{t-1}] + \mathbb{E}[U(\mathbf{p}_t, \mathbf{x}_t, \hat{\theta}_t^\lambda) - R(\mathbf{p}_t, \mathbf{x}_t, \theta_*)|\mathcal{F}_{t-1}]
\end{aligned}$$

where the final inequality holds by noting that $\mathbb{E}[U(\mathbf{p}_*, \mathbf{x}_*, \hat{\theta}_t^\lambda) - U(\mathbf{p}_t, \mathbf{x}_t, \hat{\theta}_t^\lambda)|\mathcal{F}_{t-1}] = 0$ since $\hat{\theta}_t^\lambda$ is deterministic conditioned on $\mathcal{F}_{t-1}$ and $\mathbf{p}_*$ and $\mathbf{p}_t$ as well as $\mathbf{x}_*$ and $\mathbf{x}_t$ are identically distributed since $\mathbf{p}_t$ and $\mathbf{x}_t$ are sampled according to the posterior at time t. Therefore,

$$\begin{aligned}
\mathbb{E}[\text{Regret}(T)] &= \sum_{t=1}^T \mathbb{E}[\text{Regret}(t) - \text{Regret}(t-1)|\mathcal{F}_{t-1}] \\
&= \sum_{t=1}^T \mathbb{E}[R(\mathbf{p}_*, \mathbf{x}_*, \theta_*) - U(\mathbf{p}_*, \mathbf{x}_*, \hat{\theta}_t^\lambda)|\mathcal{F}_{t-1}] + \mathbb{E}[U(\mathbf{p}_t, \mathbf{x}_t, \hat{\theta}_t^\lambda) - R(\mathbf{p}_t, \mathbf{x}_t, \theta_*)|\mathcal{F}_{t-1}] \\
&= \sum_{t=1}^T \mathbb{E}[R(\mathbf{p}_*, \mathbf{x}_*, \theta_*) - U(\mathbf{p}_*, \mathbf{x}_*, \hat{\theta}_t^\lambda)|\mathcal{F}_{t-1}] + \sum_{t=1}^T \mathbb{E}[U(\mathbf{p}_t, \mathbf{x}_t, \hat{\theta}_t^\lambda) - R(\mathbf{p}_t, \mathbf{x}_t, \theta_*)|\mathcal{F}_{t-1}].
\end{aligned}$$

We refer to the first summation as R_1 and the second as R_2 and bound them independently.

Bounding R_1: On the good event $\mathcal{E}$ defined in Lemma 2 we have that

$$\theta_*^\top \overline{\mathbf{p}}_i \le (\hat{\theta}_t^\lambda)^\top \overline{\mathbf{p}}_i + \psi_t(\delta)\|\mathbf{p}\|_{V_t(\hat{\theta}_t^\lambda)^{-1}}, \text{for all } i \in [K], t \ge 1$$

Thus by Lemma 3 we have that $R(\mathbf{p}_*, \mathbf{x}_*, \theta_*) - U(\mathbf{p}_*, \mathbf{x}_*, \hat{\theta}_t^\lambda) \le 0$. Thus,

$$\sum_{t=1}^T \mathbb{E}[R(\mathbf{p}_*, \mathbf{x}_*, \theta_*) - U(\mathbf{p}_t, \mathbf{x}_t, \hat{\theta}_t^\lambda)| \cap \mathcal{F}_{t-1}] \le u\mathbb{P}(\mathcal{E}^c) \le u\delta \le O(u)$$

where u is an upper bound on the largest price.
Bound R_2: As in the bound on R_1,

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}[U(\mathbf{p}_t, \mathbf{x}_t, \hat{\theta}_t^\lambda) - R(\mathbf{p}_t, \mathbf{x}_t, \theta_*)|\mathcal{F}_{t-1}] &= \sum_{t=1}^T \mathbb{E}[\mathbf{1}(\mathcal{E})(U(\mathbf{p}_t, \mathbf{x}_t, \hat{\theta}_t^\lambda) - R(\mathbf{p}_t, \mathbf{x}_t, \theta_*))|\mathcal{F}_{t-1}] \\ &\quad + \sum_{t=1}^T \mathbb{E}[\mathbf{1}(\mathcal{E}^c)(U(\mathbf{p}_t, \mathbf{x}_t, \hat{\theta}_t^\lambda) - R(\mathbf{p}_t, \mathbf{x}_t, \theta_*))|\mathcal{F}_{t-1}] \\ &\lesssim u + \sum_{t=1}^T \mathbb{E}[\mathbf{1}(\mathcal{E})(U(\mathbf{p}_t, \mathbf{x}_t, \hat{\theta}_t^\lambda) - R(\mathbf{p}_t, \mathbf{x}_t, \theta_*))|\mathcal{F}_{t-1}].\end{aligned}$$

Let $\mathbb{E}[\cdot|\mathcal{F}_{t-1}] =: \mathbb{E}_t[\cdot]$ for brevity. Then we apply the Lipschitz property from Lemma 3 to show that

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}_t[\mathbf{1}(\mathcal{E})(U(\mathbf{p}_t, \mathbf{x}_t, \hat{\theta}_t^\lambda) - R(\mathbf{p}_t, \mathbf{x}_t, \theta_*))] &\le u\sum_{t=1}^T \mathbb{E}_t[\mathbf{1}(\mathcal{E}) \max_\ell |\overline{\mathbf{p}}_\ell^\top(\hat{\theta}_t^\lambda - \theta_*)|] \\ &= u\sum_{t=1}^T \mathbb{E}_t[\mathbf{1}(\mathcal{E}) \max_\ell \overline{\mathbf{p}}_\ell^\top(\hat{\theta}_t^\lambda - \theta_*)]\end{aligned}$$

where the final inequality holds due to $\mathcal{E}$ since

$$\theta_*^\top \overline{\mathbf{p}}_i \le (\hat{\theta}_t^\lambda)^\top \overline{\mathbf{p}}_i + \psi_t(\delta)\|\mathbf{p}\|_{V_t(\hat{\theta}_t^\lambda)^{-1}}, \text{for all } i \in [K], t \ge 1$$

on $\mathcal{E}$. Then by Cauchy-Schwarz,

$$\begin{aligned}u\sum_{t=1}^T \mathbb{E}_t[\mathbf{1}(\mathcal{E}) \max_\ell \overline{\mathbf{p}}_\ell^\top(\hat{\theta}_t^\lambda - \theta_*)] &\le u\sum_{t=1}^T \mathbb{E}_t[\mathbf{1}(\mathcal{E}) \max_\ell \|\overline{\mathbf{p}}_\ell\|_{V_t(\hat{\theta}_t^\lambda)^{-1}} \|\hat{\theta}_t^\lambda - \theta_*\|_{V_t(\hat{\theta}_t^\lambda)}] \\ &\le u\sum_{t=1}^T \psi_t(\delta)\mathbb{E}_t[\max_\ell \|\overline{\mathbf{p}}_\ell\|_{V_t(\hat{\theta}_t^\lambda)^{-1}}] \\ &\le u\psi_T(\delta)\sum_{t=1}^T \mathbb{E}_t[\max_\ell \|\overline{\mathbf{p}}_\ell\|_{V_t(\hat{\theta}_t^\lambda)^{-1}}]\end{aligned}$$

where the penultimate inequality follows from the definition of $\mathcal{E}$ and the final inequality holds since $\psi_t(\delta)$ is an increasing sequence in t. By second application of Cauchy-Schwarz,

$$u\psi_T(\delta)\sum_{t=1}^T \mathbb{E}_t[\max_\ell \|\overline{\mathbf{p}}_\ell\|_{V_t(\hat{\theta}_t^\lambda)^{-1}}] \le u\psi_T(\delta)\sqrt{T\sum_{t=1}^T \mathbb{E}_t[\max_\ell \|\overline{\mathbf{p}}_\ell\|^2_{V_t(\hat{\theta}_t^\lambda)^{-1}}]}$$

Next, note that

$$V_t(\hat{\theta}_t^\lambda) = \sum_{s=1}^{t-1} \dot{\mu}_i(\hat{\theta}_t^\lambda, \overline{\mathbf{p}}_s)\overline{\mathbf{p}}_s\overline{\mathbf{p}}_s^\top + \lambda \mathbf{I} \succeq \kappa^{-1} \sum_{s=1}^{t-1} \overline{\mathbf{p}}_s\overline{\mathbf{p}}_s^\top + \lambda \mathbf{I} =: \kappa^{-1} W_t.$$

for $\kappa = 1/\min_{\overline{\mathbf{p}}} \dot{\mu}_i(\theta, \overline{\mathbf{p}})$. Hence, $V_t(\hat{\theta}_t^\lambda)^{-1} \preceq \kappa W_t^{-1}$ for $W_t := \sum_{s=1}^{t-1} \overline{\mathbf{p}}_s\overline{\mathbf{p}}_s^\top + \lambda \mathbf{I}$ which implies

$$u\psi_T(\delta)\sqrt{T\sum_{t=1}^{T} \mathbb{E}_t[\max_\ell \|\overline{\mathbf{p}}_\ell\|^2_{V_t(\hat{\theta}_t^\lambda)^{-1}}]} \leq u\psi_T(\delta)\sqrt{\kappa T\sum_{t=1}^{T} \mathbb{E}_t[\max_\ell \|\overline{\mathbf{p}}_\ell\|^2_{W_t^{-1}}]}.$$

Next, note that $\overline{\mathbf{p}}_\ell$ can be written as $P_\ell \overline{\mathbf{p}}$ where $P_\ell : \mathbb{R}^3 \to \mathbb{R}^{3K}$ projects onto the ℓ^{th} set of 3 coordinates. Hence the above is bounded by

$$u\psi_T(\delta)\sqrt{\kappa T\sum_{t=1}^{T} \mathbb{E}_t[\max_\ell \|P_\ell\overline{\mathbf{p}}\|^2_{W_t^{-1}}]} = u\psi_T(\delta)\sqrt{\kappa T\sum_{t=1}^{T} \mathbb{E}_t[\|\overline{\mathbf{p}}\|^2_{W_t^{-1}}]} = u\psi_T(\delta)\sqrt{\kappa T\mathbb{E}\left[\sum_{t=1}^{T} \|\overline{\mathbf{p}}\|^2_{W_t^{-1}}|\mathcal{F}_{t-1}\right]}$$

Finally, we rely on the standard elliptical potential lemma (cf., Lemma 19.4 of Lattimore and Szepesvári (2020)) to show that

$$\mathbb{E}\left[\sum_{t=1}^{T} \|\overline{\mathbf{p}}\|^2_{W_t^{-1}}|\mathcal{F}_{t-1}\right] \leq 6K \log\left(\frac{3\lambda + T(u^2+B^2+1)}{3\lambda^{1/3K}}\right)$$

where we recall that we have assumed $\max_{x\in\mathcal{X}} \|x\|_\infty \leq B$. Plugging this in alongside the definition of $\psi_T(\delta)$ completes the proof.

Lemma 3. *Let $p_\ell \leq U$ for all $\ell \in [K]$. For $\theta = [\alpha_1, \beta_1, \gamma_1, \ldots, \alpha_K, \beta_K, \gamma_K] \in R^{3K}$ and $\theta' = [\alpha_1', \beta_1', \gamma_1' \ldots, \alpha_K', \beta_K', \gamma_K'] \in R^{3K}$, price vector $\mathbf{p}$, and marketing mix $\mathbf{x}$*

$$R(\theta', \mathbf{p}, \mathbf{x}) - R(\theta, \mathbf{p}, \mathbf{x}) \leq U \max_\ell |(1, p_\ell, x_\ell)^\top(\alpha_\ell' - \alpha_\ell, \beta_\ell' - \beta_\ell, \gamma_\ell - \gamma_\ell')|$$

In particular, if $\alpha_\ell - \beta_\ell p_\ell - \gamma_\ell x_\ell \geq \alpha_\ell' - \beta_\ell' - \gamma_\ell' x_\ell p_\ell$ then

$$R(\theta', \mathbf{p}) - R(\theta, \mathbf{p}) \leq U \max_\ell (1, p_\ell)^\top(\alpha_\ell' - \alpha_\ell, \beta_\ell' - \beta_\ell, \gamma_\ell - \gamma_\ell')$$

Proof of Lemma 3. Let $\theta = [\alpha_1, \beta_1, \gamma_1 \cdots, \alpha_K, \beta_K, \gamma_K] \in \mathbb{R}^{3K}$, $\mathbf{p} = (p_1, \cdots, p_K) \in \mathbb{R}^K$, and $x = (x_1, \cdots, x_K) \in \mathbb{R}^K$. Define $\theta_\ell = (\alpha_\ell, \beta_\ell, \gamma_\ell)^\top$ and $\mathbf{p}_\ell = (1, p_\ell, x_\ell)^\top$, and functions

$$q_\ell(\theta, p, x) := \frac{e^{\theta_\ell^\top \mathbf{p}_\ell}}{1 + \sum_{i=1}^K e^{\theta_i^\top \mathbf{p}_i}}$$

$$R(\theta, p, x) := \frac{\sum_{i=1}^K p_i e^{\theta_i^\top \mathbf{p}_i}}{1 + \sum_{i=1}^K e^{\theta_i^\top \mathbf{p}_i}}$$

By the mean value theorem, for some $\bar{\theta}$ on the line between θ', θ

$$R(\theta', \mathbf{p}, \mathbf{x}) - R(\theta, \mathbf{p}, \mathbf{x}) = \nabla_\theta R(\tilde{\theta}, \mathbf{p}, \mathbf{x})^\top(\theta' - \theta)$$

Let's now compute the righthand side. Note that

$$\nabla_{\theta_\ell} R(\theta, \mathbf{p}, \mathbf{x}) = \frac{p_\ell e^{\theta_\ell^\top \mathbf{p}_\ell}(1 + \sum_{i=1}^K e^{\theta_i^\top \mathbf{p}_i}) - (\sum_{i=1}^K p_i e^{\theta_i^\top \mathbf{p}_i}) e^{\theta_\ell^\top \bar{p}_\ell}}{(1 + \sum_{i=1}^K e^{\theta_i^\top \mathbf{p}_i})^2} \mathbf{p}_\ell$$

So then

$$\begin{aligned}
(\nabla_\theta R(\tilde{\theta}, \mathbf{p}, \mathbf{x}))^\top (\theta' - \theta) &= \frac{\sum_{\ell=1}^K [p_\ell e^{\tilde{\theta}_\ell^\top \mathbf{p}_\ell}(1 + \sum_{i=1}^K e^{\tilde{\theta}_i^\top \mathbf{p}_i}) - (\sum_{i=1}^K p_i e^{\tilde{\theta}_i^\top \mathbf{p}_i}) e^{\tilde{\theta}_\ell^\top \mathbf{p}_\ell}] \mathbf{p}_\ell^\top (\theta'_\ell - \theta_\ell)}{(1 + \sum_{i=1}^K e^{\tilde{\theta}_i^\top \mathbf{p}_i})^2} \\
&= \sum_{l=1}^K p_\ell q_\ell(\tilde{\theta}, p, x) \mathbf{p}_\ell^\top (\theta'_\ell - \theta_\ell) - R(\tilde{\theta}, \mathbf{p}, \mathbf{x}) \sum_{\ell=1}^K q_\ell(\tilde{\theta}, p, x) \mathbf{p}_\ell^\top (\theta'_\ell - \theta_\ell) \\
&= \sum_{l=1}^K (p_\ell - R(\tilde{\theta}, \mathbf{p}, \mathbf{x})) q_\ell(\tilde{\theta}, p, x) \mathbf{p}_\ell^\top (\theta'_\ell - \theta_\ell)
\end{aligned}$$

In general, we can bound

$$\begin{aligned}
&\leq U \sum_{\ell=1}^K q_\ell(\tilde{\theta}, p) |\mathbf{p}_\ell^\top (\theta'_\ell - \theta_\ell)| \\
&\leq U \max_\ell |p_\ell^\top (\theta'_\ell - \theta_\ell)|
\end{aligned}$$

where the final inequality holds since q_ℓ defines a probability distribution. If $\mathbf{p}_\ell^\top (\theta'_\ell - \theta_\ell) \geq 0$ (component-wise) the previous inequality is still true. □

D Proof of Theorem 3

Theorem 5. *Assume that* $\max_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x}\|_\infty \leq B$*, the largest price is bounded by* u*,* $\|\theta\| \leq S$ *with probability 1. With probability at least* $1 - \delta$*, the regret of Algorithm 1 in the contextual setting is bounded by*

$$Reg_T^B \lesssim O\left(SudK \log(dKT/\delta) \sqrt{\kappa T \log\left(3\kappa + T(u^2 + B^2 + 1)\right)}\right).$$

Proof. Let μ_i denote the multinomial probability of item i

$$\mu_i(\theta, \overline{\mathbf{p}}_t) = \mathbb{P}_\theta(I = i | \mathbf{p}, \mathbf{x}, \mathbf{c}_t) = \frac{\exp(\langle \alpha_i, \mathbf{c}_t \rangle - \langle \beta_i, \mathbf{c}_t \rangle p_i + \langle \gamma_i, \mathbf{c}_t \rangle x_i)}{1 + \sum_{k=1}^K \exp(\langle \alpha_k, \mathbf{c}_t \rangle - \langle \beta_k, \mathbf{c}_t \rangle p_k + \langle \gamma_k, \mathbf{c}_t \rangle x_k)}. \tag{A5}$$

We begin by formally defining the λ-regularized maximum likelihood estimator at time t, $\hat{\theta}_t^\lambda$. Given a dataset $\{\mathbf{p}_s, \mathbf{x}_s, I_s\}_{s=1}^t$ where $\mathbf{p}_s$ denotes the price vector played at time s, $\mathbf{x}_s$ the marketing mix, and $I_s \in \{0, 1, \ldots, K\}$ is the item selected,

$$\hat{\theta}_t^\lambda := \arg \min_{\theta \in \mathbb{R}^{3dK}} \sum_{s=0}^t \sum_{i=0}^K \mathbf{1}\{I_s = i\} \log\left(\mu_i\left(\theta, \overline{\mathbf{p}}_s\right)\right) - \frac{\lambda}{2} \|\theta\|_2^2 \tag{A6}$$

where $\overline{\mathbf{p}}_{t,i} = [\mathbf{c}_t, p_i \mathbf{c}_t, \mathbf{x}_t] \in \mathbb{R}^{3dK}$ and $\theta_* := [\alpha, \beta, \gamma] \in \mathbb{R}^{3dK}$. We define the confidence set as

$$C_t(\delta) := \left\{\theta \in \Theta : \|\theta - \hat{\theta}_t^\lambda\|_{V_t(\hat{\theta}_t^\lambda)} \leq \overline{\psi}_t(\delta)\right\}$$

where

$$V_t(\theta) := \sum_{s=1}^{t-1} \dot{\mu}_i(\theta, \overline{\mathbf{p}}_s)\overline{\mathbf{p}}_s\overline{\mathbf{p}}_s^\top + \lambda \mathbf{I}$$

for $\dot{\mu}_i := \nabla_\theta \mu_i(\theta, \overline{\mathbf{p}})$ and

$$\overline{\psi}_t(\delta) := 2(1+2S)\left(\frac{\sqrt{\lambda}}{2} + \frac{2}{\sqrt{\lambda}}\log\left(\frac{T(\lambda + t/3)^{1.5dK}\lambda^{1.5dK}}{\delta}\right) + \frac{6dK}{\sqrt{\lambda}}\log(2)\right).$$

Also define the UCB given some p and x as

$$U_t(p, x) = \max_{\theta \in C_t(\delta)} \sum_{i=1}^{K} \frac{p_i \exp(\theta^\top \overline{\mathbf{p}}_i)}{1 + \sum_{j=1}^{K} \exp(\theta^\top \overline{\mathbf{p}}_j)}.$$

We let Reg_T denote the cumulated regret at time T. Then

$$\begin{aligned}
&\mathbb{E}[\mathrm{Reg}_t - \mathrm{Reg}_{t-1} | \mathcal{F}_{t-1}] \\
&= \mathbb{E}[R_{\theta_*}(\mathbf{p}_*, \mathbf{x}_*) - R_{\theta_*}(\mathbf{p}_t, \mathbf{x}_t) | \mathcal{F}_{t-1}] \\
&= \mathbb{E}[R_{\theta_*}(\mathbf{p}_*, \mathbf{x}_*) - U_t(\mathbf{p}_*, \mathbf{x}_*) | \mathcal{F}_{t-1}] + \mathbb{E}[U_t(\mathbf{p}_*, \mathbf{x}_*) - U_t(\mathbf{p}_t, \mathbf{x}_t) | \mathcal{F}_{t-1}] \\
&\quad + \mathbb{E}[U_t(\mathbf{p}_t, \mathbf{x}_t) - R_{\theta_*}(\mathbf{p}_t, \mathbf{x}_t) | \mathcal{F}_{t-1}] \\
&= \mathbb{E}[R_{\theta_*}(\mathbf{p}_*, \mathbf{x}_*) - U_t(\mathbf{p}_*, \mathbf{x}_*) | \mathcal{F}_{t-1}] + \mathbb{E}[U_t(\mathbf{p}_t, \mathbf{x}_t) - R_{\theta_*}(\mathbf{p}_t, \mathbf{x}_t) | \mathcal{F}_{t-1}]
\end{aligned}$$

where the second term has expectation zero since $(\mathbf{p}_t, \mathbf{x}_t)$ follows the same distribution as $(\mathbf{p}_*, \mathbf{x}_*)$ given $\mathcal{F}_{t-1}$. Therefore, the cumulated regret

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}_T] &= \sum_{t=1}^{T} \mathbb{E}[\mathrm{Reg}_t - \mathrm{Reg}_{t-1} | \mathcal{F}_{t-1}] \\
&= \sum_{t=1}^{T} \mathbb{E}[R_{\theta_*}(\mathbf{p}_*, \mathbf{x}_*) - U_t(\mathbf{p}_*, \mathbf{x}_*) | \mathcal{F}_{t-1}] + \mathbb{E}[U_t(\mathbf{p}_t, \mathbf{x}_t) - R_{\theta_*}(\mathbf{p}_t, \mathbf{x}_t) | \mathcal{F}_{t-1}] \\
&= \sum_{t=1}^{T} \mathbb{E}[R_{\theta_*}(\mathbf{p}_*, \mathbf{x}_*) - U_t(\mathbf{p}_*, \mathbf{x}_*) | \mathcal{F}_{t-1}] + \sum_{t=1}^{T} \mathbb{E}[U_t(\mathbf{p}_t, \mathbf{x}_t) - R_{\theta_*}(\mathbf{p}_t, \mathbf{x}_t) | \mathcal{F}_{t-1}].
\end{aligned}$$

We will bound the two terms together. Define the good event $\mathcal{E}$ as

$$\mathcal{E} := \{\forall t \geq 1, \theta_* \in C_t(\delta)\}.$$

First, by Lemma 11 from Agrawal et al. (2020), we have $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$. Also, for any p, x, we have

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[U_t(\mathbf{p}, \mathbf{x}) - R_{\theta_*}(\mathbf{p}, \mathbf{x}) | \mathcal{F}_{t-1}] &= \sum_{t=1}^{T} \mathbb{E}[\mathbf{1}(\mathcal{E})(U_t(\mathbf{p}, \mathbf{x}) - R_{\theta_*}(\mathbf{p}, \mathbf{x})) | \mathcal{F}_{t-1}] \\
&\quad + \sum_{t=1}^{T} \mathbb{E}[\mathbf{1}(\mathcal{E}^c)(U_t(\mathbf{p}, \mathbf{x}) - R_{\theta_*}(\mathbf{p}, \mathbf{x})) | \mathcal{F}_{t-1}] \\
&\lesssim u\delta + \sum_{t=1}^{T} \mathbb{E}[\mathbf{1}(\mathcal{E})(U_t(\mathbf{p}, \mathbf{x}) - R_{\theta_*}(\mathbf{p}, \mathbf{x})) | \mathcal{F}_{t-1}].
\end{aligned}$$

Consider some time t and some $\theta \in C_t(\delta)$. First,

$$\begin{aligned}\|\theta - \theta_*\|_{V_t(\hat{\theta}_t^\lambda)} &\le \|\theta - \hat{\theta}_t^\lambda\|_{V_t(\hat{\theta}_t^\lambda)} + \|\hat{\theta}_t^\lambda - \theta_*\|_{V_t(\hat{\theta}_t^\lambda)} \\ &\le 2\overline{\psi}_t(\delta)\end{aligned}$$

since both θ and θ_* are in the confidence set. Then, by the Lipschitz property from Lemma 3, we have that under $\mathcal{E}$,

$$\begin{aligned}R_\theta(\mathbf{p}, \mathbf{x}) - R_{\theta_*}(\mathbf{p}, \mathbf{x}) &\le u \max_\ell |\overline{\mathbf{p}}_\ell^\top (\theta - \theta_*)|] \\ &\le u \max_\ell \|\overline{\mathbf{p}}_\ell\|_{V_t(\hat{\theta}_t^\lambda)^{-1}} \|\theta - \theta_*\|_{V_t(\hat{\theta}_t^\lambda)} \\ &\le 2u\psi_t(\delta) \max_\ell \|\overline{\mathbf{p}}_\ell\|_{V_t(\hat{\theta}_t^\lambda)^{-1}}.\end{aligned}$$

Note that the right-hand side is a quantity independent of θ, by taking maximum over $\theta \in C_t(\delta)$, we have that

$$U_t(\mathbf{p}, \mathbf{x}) - R_{\theta_*}(\mathbf{p}, \mathbf{x}) \le 2u\psi_t(\delta) \max_\ell \|\overline{\mathbf{p}}_\ell\|_{V_t(\hat{\theta}_t^\lambda)^{-1}}.$$

Let $\mathbb{E}[\cdot|\mathcal{F}_{t-1}] =: \mathbb{E}_t[\cdot]$ for brevity. Therefore,

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}_t[\mathbf{1}(\mathcal{E})(U_t(\mathbf{p}, \mathbf{x}) - R_{\theta_*}(\mathbf{p}, \mathbf{x}))] &\le 2u \sum_{t=1}^T \overline{\psi}_t(\delta) \mathbb{E}_t[\max_\ell \|\overline{\mathbf{p}}_\ell\|_{V_t(\hat{\theta}_t^\lambda)^{-1}}] \\ &\le 2u\overline{\psi}_T(\delta) \sum_{t=1}^T \mathbb{E}_t[\max_\ell \|\overline{\mathbf{p}}_\ell\|_{V_t(\hat{\theta}_t^\lambda)^{-1}}].\end{aligned}$$

where the final inequality holds since $\overline{\psi}_t(\delta)$ is an increasing sequence in t. By a second application of Cauchy-Schwarz,

$$u\psi_T(\delta) \sum_{t=1}^T \mathbb{E}_t[\max_\ell \|\overline{\mathbf{p}}_\ell\|_{V_t(\hat{\theta}_t^\lambda)^{-1}}] \le u\psi_T(\delta) \sqrt{T \sum_{t=1}^T \mathbb{E}_t[\max_\ell \|\overline{\mathbf{p}}_\ell\|^2_{V_t(\hat{\theta}_t^\lambda)^{-1}}]}$$

Next, note that

$$V_t(\hat{\theta}_t^\lambda) = \sum_{s=1}^{t-1} \dot{\mu}_i(\hat{\theta}_t^\lambda, \overline{\mathbf{p}}_s) \overline{\mathbf{p}}_s \overline{\mathbf{p}}_s^\top + \lambda \mathbf{I} \succeq \kappa^{-1} \left(\sum_{s=1}^{t-1} \overline{\mathbf{p}}_s \overline{\mathbf{p}}_s^\top + \kappa\lambda \mathbf{I} \right) =: \kappa^{-1} W_t.$$

for $\kappa = 1/\min_{\overline{\mathbf{p}}} \dot{\mu}_i(\theta, \overline{\mathbf{p}})$. Hence, $V_t(\hat{\theta}_t^\lambda)^{-1} \preceq \kappa W_t^{-1}$ for $W_t := \sum_{s=1}^{t-1} \overline{\mathbf{p}}_s \overline{\mathbf{p}}_s^\top + \kappa\lambda \mathbf{I}$ which implies

$$u\overline{\psi}_T(\delta) \sqrt{T \sum_{t=1}^T \mathbb{E}_t[\max_\ell \|\overline{\mathbf{p}}_\ell\|^2_{V_t(\hat{\theta}_t^\lambda)^{-1}}]} \le u\overline{\psi}_T(\delta) \sqrt{\kappa T \sum_{t=1}^T \mathbb{E}_t[\max_\ell \|\overline{\mathbf{p}}_\ell\|^2_{W_t^{-1}}]}.$$

Next, note that $\overline{\mathbf{p}}_\ell$ can be written as $P_\ell \overline{\mathbf{p}}$ where $P_\ell : \mathbb{R}^3 \to \mathbb{R}^{3K}$ projects onto the ℓ^{th} set of 3 coordinates. Hence the above is bounded by

$$u\overline{\psi}_T(\delta) \sqrt{\kappa T \sum_{t=1}^T \mathbb{E}_t[\max_\ell \|\overline{\mathbf{p}}\|^2_{W_t^{-1}}]} = u\overline{\psi}_T(\delta) \sqrt{\kappa T \sum_{t=1}^T \mathbb{E}_t[\|\overline{\mathbf{p}}\|^2_{W_t^{-1}}]} = u\overline{\psi}_T(\delta) \sqrt{\kappa T \mathbb{E}\left[\sum_{t=1}^T \|\overline{\mathbf{p}}\|^2_{W_t^{-1}} | \mathcal{F}_{t-1} \right]}$$

Finally, we rely on the standard elliptical potential lemma (cf., Lemma 19.4 of Lattimore and Szepesvári (2020)) to show that

$$\mathbb{E}\left[\sum_{t=1}^{T} \|\overline{\mathbf{p}}\|^2_{W_t^{-1}} | \mathcal{F}_{t-1}\right] \leq 6dK \log\left(\frac{3\kappa\lambda + T(u^2 + B^2 + 1)}{3\kappa^{1/3dK}\lambda^{1/3dK}}\right)$$

where we recall that we have assumed $\max_{x\in\mathcal{X}} \|x\|_\infty \leq B$. Plugging this in alongside the definition of $\overline{\psi}_T(\delta)$ completes the proof. □

E Baseline Algorithms' Descriptions

Algorithm 6 Greedy for Multinomial Demand Model

Input: $\mathcal{P} \times \mathcal{X}$, τ - Number of initial random samples
  for $t = 0, 1, 2, \cdots$, **do**
    if $t < \tau$ **then**
      Sample $\mathbf{p}_t, \mathbf{x}_t$ uniformly from $\mathcal{P} \times \mathcal{X}$
    else
      Set $\theta_t = \arg\max_\theta \mathcal{L}_t(\theta|\mathcal{H}_{t-1})$
      Set $\mathbf{p}_t, \mathbf{x}_t = \arg\max_{\mathbf{p}\in\mathcal{P},\mathbf{x}\in\mathcal{X}} R_{\theta_t}(\mathbf{p}, \mathbf{x})$
    end if
    Observe I_t and $r_t := p_{I_t} - m_{I_t}$
  end for

Algorithm 7 M3P for Multinomial Demand Model

Input: $\mathcal{P} \times \mathcal{X}$
  Set $K = dim(\mathcal{P})$
  Set $t_{next} = 0$
  for $b = 1, 2, \cdots$, **do**
    Set $l_b = K + b$ as the number of periods in the b-th block
    for $t = t_{next}, t_{next} + 1, \cdots, t_{next} + l_b - 1$ **do**
      Select $\mathbf{p}_t$ as follows
        Exploration: For the first K periods ($t - t_{next} < K$) sample $\mathbf{p}_t, \mathbf{x}_t$ uniformly from $\mathcal{P} \times \mathcal{X}$
        Learning: After these K exploration samples find $\theta_k = \arg\max_\theta \mathcal{L}_t(\theta|\mathcal{H}_{t-1})$
        Exploitation: For the remaining $b - K$ periods set $\mathbf{p}_t, \mathbf{x}_t = \arg\max_{\mathbf{p}\in\mathcal{P},\mathbf{x}\in\mathcal{X}} R_{\theta_k}(\mathbf{p}, \mathbf{x})$
      In each period t, observe I_t and $r_t := p_{I_t} - m_{I_t}$
    end for
    Set $t_{next} = t_{next} + l_b$
  end for

F Equivalence of Linear Models

A common model considered for linear contextual pricing in the literature Ban and Keskin (2021); Javanmard et al. (2020) is as follows. At each time we assume that we observe a context $\mathbf{c}_{ti} \in \mathbb{R}^d$ for product i. The utility of product at time t is given by

$$U_{it}(\mathbf{p}_t, \mathbf{c}_{ti}) = \langle \mathbf{c}_{ti}, \alpha \rangle - \langle \mathbf{c}_{it}, \beta \rangle p_{ti} \tag{A7}$$

for some fixed parameter vectors $\alpha, \beta \in \mathbb{R}^d$. We claim that this model captures our setting above:

- In the case where we have no context (Section 4.2), we can set $\mathbf{c}_{ti} = \mathbf{e}_i \in \mathbb{R}^K$, and α, β above. That is, the context that arrives is constant in each round.
- In our setting of the contextual case (Section 6.1), we set $\mathbf{c}_{ti} = \mathbf{c}_t \otimes \mathbf{e}_i \in \mathbb{R}^{dK}$, and $\alpha = [\alpha_1, \cdots, \alpha_k] \in \mathbb{R}^{dK}$ and similarly for β.

As M3P (Javanmard et al., 2020) uses the utility model stated in Equation (A7), in the case of contextual pricing it should have exploration phases equal to the dimension of $\mathbf{c}_{ti}$ which is $d \times K$. Algorithm 8 shows our adaptation of the M3P algorithm for the contextual setting with promotions. We point out again, that M3P has no theoretical guarantee when promotions are included.

Algorithm 8 M3P for Linear Contextual Pricing with Multinomial Demand Model

Input: $\mathcal{P} \times \mathcal{X}$, dimension of the context vectors d, regularization factor λ

Set $K = \dim(\mathcal{P})$
Set $t_{next} = 0$
for $b = 1, 2, \cdots,$ **do**
  Set $l_b = dK + b$ as the number of periods in the bth block
  for $t = t_{next}, t_{next} + 1, \cdots, t_{next} + l_b - 1$ **do**
    Select p_t as follows
      Exploration: For the first dK periods ($t - t_{next} < dK$) sample $\mathbf{p}_t, \mathbf{x}_t$ uniformly from $\mathcal{P} \times \mathcal{X}$
      Learning: After dK exploration samples, learn the parameters:
        Define $\mathcal{L}_t(\{(\mathbf{p}_s, \mathbf{x}_s, \mathbf{c}_s, I_s)\}_{s=1}^t, \theta) = \sum_{s=1}^t \log(\mathbb{P}_\theta(I_s|\mathbf{p}_s, \mathbf{x}_s, \mathbf{c}_s)) + \frac{\lambda}{2}\|\theta\|_2^2$
        Find $\theta_b = \arg\max_\theta \mathcal{L}_t(\{(\mathbf{p}_s, \mathbf{x}_s, \mathbf{c}_s, I_s)\}_{s=1}^t, \theta)$
      Exploitation:
        For the remaining b periods observe context $\mathbf{c}_t$ and set

$$\mathbf{p}_t, \mathbf{x}_t = \arg \max_{\mathbf{p} \in \mathcal{P}, \mathbf{x} \in \mathcal{X}} R_{\theta_b}(\mathbf{p}, \mathbf{x}, \mathbf{c_t}).$$

    In each period t, observe I_t and $r_t := p_{I_t} - m_{I_t}$
  end for
  Set $t_{next} = t_{next} + l_b$
end for

G Ablation Study: Importance of Promotions

One of the key distinguishing features of our proposed algorithm lies in its ability to simultaneously experiment with both prices and promotions, seeking to discover the optimal price-promotion combination. In this section, we assess the performance of our proposed Thompson Sampling algorithm which learns the optimal promotion by contrasting it with variants of the algorithm in which we fix the promotion. These variants, in particular, involve scenarios where the algorithm does not adaptively select promotions. Instead, it either chooses promotions randomly at each time step or selects a promotion at the beginning and adheres to it throughout the experiment. Notably, when the algorithm begins with the optimal promotion choice, it obviates the need for exploration in the promotion space since it already uses the best strategy.

We employ the same experimental setup as described in Section 5.1. This configuration consists of three products, with the optimal promotion strategy being the allocation of all promotion weight to the first product, as presented in Table 2. To elucidate the significance of learning and adapting to the best promotion, we examine the following variants of Thompson Sampling:

- Our proposed Thompson Sampling, which does not possess prior knowledge of the optimal promotion

vector.

- A version of Thompson Sampling that rigidly employs a fixed promotion strategy at all time steps. We evaluate four different variants:
  - $x = [0, 0, 0]$ - i.e., no promotion
  - $x = [1, 0, 0] = x^*$
  - $x = [0, 1, 0]$
  - $x = [0, 0, 1]$
- A version of Thompson Sampling that randomly selects a promotion strategy from the aforementioned four vectors.

The cumulative regret plot for this setting is displayed in Figure A2. Notably, when the promotion vector remains fixed, the algorithm's inability to adapt to the optimal promotion strategy becomes evident, resulting in linear regret when the wrong promotion is chosen. Similarly, randomly selecting a promotion vector does not yield favorable outcomes, as the learner ends up with suboptimal choices with a high probability. Lastly, we compare the case where the learner fixes the promotion at the optimum promotion from the beginning to the proposed version that actively explores and learns the promotion vector. Interestingly, the Thompson Sampling regret line completely matches the case where the method fixes the optimum promotion. This finding underscores the efficacy of our approach in efficiently acquiring knowledge about the optimal promotion with minimal exploration, ultimately achieving a cumulative regret on par with a scenario in which the optimal promotion is known in advance.

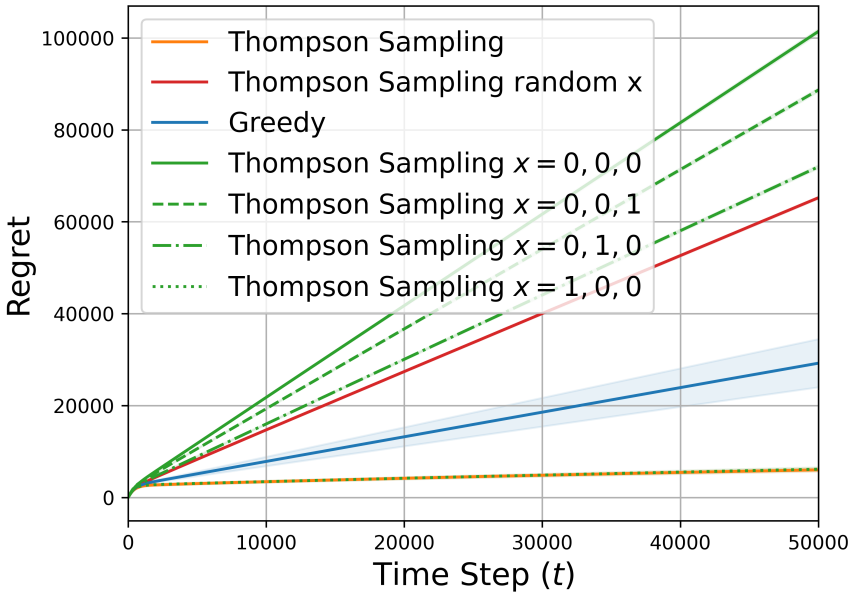


Figure A2: We compare the performance of Thompson Sampling to variants of Thompson Sampling that the promotion part is disabled and the learner either uses a fixed promotion vector all the time or randomly selects one at each step.

H Appendix for Numerical Experiments based on Nielsen Data for Non-Contextual Settings 5.2

H.1 Descriptive Statistics of Coffee Category in NielsenIQ Data

Brand	Mean	Std. Dev.	Min	Max	No. Obs.
			Price per ounce		
Peet's Coffee	0.6423	0.1168	0.4739	0.8998	52.0
Starbucks	0.6432	0.0871	0.5225	0.8231	52.0
Seattle's Best	0.4542	0.0664	0.3298	0.6109	52.0
Stumptown	1.1608	0.0964	0.9602	1.2951	52.0
Tony's Coffee	0.8126	0.0389	0.6685	0.8325	52.0
Folgers	0.3311	0.0396	0.2568	0.4099	52.0
Cutters Point	0.8000	0.0277	0.7446	0.8325	52.0
CTL BR	0.3857	0.0678	0.2626	0.6424	52.0
Gevalia Kaffe	0.5942	0.0304	0.5183	0.6991	52.0
Other	0.7494	0.0544	0.5787	0.8509	52.0
			Feature per ounce		
Peet's Coffee	0.3905	0.4325	0.0	1.0000	52.0
Starbucks	0.3833	0.3919	0.0	1.0000	52.0
Seattle's Best	0.2079	0.3652	0.0	1.0000	52.0
Stumptown	0.0577	0.2354	0.0	1.0000	52.0
Tony's Coffee	0.0000	0.0000	0.0	0.0000	52.0
Folgers	0.0408	0.1679	0.0	0.8324	52.0
Cutters Point	0.0000	0.0000	0.0	0.0000	52.0
CTL BR	0.0136	0.0774	0.0	0.5451	52.0
Gevalia Kaffe	0.0596	0.2119	0.0	0.9126	52.0
Other	0.0225	0.0774	0.0	0.4982	52.0
			Display per ounce		
Peet's Coffee	0.2377	0.2070	0.0	0.6990	52.0
Starbucks	0.2114	0.1662	0.0	0.5784	52.0
Seattle's Best	0.0546	0.1366	0.0	0.5562	52.0
Stumptown	0.0846	0.2449	0.0	1.0000	52.0
Tony's Coffee	0.0301	0.1230	0.0	0.5714	52.0
Folgers	0.0000	0.0000	0.0	0.0000	52.0
Cutters Point	0.1184	0.2346	0.0	0.6667	52.0
CTL BR	0.0784	0.1261	0.0	0.5451	52.0
Gevalia Kaffe	0.0000	0.0000	0.0	0.0000	52.0
Other	0.0346	0.0775	0.0	0.4582	52.0
			Ounces sold		
Peet's Coffee	3060.4519	1324.0975	923.00	5883.50	52.0
Starbucks	2892.1919	1065.3690	1245.00	5978.00	52.0
Seattle's Best	2032.7692	878.0696	740.00	4056.00	52.0
Stumptown	544.3846	253.8145	228.00	1464.00	52.0
Tony's Coffee	443.7692	114.9156	264.00	792.00	52.0
Folgers	763.3038	220.1129	357.00	1431.40	52.0
Cutters Point	286.8462	83.6888	132.00	456.00	52.0
CTL BR	557.2779	150.3929	322.90	846.60	52.0
Gevalia Kaffe	318.8846	111.8244	82.00	628.00	52.0
Other	1969.2338	452.0854	1328.42	3954.21	52.0

Table A2: Summary statistics of the weekly data used in Section 5.2 for the nine top brands and other brands aggregated to *Other*. We show the statistics on price, feature, and display per ounce, and total ounces sold.

H.2 Parameter Estimates for Single Store in King County

Table A3 shows the α, β, γ parameters estimated for coffee demand in one of the stores in King County. The estimation procedure is discussed in Section 5.2.1.

	coef	std err	t	P> |t|	[0.025	0.975]
$\alpha_{Peet'sCoffee}$	1.4967	0.173	8.659	0.000	1.158	1.835
$\alpha_{Starbucks}$	1.2047	0.257	4.685	0.000	0.701	1.709
$\alpha_{Seattle'sBest}$	1.4439	0.195	7.417	0.000	1.062	1.825
$\alpha_{Stumptown}$	1.6104	0.171	9.423	0.000	1.275	1.945
$\alpha_{Tony'sCoffee}$	-0.9526	0.262	-3.641	0.000	-1.465	-0.440
$\alpha_{Folgers}$	-0.5087	0.111	-4.602	0.000	-0.725	-0.292
$\alpha_{CuttersPoint}$	2.1615	0.375	5.768	0.000	1.427	2.896
α_{CTLBR}	-1.9331	0.093	-20.705	0.000	-2.116	-1.750
$\alpha_{GevaliaKaffe}$	0.9966	0.273	3.645	0.000	0.461	1.533
α_{Other}	-0.2887	0.203	-1.423	0.155	-0.686	0.109
$\beta_{Peet'sCoffee}$	4.1954	0.210	19.963	0.000	3.783	4.607
$\beta_{Starbucks}$	3.7477	0.330	11.361	0.000	3.101	4.394
$\beta_{Seattle'sBest}$	6.5896	0.395	16.683	0.000	5.815	7.364
$\beta_{Stumptown}$	3.8422	0.146	26.289	0.000	3.556	4.129
$\beta_{Tony'sCoffee}$	2.4883	0.321	7.747	0.000	1.859	3.118
$\beta_{Folgers}$	5.8400	0.250	23.381	0.000	5.350	6.330
$\beta_{CuttersPoint}$	7.0085	0.465	15.071	0.000	6.097	7.920
β_{CTLBR}	2.1558	0.175	12.324	0.000	1.813	2.499
$\beta_{GevaliaKaffe}$	7.3040	0.403	18.114	0.000	6.514	8.094
β_{Other}	1.5894	0.238	6.682	0.000	1.123	2.056
$\gamma_{feature}$	0.0742	0.070	1.060	0.289	-0.063	0.211
$\gamma_{display}$	0.2195	0.055	4.009	0.000	0.112	0.327
Number of Observations: 520						
R-squared: 0.958						
Adjusted R-squared: 0.951						
Log-Likelihood: 100.81						

Table A3: Parameter estimates for ground coffee category for a large store in King County, WA. Week dummies are not shown and the standard errors are clustered at the brand-level.

H.3 Parameters and Optimal Prices and Promotions for the Experiments

Table A4 show the α, β, γ parameters used in the experiments, alongside the optimal price and promotion for each brand. The estimated α, β, and γ values from Table A3 are used here, and to adopt the 32-ounce coffee bag size, the values of β are divided by 32. This is because our estimated β values are for prices per ounce, and with a bag size of 32, the final prices for 32 ounces would be 32 times the price per ounce. Hence, to keep the demand model the same as in the estimation, we should divide the β by 32.

	Peet's Coffee	Starbucks	Seattle's Best	Stumptown	Tony's Coffee	Folgers	Cutters Point	CTL BR	GevaliaKaffe
α	1.497	1.205	1.444	1.610	-0.953	-0.509	2.162	-1.933	0.997
β	0.131	0.117	0.206	0.120	0.078	0.182	0.219	0.067	0.228
γ	0.386	0.237	0.256	0.112	0.324	0.239	0.062	0.307	0.192
$\mathbf{p}_*$	20.071	20.982	17.300	20.772	25.304	17.923	17.010	27.288	16.825
$\mathbf{x}_*$	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table A4: Parameters and Optimal Prices and Promotions used for Numerical Experiments based on Nielsen Data. Note that these parameters are for the 32-ounce coffee bags; hence we have divided the estimated β parameters by 32.

H.4 Simulation Running Time

The simulations were conducted on an Ubuntu machine equipped with a 64-core Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz. Utilizing 48 cores of the CPU, each run of every method was executed on a separate thread. The average running times per time step in seconds for each method are presented in Table A5,

accompanied by the standard deviation across different runs in parentheses. It is noteworthy that all methods exhibit a small computational burden, with each time step requiring a fraction of a second.

Thompson Sampling with Laplace demonstrates comparable running times to Greedy, as both methods employ the same optimization process, with Thompson Sampling incorporating noise in the final step. On the other hand, the performance of Thompson Sampling with Langevin dynamics is contingent on the number of Langevin steps. Each Langevin step involves a gradient descent followed by a noise addition. In the reported measurements, these methods undertake $N_t = 50$ Langevin steps per iteration.

Batch Size	Algorithm	Mean (Std)
	M3P	0.0387 (0.0008)
10	Greedy	0.0839 (0.0036)
	TS-Laplace	0.0841 (0.0035)
	TS-Langevin	0.2350 (0.0252)
200	Greedy	0.0409 (0.0006)
	TS-Laplace	0.0413 (0.0006)
	TS-Langevin	0.2295 (0.0203)

Table A5: Average running times per time step in seconds for each method.

H.5 Implemention Details

- *Batch Size* In practice, it is extremely difficult for firms to do real-time updates of adaptive experimentation algorithms due to computational and engineering costs (Jamieson et al., 2015). Instead, it's common to do updates in fixed time-intervals or in fixed batches. For example, if the batch size is 200, we assume that there are 200 purchase decisions made per period, and all the parameter estimates and the $\{\mathbf{p}_t, \mathbf{x}_t\}$ are updated at the start of each batch and then fixed for the rest of the batch. Depending on the retailer's volume, each batch could translate to a few hours, a day, or a week. [17] We consider two settings, a batch size of 10 (frequent updating) and a batch size of 200.
- *Range of decision variables:* In all our simulations, we allow the price of a two-pound bag of coffee to range from $\ell = \$0.0$ to $u = \$35$.[18] This large range ensures that we are exploring over a sufficiently large range of prices and gives us conservative estimates of regret (since exploration far from the optimal prices is costly). Next, we assume that the display variable is constrained to be in the simplex i.e. $\mathcal{X} = \{x \in \mathbb{R}^K_{\geq 0} : \sum_{i=1}^K x_i = 1\}$, where here, K is the number of brands. Theorem 1 implies that the optimal promotion will be a vertex of the simplex. Therefore, in our simulations, we restrict to $\mathcal{X} = \{\mathbf{e}_i : 1 \leq i \leq K\} \cup \{\mathbf{0}\}$. Again, this corresponds to a setting where we can choose to promote one item or promote none at all.
- *Initial Exploration Phase ($\tau_{explore}$):* For each algorithm, we begin training them with 10 price and promotion vectors uniformly chosen. In practice, if historical data has sufficient variation, this data could be used to initialize models.
- *Parameter Ranges:* As described in Section 3.1, we assume that the true parameters are bounded in absolute value by some constant M. In general, we have not made any assumptions on our prior or

[17] If we were to make a very conservative assumption that there are only 200 consumers in this category in a given day at the retailer, then the timespan of the data is 20000/200 = 100 days. In most large retailers, the number of daily customers is much larger, and this would represent anyone from just one or two days (e.g., Amazon) to a couple of weeks (for smaller retailers with one or two stores). Note that these numbers reflect the anecdote that Amazon changes prices hourly Mehta et al. (2018).

[18] The maximum optimal price is $27.28; see Table A4 in the Web Appendix H.3. Thus, this range covers the optimal prices for all the products in this setting.

posterior distributions when specifying Algorithm 1. However, our theory needs both of these distributions to be supported on Ω. In practice, after enough samples, since the posterior concentrates to a normal distribution with almost all of its mass on a small region around θ, it is not necessary to restrict to distributions that are supported on Ω. If we want to ensure that samples from the posterior lie in Ω in early time steps, we can solve a constrained MLE and use Laplace approximation with rejection sampling. However, we use a simpler approach. Essentially, we sample a set of parameters. If the sampled $\beta_i < 0$ for any product, we set the β_i for this product to a small positive constant (namely .01). Otherwise we compute the optimal price and clip it to the range $[\ell, u]$ for each product.

- *Number of runs:* We perform 40 replications of each algorithm in each simulation and we display the average performance in all our plots.

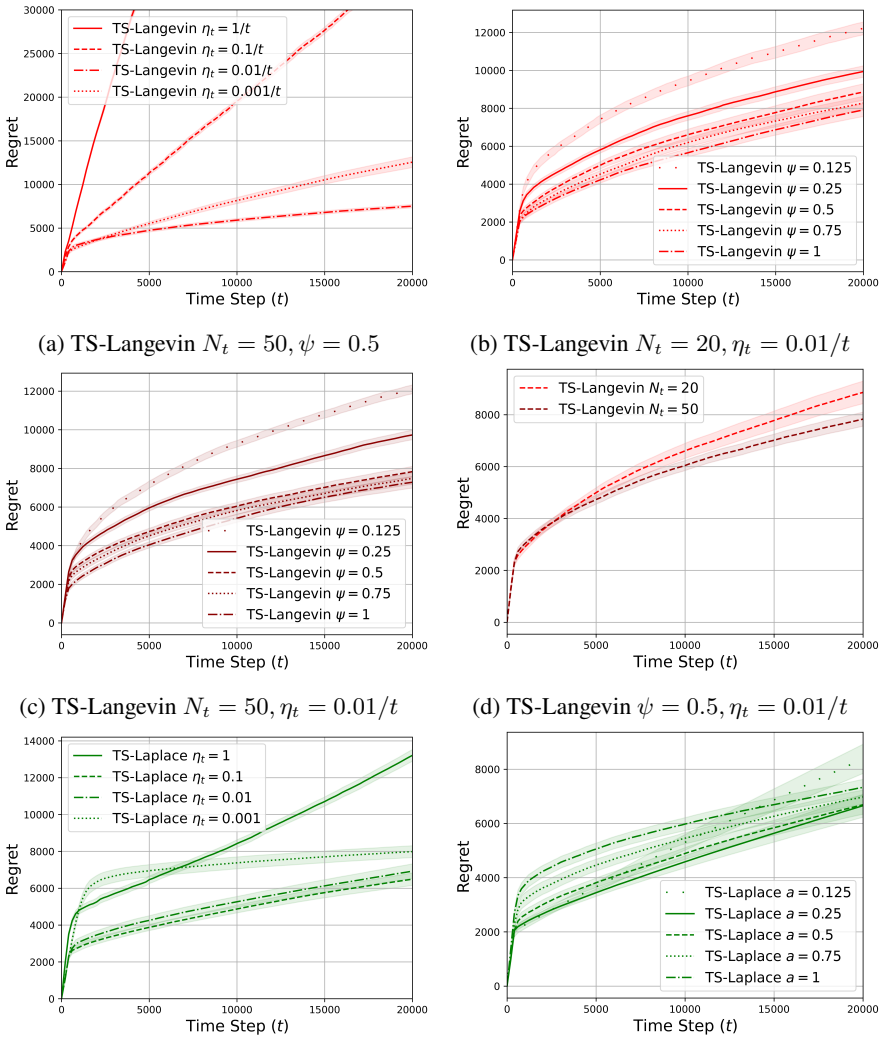
H.6 Parameter Selection and Training for Baseline and Thompson Sampling Algorithms

This section provides an overview of the parameter selection process Thompson Sampling methods in the experiments discussed in Section 5.2.

For the Thompson Sampling methods, both Laplace and Langevin are methods to approximate posterior sampling. The main hyperparameter needed are the learning rate η_t, the number of Langevin dynamics steps N_t and the exploration ψ. For Laplace, the main hyperparameters are the learning rate η_t and exploration parameter a. For the Laplace version, we employ constant learning rate functions $\eta_t = c$ with 500 gradient steps, whereas for Langevin, we utilize reciprocally decaying learning rates derived from the theory $\eta_t = c/t$ Welling and Teh (2011). In both cases $c \in \{1, 0.1, 0.01, 0.001\}$. Subsequent to the learning rate selection, we undertake a grid search, exploring the exploration rate (a for Laplace and ψ for Langevin) and the Number of Langevin Steps N_t for Langevin dynamics. Specifically, we experiment with exploration rates ψ from the set $\{0.125, 0.25, 0.5, 0.75, 1\}$ and Langevin steps $N_t \in \{20, 50\}$. The results are presented in Figure A3.

The learning rate has a significant impact on the regret, with $\eta_t = .01/t$ being the best performing. Our observations reveal that the method exhibits insensitivity to the choice of the exploration rate ψ, provided it falls within the range close to 1 (i.e., [0.5, 1]) (Figure A4.b, A4.c). For Langevin dynamics, smaller exploration rates (e.g., 0.125, 0.25) lead to insufficient exploration, resulting in suboptimal regrets. Unsurprisingly, regarding Langevin's steps, we note that a higher number of steps, such as $N_t = 50$, yields superior regret outcomes (A4.d). This improvement could be attributed to the fact that with an increased number of Langevin steps, the distribution of the next step parameters converges more closely to the posterior. We remark that in all of these simulations, the $O(\sqrt{T})$ behavior of Thompson Sampling regret can be seen unlike in Greedy which is consistently Linear (eg in Figure 4a)

We remark that in practice, choosing optimal hyperparameters for online algorithms is a difficult task. There are a variety of methods that have been proposed (e.g. Li et al. (2018)). A simple heuristic is to collect some data and then run offline simulations (similar to what we have done in this appendix) to find robust ranges.



(a) TS-Langevin $N_t = 50, \psi = 0.5$

(b) TS-Langevin $N_t = 20, \eta_t = 0.01/t$

(c) TS-Langevin $N_t = 50, \eta_t = 0.01/t$

(d) TS-Langevin $\psi = 0.5, \eta_t = 0.01/t$

(e) TS-Laplace $a = 0.5$

(f) TS-Laplace $\eta_t = 0.1$

Figure A3: Regret plots for grid search over exploration rate and number of Langevin steps for Thompson Sampling method.

I Supplementary Materials for Synthetic Linear Contextual Experiments Section 7.1

I.1 Linear Contextual Model Parameters

In this Appendix we first explain how our parameters $\alpha_i, \beta_i, \gamma_i \in \mathbb{R}^4, 1 \leq i \leq 9$ were chosen. The first column of Table A6 captures the first dimension of each parameter. The second, third, and fourth column are just three randomly chosen permutations of the first column, i.e. as an example $\beta_{12}, \cdots, \beta_{92}$ are just a permutation of $\beta_{11}, \cdots, \beta_{91}$. In particular if at time t the context vector observed is $\mathbf{c}_t = (1, 0, 0, 0)$ then the underlying parameters of the underlying demand model are given by the first column of Table A6 and for $\mathbf{c}_t = (0, 0, 0, 1)$ the set of parameters given by the fourth column. Table A7 shows the optimal price and promotion for each of the products in the case where $\mathbf{c}_t = (1, 0, 0, 0)$. The optimal prices and promotions for each of the other contexts in the ORTHOGONALGROUP case are just permutations of these values. Since each context distribution is a positive vector and in addition, the parameters are positive, this ensures that the demand function is always monotonically decreasing.

Parameter	Feature 1	Feature 2	Feature 3	Feature 4
α_1	1.00	1.00	1.00	1.00
α_2	1.00	1.00	1.00	1.00
α_3	1.00	1.00	1.00	1.00
α_4	1.00	1.00	1.00	1.00
α_5	1.00	1.00	1.00	1.00
α_6	1.00	1.00	1.00	1.00
α_7	1.00	1.00	1.00	1.00
α_8	1.00	1.00	1.00	1.00
α_9	1.00	1.00	1.00	1.00
β_1	0.10	0.25	0.50	0.50
β_2	0.15	0.45	0.15	0.45
β_3	0.20	0.15	0.20	0.40
β_4	0.25	0.10	0.30	0.35
β_5	0.30	0.20	0.25	0.30
β_6	0.35	0.30	0.35	0.25
β_7	0.40	0.35	0.10	0.20
β_8	0.45	0.40	0.45	0.15
β_9	0.50	0.50	0.40	0.10
γ_1	0.80	0.20	0.10	0.10
γ_2	0.30	0.20	0.30	0.20
γ_3	0.50	0.30	0.50	0.50
γ_4	0.20	0.80	0.80	0.30
γ_5	0.80	0.50	0.20	0.80
γ_6	0.30	0.80	0.30	0.20
γ_7	0.50	0.30	0.80	0.50
γ_8	0.20	0.50	0.20	0.30
γ_9	0.10	0.10	0.50	0.80

Table A6: Table of model parameters for linear contextual setting with 9 products and 4 context dimensions.

α	1	1	1	1	1	1	1	1	1
β	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
γ	0.8	0.3	0.5	0.2	0.8	0.3	0.5	0.2	0.1
$\mathbf{p}_*$	\$20.44	\$17.10	\$15.44	\$14.44	\$13.77	\$13.29	\$12.94	\$12.66	\$12.44
$\mathbf{x}_*$	1	0	0	0	0	0	0	0	0

Table A7: Setting of Nine Demand Parameters Used for Linear contextual setting in the case where the context vector is $\mathbf{c}_t = (1, 0, 0, 0)$. The optimal revenue $R(\mathbf{p}_*, \mathbf{x}_*) = \$\mathbf{10.44}$.

I.2 Implementation Details

We also set the following:

- *Batch Size:* We chose the batch size for retraining our models to be 1 in both OrthogonalGroup and WeightedAverages cases. That is we retrain our models after each sample.
- *Range of decision variables:* For simulations in this section we use $\ell = \$0$ to $u = \$30$ for prices and display variables are constrained to be in the simplex, i.e. $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^9_{\geq 0} : \sum_{i=1}^9 \mathbf{x}_i = 1\}$, where $\mathcal{K}$ is the set of brands considered.
- *Initial Exploration Phase ($\tau_{explore}$):* For each algorithm, we begin their training with 10 random context vectors uniformly chosen.
- *Number of runs:* We perform 20 replications of each algorithm.
- *Langevin Parameters:* We set $\eta_t = .03/t$, $\psi_t = \psi = 1$, and $N_t = 100$.

I.3 Box Context Vector Distribution

In this section, we show the experiment results for the Box context distribution. In the Box setting, we use a uniform sample over a box in $\mathbb{R}^4$. We use $[0.5, 1.5]^4/K$ where $K = 9$ is the number of products and the division is to keep the final magnitude of the parameters in the same range as the the previous two cases. Moreover, other implementation details are the same as in previous cases. This setting allows customer heterogeneity along all of the four dimensions.

Figure A4 shows the regret comparison between the non-contextual and contextual versions of the Thompson Sampling method. Aligned with the results for the other two context vector distributions, namely OrthogonalGroups and WeightedAverages, we see that using context vectors is critical to achieving an optimal regret. Furthermore, we compare the contextual version of Thompson Sampling to the contextual version of Greedy and M3P baselines in Figure A5. We could again see that Thompson Sampling outperforms M3P and Greedy in this time horizon. By the time 20,000, the simple regret of Greedy is more than twice the TS simple regret. In this setting, Greedy is even more competitive in the early stages compared to previous settings. We believe this is due that information sharing along context dimensions is higher than OrthogonalGroups and WeightedAverages settings because the context distribution spans $\mathbb{R}^4$.

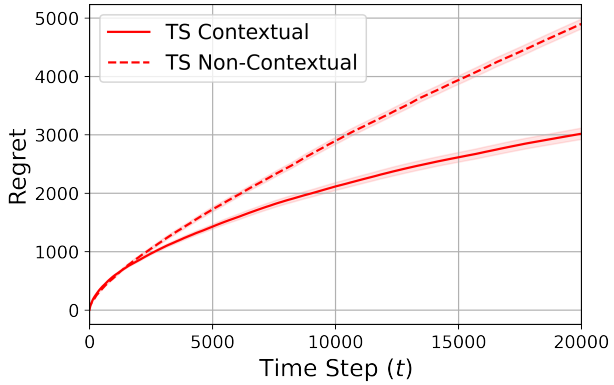
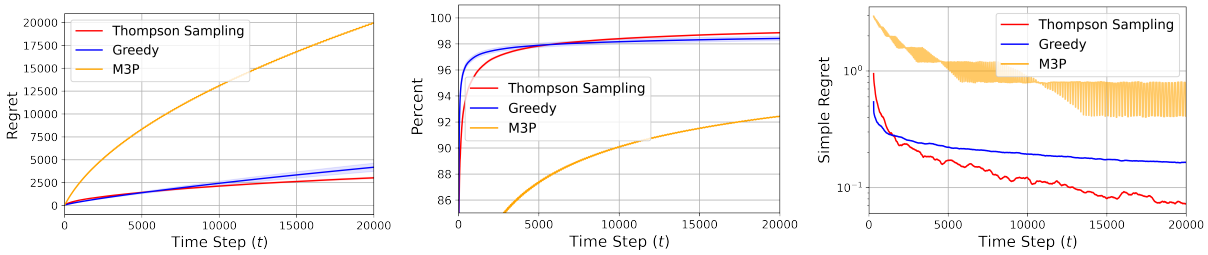


Figure A4: We compare the cumulative regret of the contextual versions of Thompson Sampling (TS Contextual) with its non-contextual counterparts (TS Non-Contextual) for the Box context distributions



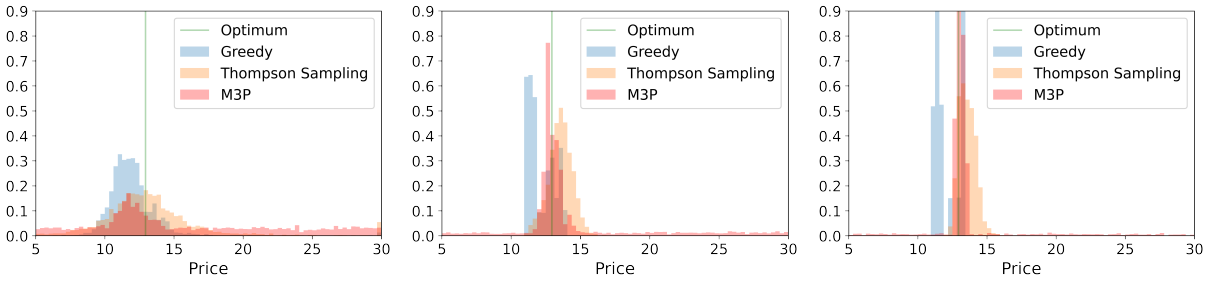
(a) Cumulative Regret (b) Percentage of Optimal Revenue (c) Simple Regret

Figure A5: Comparison of the performances of three contextual methods – Thompson Sampling (TS), M3P, and Greedy for the Box context vector distribution.

I.4 Price Distributions on Synthetic Experiments for the Linear Contextual Setting

Here, we add the price distribution of the M3P method and discuss the properties of the distribution compared to Greedy and Thompson Sampling methods. Figure A6 shows the distribution of prices played that contains Thompson Sampling method and both Greedy and M3P baselines in the OrthogonalGroups context

distribution when the context vector is the first basis vector ($\mathbf{c} = e_1$). While M3P ultimately identifies the optimal price, we can observe that it explores a broad spectrum of prices across all stages, resulting in higher total regret.



(a) Prices $T < 1000$ (b) Prices $5,000 \leq T \leq 6000$ (c) Prices $19,000 \leq T$

Figure A6: A version of Figure 7 that includes the Thompson Sampling method along with both Greedy and M3P baselines.

J Supplementary Materials for NielsenIQ Linear Contextual Experiments in Section 7.2

Table A8 consists of the estimates that we obtained for the model in Section 7.2, along with our modifications to estimates that were zero or negative. We describe our changes now. Recall that the set of brands is $\mathcal{K} = \{$Peet's Coffee, CTLBR, Starbucks, Stumptown, Seattle's Best, Tony's coffee, Caffe Umbria, Ladro, Caffee Vita, Other $\}$.

- Several of the coefficients were zero due to no promotion in those quarters for specific brands (namely Seattle's Best in Q_2 $\gamma_{Seattle'sBest,Q_2}$, Caffe Umbria in Q_1 $\gamma_{CaffeUmbria,Q_1}$, Ladro in Q_1 and Q_4 $\gamma_{Ladro,Q_1}, \gamma_{Ladro,Q_4}$, and Caffe Vita in Q_1 $\gamma_{CaffeVita,Q_1}$). For any estimated coefficients, we replaced its value with the average of the promotion coefficients of all the other brands in the same quarter.
- For the γ_{Ladro,S_1} estimate was zero because the Ladro brand in Store 1 did not have any promotions. So we replaced this estimate with the average of the gamma coefficients of other brands for Store 1 $\gamma_{i',S_1}, i' \in \mathcal{K}, i' \neq Ladro$.
- Two of the brands, Starbucks and Seattle's Best, were not present in Store 2 and hence resulted in zero estimates on all of the coefficients $\alpha_{i,S_2}, \beta_{i,S_2}, \gamma_{i,S_2}$, $i \in \{$*Starbucks, Seattle's Best*$\}$. For each of these coefficients, we replaced its estimated value with the average of the coefficients of the other brands. This modification led to unrealistic market shares for Starbucks and Seattle's Best in store 2, which did not match the rest of the data. To correct this, we subtracted 2 from each coefficient.
- Finally, for some of the contexts $\mathbf{c}$, $\beta_i(\mathbf{c})$,$i \in \{$*Ladro, CaffeVita*$\}$ is negative using the estimated values. This leads to an upwards sloping demand curve which violates our theory. This is due to negative estimated β_{Ladro,Q_1}, β_{Ladro,Q_3}, $\beta_{CaffeVita,Q_3}$, $\beta_{CaffeVita,Q_4}$ coefficients. To deal with this, we increased the estimated values for these coefficients by either two or four to ensure that $\beta_i(\mathbf{c})$, $i \in \{$*Ladro, CaffeVita*$\}$ is still positive for all possible eight context vectors. Precisely, we added two to coefficients β_{Ladro,Q_1}, β_{Ladro,Q_3}, $\beta_{CaffeVita,Q_3}$; and added four to $\beta_{CaffeVita,Q_4}$.

Experimental Setting: We now describe the experimental setting we use for this experiment.

- *Batch Size:* We chose the batch size for retraining our models to be 1. That is we retrain our models after each sample.
- *Range of decision variables:* For simulations in this section we use $\ell = \$0.25$ to $u = \$40$ for prices and display variables are constrained to be in the simplex i.e. $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^{10}_{\geq 0} : \sum_{i=1}^{10} \mathbf{x}_i = 1\}$ as we had in previous simulations.
- *Initial Exploration Phase ($\tau_{explore}$):* For each algorithm, we begin their training them with eight price

Parameter	Estimated	Modified	Parameter	Estimated	Modified	Parameter	Estimated	Modified
$\alpha_{Peet'sCoffee,Q_1}$	1.167	1.167	$\beta_{Peet'sCoffee,Q_1}$	1.989	1.989	$\gamma_{Peet'sCoffee,Q_1}$	-0.280	-0.280
$\alpha_{Peet'sCoffee,Q_2}$	0.458	0.458	$\beta_{Peet'sCoffee,Q_2}$	1.375	1.375	$\gamma_{Peet'sCoffee,Q_2}$	0.081	0.081
$\alpha_{Peet'sCoffee,Q_3}$	0.442	0.442	$\beta_{Peet'sCoffee,Q_3}$	1.449	1.449	$\gamma_{Peet'sCoffee,Q_3}$	0.347	0.347
$\alpha_{Peet'sCoffee,Q_4}$	1.259	1.259	$\beta_{Peet'sCoffee,Q_4}$	2.156	2.156	$\gamma_{Peet'sCoffee,Q_4}$	-0.182	-0.182
$\alpha_{Peet'sCoffee,S_1}$	0.864	0.864	$\beta_{Peet'sCoffee,S_1}$	2.749	2.749	$\gamma_{Peet'sCoffee,S_1}$	0.219	0.219
$\alpha_{Peet'sCoffee,S_2}$	2.462	2.462	$\beta_{Peet'sCoffee,S_2}$	4.220	4.220	$\gamma_{Peet'sCoffee,S_2}$	-0.253	-0.253
α_{CTLBR,Q_1}	0.626	0.626	β_{CTLBR,Q_1}	1.986	1.986	γ_{CTLBR,Q_1}	-0.326	-0.326
α_{CTLBR,Q_2}	-0.050	-0.050	β_{CTLBR,Q_2}	0.630	0.630	γ_{CTLBR,Q_2}	0.581	0.581
α_{CTLBR,Q_3}	-0.161	-0.161	β_{CTLBR,Q_3}	0.680	0.680	γ_{CTLBR,Q_3}	1.064	1.064
α_{CTLBR,Q_4}	0.308	0.308	β_{CTLBR,Q_4}	1.209	1.209	γ_{CTLBR,Q_4}	-1.781	-1.781
α_{CTLBR,S_1}	-2.181	-2.181	β_{CTLBR,S_1}	0.853	0.853	γ_{CTLBR,S_1}	0.159	0.159
α_{CTLBR,S_2}	2.905	2.905	β_{CTLBR,S_2}	3.653	3.653	γ_{CTLBR,S_2}	-0.620	-0.620
$\alpha_{Starbucks,Q_1}$	0.203	0.203	$\beta_{Starbucks,Q_1}$	0.475	0.475	$\gamma_{Starbucks,Q_1}$	0.354	0.354
$\alpha_{Starbucks,Q_2}$	0.059	0.059	$\beta_{Starbucks,Q_2}$	0.598	0.598	$\gamma_{Starbucks,Q_2}$	0.212	0.212
$\alpha_{Starbucks,Q_3}$	0.044	0.044	$\beta_{Starbucks,Q_3}$	0.493	0.493	$\gamma_{Starbucks,Q_3}$	-0.483	-0.483
$\alpha_{Starbucks,Q_4}$	0.842	0.842	$\beta_{Starbucks,Q_4}$	1.798	1.798	$\gamma_{Starbucks,Q_4}$	0.307	0.307
$\alpha_{Starbucks,S_1}$	1.148	1.148	$\beta_{Starbucks,S_1}$	3.364	3.364	$\gamma_{Starbucks,S_1}$	0.389	0.389
$\alpha_{Starbucks,S_2}$	**0.000**	**-1.182**	$\beta_{Starbucks,S_2}$	**0.000**	**4.071**	$\gamma_{Starbucks,S_2}$	**0.000**	**-2.145**
$\alpha_{Stumptown,Q_1}$	3.670	3.670	$\beta_{Stumptown,Q_1}$	3.680	3.680	$\gamma_{Stumptown,Q_1}$	-0.076	-0.076
$\alpha_{Stumptown,Q_2}$	1.035	1.035	$\beta_{Stumptown,Q_2}$	1.550	1.550	$\gamma_{Stumptown,Q_2}$	-0.226	-0.226
$\alpha_{Stumptown,Q_3}$	-0.723	-0.723	$\beta_{Stumptown,Q_3}$	0.137	0.137	$\gamma_{Stumptown,Q_3}$	-0.161	-0.161
$\alpha_{Stumptown,Q_4}$	-0.612	-0.612	$\beta_{Stumptown,Q_4}$	-0.005	-0.005	$\gamma_{Stumptown,Q_4}$	0.825	0.825
$\alpha_{Stumptown,S_1}$	0.957	0.957	$\beta_{Stumptown,S_1}$	2.621	2.621	$\gamma_{Stumptown,S_1}$	0.072	0.072
$\alpha_{Stumptown,S_2}$	2.412	2.412	$\beta_{Stumptown,S_2}$	2.741	2.741	$\gamma_{Stumptown,S_2}$	0.290	0.290
$\alpha_{Seattle'sBest,Q_1}$	1.779	1.779	$\beta_{Seattle'sBest,Q_1}$	4.294	4.294	$\gamma_{Seattle'sBest,Q_1}$	-0.413	-0.413
$\alpha_{Seattle'sBest,Q_2}$	-0.369	-0.369	$\beta_{Seattle'sBest,Q_2}$	0.251	0.251	$\gamma_{Seattle'sBest,Q_2}$	**0.000**	**0.096**
$\alpha_{Seattle'sBest,Q_3}$	-0.516	-0.516	$\beta_{Seattle'sBest,Q_3}$	-0.605	-0.605	$\gamma_{Seattle'sBest,Q_3}$	0.385	0.385
$\alpha_{Seattle'sBest,Q_4}$	0.408	0.408	$\beta_{Seattle'sBest,Q_4}$	1.720	1.720	$\gamma_{Seattle'sBest,Q_4}$	0.357	0.357
$\alpha_{Seattle'sBest,S_1}$	1.301	1.301	$\beta_{Seattle'sBest,S_1}$	5.660	5.660	$\gamma_{Seattle'sBest,S_1}$	0.329	0.329
$\alpha_{Seattle'sBest,S_2}$	**0.000**	**-1.404**	$\beta_{Seattle'sBest,S_2}$	**0.000**	**4.293**	$\gamma_{Seattle'sBest,S_2}$	**0.000**	**-2.367**
$\alpha_{Tony'sCoffee,Q_1}$	0.990	0.990	$\beta_{Tony'sCoffee,Q_1}$	2.097	2.097	$\gamma_{Tony'sCoffee,Q_1}$	0.034	0.034
$\alpha_{Tony'sCoffee,Q_2}$	0.015	0.015	$\beta_{Tony'sCoffee,Q_2}$	1.087	1.087	$\gamma_{Tony'sCoffee,Q_2}$	-0.218	-0.218
$\alpha_{Tony'sCoffee,Q_3}$	-0.504	-0.504	$\beta_{Tony'sCoffee,Q_3}$	0.596	0.596	$\gamma_{Tony'sCoffee,Q_3}$	0.059	0.059
$\alpha_{Tony'sCoffee,Q_4}$	-0.430	-0.430	$\beta_{Tony'sCoffee,Q_4}$	0.327	0.327	$\gamma_{Tony'sCoffee,Q_4}$	0.296	0.296
$\alpha_{Tony'sCoffee,S_1}$	-0.190	-0.190	$\beta_{Tony'sCoffee,S_1}$	2.434	2.434	$\gamma_{Tony'sCoffee,S_1}$	0.034	0.034
$\alpha_{Tony'sCoffee,S_2}$	0.260	0.260	$\beta_{Tony'sCoffee,S_2}$	1.673	1.673	$\gamma_{Tony'sCoffee,S_2}$	0.137	0.137
$\alpha_{CaffeUmbria,Q_1}$	2.480	2.480	$\beta_{CaffeUmbria,Q_1}$	3.973	3.973	$\gamma_{CaffeUmbria,Q_1}$	**0.000**	**-0.143**
$\alpha_{CaffeUmbria,Q_2}$	0.819	0.819	$\beta_{CaffeUmbria,Q_2}$	2.226	2.226	$\gamma_{CaffeUmbria,Q_2}$	-0.244	-0.244
$\alpha_{CaffeUmbria,Q_3}$	-0.548	-0.548	$\beta_{CaffeUmbria,Q_3}$	0.644	0.644	$\gamma_{CaffeUmbria,Q_3}$	0.109	0.109
$\alpha_{CaffeUmbria,Q_4}$	-0.043	-0.043	$\beta_{CaffeUmbria,Q_4}$	0.808	0.808	$\gamma_{CaffeUmbria,Q_4}$	0.211	0.211
$\alpha_{CaffeUmbria,S_1}$	1.895	1.895	$\beta_{CaffeUmbria,S_1}$	5.267	5.267	$\gamma_{CaffeUmbria,S_1}$	-0.085	-0.085
$\alpha_{CaffeUmbria,S_2}$	0.814	0.814	$\beta_{CaffeUmbria,S_2}$	2.384	2.384	$\gamma_{CaffeUmbria,S_2}$	0.161	0.161
α_{Ladro,Q_1}	-4.237	-4.237	β_{Ladro,Q_1}	**-2.252**	**-0.252**	γ_{Ladro,Q_1}	**0.000**	**-0.143**
α_{Ladro,Q_2}	3.774	3.774	β_{Ladro,Q_2}	4.364	4.364	γ_{Ladro,Q_2}	-0.359	-0.359
α_{Ladro,Q_3}	2.304	2.304	β_{Ladro,Q_3}	3.048	3.048	γ_{Ladro,Q_3}	0.422	0.422
α_{Ladro,Q_4}	-3.812	-3.812	β_{Ladro,Q_4}	**-2.284**	**-0.284**	γ_{Ladro,Q_4}	**0.000**	**-0.222**
α_{Ladro,S_1}	-1.430	-1.430	β_{Ladro,S_1}	1.846	1.846	γ_{Ladro,S_1}	**0.000**	**0.232**
α_{Ladro,S_2}	-0.542	-0.542	β_{Ladro,S_2}	1.030	1.030	γ_{Ladro,S_2}	0.063	0.063
$\alpha_{CaffeVita,Q_1}$	0.372	0.372	$\beta_{CaffeVita,Q_1}$	1.134	1.134	$\gamma_{CaffeVita,Q_1}$	**0.000**	**-0.143**
$\alpha_{CaffeVita,Q_2}$	2.026	2.026	$\beta_{CaffeVita,Q_2}$	2.608	2.608	$\gamma_{CaffeVita,Q_2}$	-0.242	-0.242
$\alpha_{CaffeVita,Q_3}$	-2.182	-2.182	$\beta_{CaffeVita,Q_3}$	**-0.357**	**1.643**	$\gamma_{CaffeVita,Q_3}$	-0.110	-0.110
$\alpha_{CaffeVita,Q_4}$	-3.875	-3.875	$\beta_{CaffeVita,Q_4}$	**-1.927**	**2.073**	$\gamma_{CaffeVita,Q_4}$	0.717	0.717
$\alpha_{CaffeVita,S_1}$	-0.826	-0.826	$\beta_{CaffeVita,S_1}$	1.857	1.857	$\gamma_{CaffeVita,S_1}$	0.046	0.046
$\alpha_{CaffeVita,S_2}$	-2.833	-2.833	$\beta_{CaffeVita,S_2}$	-0.400	-0.400	$\gamma_{CaffeVita,S_2}$	0.319	0.319
α_{Other,Q_1}	0.502	0.502	β_{Other,Q_1}	0.867	0.867	γ_{Other,Q_1}	-0.294	-0.294
α_{Other,Q_2}	-0.011	-0.011	β_{Other,Q_2}	0.399	0.399	γ_{Other,Q_2}	1.283	1.283
α_{Other,Q_3}	-0.080	-0.080	β_{Other,Q_3}	0.329	0.329	γ_{Other,Q_3}	1.430	1.430
α_{Other,Q_4}	0.102	0.102	β_{Other,Q_4}	0.260	0.260	γ_{Other,Q_4}	-2.747	-2.747
α_{Other,S_1}	-0.556	-0.556	β_{Other,S_1}	0.592	0.592	γ_{Other,S_1}	0.926	0.926
α_{Other,S_2}	1.069	1.069	β_{Other,S_2}	1.263	1.263	γ_{Other,S_2}	-1.255	-1.255

Table A8: Linear model estimates and modified estimates on Nielsen data with quarter and store dummy contexts. Coefficients that have been modified are shown in bold.

vectors uniformly chosen.

- *Parameters:* The parameters we use are the modified parameters based on two stores of Nielsen data that we discussed earlier. The parameters are listed in Table A8.
- *No. of runs:* We perform 10 replications of each algorithm.
- *Langavin parameters:* We set $\eta_t = .03/t$, $\psi_t = \psi = 1$, and $N_t = 100$

Simulation Running Time

On an Ubuntu machine equipped with a 64-core Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz, we utilized 48 cores for the simulations. Each run of every method operated on a separate thread. Table A9 presents the average running times per time step in seconds for each method. Notably, TS-Langevin demonstrates nearly identical running times to Greedy and M3P. Both Greedy and TS-Langevin employ the same number of optimization steps ($N_t = 100$) and batch size.

Algorithm	Mean (Std)
Greedy	3.6123 (0.0417)
TS-Langevin	3.5460 (0.0436)
M3P	2.9417 (0.0416)

Table A9: Average running times per time step in seconds for each method.

K Supplementary Material for NonLinear Contextual Experiments 7.3

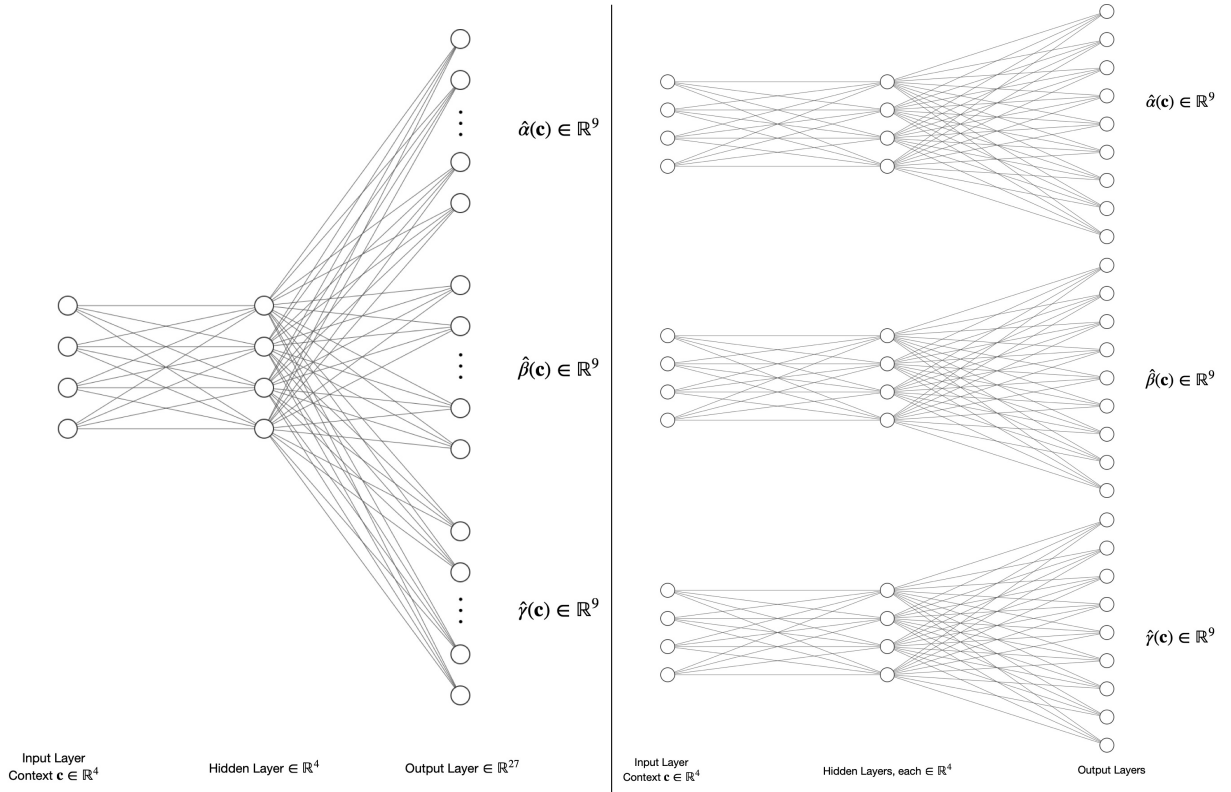
Cluster Center	Point in $\mathbb{R}^4$
C_1	(1.5333, 0.2791, -0.3047, 0.3206)
C_2	(1.2733, 1.1628, -1.8624, -0.1311)
C_3	(-0.8562, 0.9631, 1.1349, 0.6243)
C_4	0.5939, 1.1283, -0.8742, -0.1140)
C_5	(-0.2955, -0.0915, 1.4045, 0.6518)
C_6	(-0.8897, 1.0243, -1.8955, -1.8797)
C_7	(-0.1244, -1.3115, 1.4395, 0.0780)
C_8	(-0.8430, 0.5344, -0.1450, 0.6617)

Table A10: Centers of the Gaussian mixture model for non-linear contextual setup

We consider a case with $K = 9$ products. To generate the parameters $(\bar{\alpha}_g, \bar{\beta}_g, \bar{\gamma}_g)_{g\in\{1,\cdots,8\}}$ for each group we use the same set of parameters as we used in 6.2. For setting the demand parameters of each group for each product, we randomly select one of the columns in the table. The final realization of the parameters $(\bar{\alpha}_g, \bar{\beta}_g, \bar{\gamma}_g)_{g\in\{1,\cdots,8\}}$ are shown in Table A11.

Experimental Setting: We describe the experimental setting we use for this experiment.

- *Batch Size:* We use a batch size of 1, meaning we retrain our model after each sample.
- *Range of decision variables:* In this section we use $\ell = \$0.25$ to $u = \$35$ for prices and display variables are constrained to be in the simplex i.e. $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^9_{\geq 0} : \sum_{i=1}^9 \mathbf{x}_i = 1\}$ as we had in previous simulations.
- *Initial Exploration Phase ($\tau_{explore}$):* For each algorithm, we begin the training with eight price vectors uniformly chosen in the price range for each product.
- *No. of runs:* We perform 16 replications of each algorithm.
- *Langavin parameters:* We set $\eta_t = .03/t$, $\psi_t = \psi = 1$, and $N_t = 100$



(a) Neural Network With Shared Hidden Layer

(b) Neural Network With Separate Hidden Layers

Figure A7: The two neural network architectures used for estimating the non-linear parameter function.

	Product 1	Product 2	Product 3	Product 4	Product 5	Product 6	Product 7	Product 8	Product 9
$\bar{\alpha}_1$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\bar{\alpha}_2$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\bar{\alpha}_3$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\bar{\alpha}_4$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\bar{\alpha}_5$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\bar{\alpha}_6$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\bar{\alpha}_7$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\bar{\alpha}_8$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\bar{\beta}_1$	0.2500	0.2000	0.3000	0.1000	0.2000	0.4500	0.1000	0.4000	0.4500
$\bar{\beta}_2$	0.2500	0.1000	0.1000	0.1500	0.2500	0.4500	0.4500	0.4500	0.4000
$\bar{\beta}_3$	0.5000	0.2000	0.2000	0.5000	0.2500	0.1500	0.3500	0.1500	0.4500
$\bar{\beta}_4$	0.2000	0.1000	0.5000	0.3500	0.1500	0.3500	0.1000	0.4000	0.1000
$\bar{\beta}_5$	0.5000	0.3500	0.3000	0.2500	0.4000	0.3000	0.2500	0.1000	0.1500
$\bar{\beta}_6$	0.5000	0.5000	0.1000	0.4000	0.5000	0.4500	0.5000	0.4000	0.3000
$\bar{\beta}_7$	0.1000	0.3000	0.4000	0.2000	0.1000	0.4000	0.4000	0.3000	0.5000
$\bar{\beta}_8$	0.4000	0.3000	0.2500	0.1500	0.3500	0.1000	0.3500	0.2000	0.3000
$\bar{\gamma}_1$	0.2000	0.5000	0.8000	0.8000	0.5000	0.2000	0.8000	0.5000	0.2000
$\bar{\gamma}_2$	0.2000	0.8000	0.8000	0.3000	0.2000	0.2000	0.2000	0.2000	0.5000
$\bar{\gamma}_3$	0.1000	0.5000	0.5000	0.1000	0.2000	0.3000	0.3000	0.3000	0.2000
$\bar{\gamma}_4$	0.5000	0.8000	0.1000	0.3000	0.3000	0.3000	0.8000	0.5000	0.8000
$\bar{\gamma}_5$	0.1000	0.3000	0.8000	0.2000	0.5000	0.8000	0.2000	0.8000	0.3000
$\bar{\gamma}_6$	0.1000	0.1000	0.8000	0.5000	0.1000	0.2000	0.1000	0.5000	0.8000
$\bar{\gamma}_7$	0.8000	0.8000	0.5000	0.5000	0.8000	0.5000	0.5000	0.8000	0.1000
$\bar{\gamma}_8$	0.5000	0.8000	0.2000	0.3000	0.3000	0.8000	0.3000	0.5000	0.8000

Table A11: Parameters of $\alpha(\mathbf{c}), \beta(\mathbf{c}), \gamma(\mathbf{c})$ piecewise linear function.

L Table of notations

In this section, we provide a table of notations used in this paper.

T	Time Horizon/rounds
K	number of products
$\mathbf{x}, \mathcal{X}$	marketing mix allocation
$\mathbf{p}, \mathcal{P}$	price vector
$\mathcal{H}$	history
$\theta, \alpha, \beta, \gamma$	parameter vector
B	bounds for $\mathbf{x}$
M	bounds for α, β, γ
U	utility function
$\mathcal{L}$	log-likelihood
V_t	covariance matrix
κ	problem parameter for contextual case
μ	demand function
$R_\theta(\mathbf{p}, \mathbf{x})$	objective function

Table A12: Notation used throughout

Q_i	quarter identifier $i \in \{1, 2, 3, 4\}$
C	cluster centers
$\mathcal{S}$	stores
u	bounds for price
w	feature variables
o	display variables
i	brands
Q	market share

Table A13: Notations used in experiments

e_i	$\exp(\alpha_i - \beta_i p_i + \gamma_i x_i)$
λ	regularization
$G(e)$	$\sum_{i \in [K]} e_i$
H	Hessian of $R_c(\mathbf{p}, \mathbf{x})$
μ_i	multinomial probability of item i
S	upper bound for $\|\theta\|$
$\psi_t(\delta)$	upper bound for the bad event $\mathcal{E}$
W_t	empirical covariance matrix

Table A14: Notations used in proof

References

P. Agrawal, V. Avadhanula, and T. Tulabandhula. A tractable online learning algorithm for the multinomial logit contextual bandit. *arXiv preprint arXiv:2011.14033*, 2020.

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

G.-Y. Ban and N. B. Keskin. Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science*, 2021.

H. Bastani, D. Simchi-Levi, and R. Zhu. Meta dynamic pricing: Transfer learning across experiments. *Management Science*, 2021.

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

O. Besbes and A. Zeevi. On the (surprising) sufficiency of linear models for dynamic pricing with demand learning. *Management Science*, 61(4):723–739, 2015.

J. Broder and P. Rusmevichientong. Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980, 2012.

W. C. Cheung, D. Simchi-Levi, and H. Wang. Dynamic pricing and demand learning with limited price experimentation. *Operations Research*, 65(6):1722–1731, 2017.

A. V. Den Boer. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and management science*, 20(1):1–18, 2015.

A. V. den Boer and B. Zwart. Simultaneously learning and optimizing using controlled variance pricing. *Management science*, 60(3):770–783, 2014.

L. Faury, M. Abeille, C. Calauzènes, and O. Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

K. J. Ferreira, D. Simchi-Levi, and H. Wang. Online network revenue management using thompson sampling. *Operations research*, 66(6):1586–1602, 2018.

R. Ganti, M. Sustik, Q. Tran, and B. Seaman. Thompson sampling for dynamic pricing. *arXiv preprint arXiv:1802.03050*, 2018.

V. Goyal and N. Perivier. Dynamic pricing and assortment under a contextual mnl demand. *arXiv preprint arXiv:2110.10018*, 2021.

K. G. Jamieson, L. Jain, C. Fernandez, N. J. Glattard, and R. Nowak. Next: A system for real-world development, evaluation, and application of active learning. *Advances in neural information processing systems*, 28, 2015.

A. Javanmard. Perishability of data: dynamic pricing under varying-coefficient models. *The Journal of Machine Learning Research*, 18(1):1714–1744, 2017.

A. Javanmard and H. Nazerzadeh. Dynamic pricing in high-dimensions. *arXiv preprint arXiv:1609.07574*, 2016.

A. Javanmard, H. Nazerzadeh, and S. Shao. Multi-product dynamic pricing in high-dimensions with heterogeneous price sensitivity. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2652–2657. IEEE, 2020.

N. B. Keskin and A. Zeevi. Dynamic pricing with an unknown demand model: Asymptotically optimal semi-myopic policies. *Operations research*, 62(5):1142–1167, 2014.

R. Kleinberg and T. Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pages 594–605. IEEE, 2003.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018.

N. Mehta, P. Detroja, and A. Agashe. "amazon changes prices on its products about every 10 minutes — here's how and why they do it". *Business Insider*, Aug 2018. URL https://www.businessinsider.com/amazon-price-changes-2018-8?international=true&r=US&IR=T.

S. Miao and X. Chao. Dynamic joint assortment and pricing optimization with demand learning. *Manufacturing & Service Operations Management*, 23(2):525–545, 2021.

K. Misra, E. M. Schwartz, and J. Abernethy. Dynamic online pricing with incomplete information using multiarmed bandit experiments. *Marketing Science*, 38(2):226–252, 2019.

J. Mueller, V. Syrgkanis, and M. Taddy. Low-rank bandit methods for high-dimensional dynamic pricing. *arXiv preprint arXiv:1801.10242*, 2018.

J. W. Mueller, V. Syrgkanis, and M. Taddy. Low-rank bandit methods for high-dimensional dynamic pricing. *Advances in Neural Information Processing Systems*, 32, 2019.

S. Qiang and M. Bayati. Dynamic pricing with demand covariates. *Available at SSRN 2765257*, 2016.

V. Shah, R. Johari, and J. Blanchet. Semi-parametric dynamic contextual pricing. *Advances in Neural Information Processing Systems*, 32, 2019.

Y. Wang, B. Chen, and D. Simchi-Levi. Multimodal dynamic pricing. *Management Science*, 2021.

I. Weaver and V. Kumar. Nonparametric bandits leveraging informational externalities to learn the demand curve. Working Paper, 2022.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

J. Xu and Y.-x. Wang. Logarithmic regret in feature-based dynamic pricing. *arXiv preprint arXiv:2102.10221*, 2021.