Extended Abstract

Motivation We consider the problem of setting optimal retail prices and promotions for a multi-product category when demand information needs to be explored. Retail pricing and promotion planning are big spending items for retailers such as Safeway, Macy's and consumer packaged goods (i.e., CPG) manufacturers such as P&G and General Mills. CPG companies worldwide invest 20% of their revenue in trade promotions [7]. Current promotion planning is largely done by using heuristics and judgments from human planners. Up to 59% promotions are losing money, so there is an opportunity to use model-based reinforcement learning to improve the quality of pricing and promotion planning for retailers and consumer packaged goods manufacturers [7].

Method We developed a **causal model-based reinforcement learning (RL)** framework for **adaptive pricing and promotion planning**. The proposed RL agent can automate decision making by balancing the sample efficient exploration of demand parameters and the optimization of price and promotion (under the learned demand parameters). Specifically, we have used the discrete choice model of consumer demand as a causal world model. This helps to handle the high-dimensional action space and coupled rewards by using the mapping from (true) demand parameters to optimal (retail) prices and promotions. For exploration, we used the d-optimal design of the Fisher information matrix to select prices and promotions from candidate proposals, which maximize the learning of the demand parameters. By alternating d-optimal exploration and exploitation (i.e., price and promotion optimization conditional on demand parameters), we showed that we can achieve more sample efficient price and promotion optimization compared to a greedy method or Thomspon sampling method.

Implementation The proposed algorithm is implemented in Python and PyTorch. To reduce overall run-time, we have used Ray for parallelization. In the algorithm, We have alternated exploration and exploitation rounds. In the exploration round, We used D-optimal design with Fisher Information Matrix to select price and promotion combinations, which can maximize demand parameter learning. Demand parameters for a multi-nomial logit demand model are estimated by using maximum likelihood (i.e., MLE.) In the exploitation round, we used the learned demand model to find and set optimal prices and promotions. Based on stylized simulated dataset, we have assessed the performance of the proposed algorithm compared to the baseline of greedy algorithm.

Results After implementing the algorithm, we calculated a cumulative regret and % of optimal profit and compared the result with the baseline cumulative regret from a "greedy" method where the algorithm initially plays some random prices for a period T (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 promotion according to the estimated model.In other words, "greedy" methods only do "exploitation" after initial periods of random price exploration. The result showed that we can reduce the cumulative regret by conducting sample efficient demand parameter explorations by using the proposed method. In addition, the proposed D-optimal algorithm is more effective compared to Thompson sampling approach, which was state-of-the-art in the literature.

Discussion In the greedy setting, we found that the lack of strategic exploration in prices make it stuck in sub-optimal prices and promotions. As a result, the data it gathers are not very informative. In contrast, in our proposed d-optimal exploration algorithm, more sample efficient exploration helps to conduct more informative price and promotion exploration, which helps to recover true demand parameters. This resulted in smallest cumulative regret and large % of optimal profits among three algorithms that we compared: (1) baseline of "Greedy", (2) Thompson Sampling (TS), and (3) proposed D-Optimal exploration.

Conclusion In conclusion, this study provides an effective approach to adaptive pricing and promotions with a discrete choice model as a causal world model. Using a D-optimal exploration approach based on Fisher information matrix, we develop a regret minimizing, (or profit maximizing) algorithm for a retailer. Using simulations based on stylized settings, we show that the proposed method significantly outperforms existing baseline approaches. This approach can be used to optimize prices and promotions adaptively while simultaneously learning the demand model.

3PO - Causal Model-Based Reinforcement Learning Agent for Adaptive Pricing and Promotion Planning

Minha Hwang
Stanford Online
drsquare@stanford.edu

Abstract

We consider the problem of setting optimal retail prices and promotions for a multi-product category when demand information needs to be explored. We developed a **causal model-based reinforcement learning (RL)** framework for **adaptive pricing and promotion planning**. The proposed RL agent can automate decision making by balancing the sample efficient exploration of demand parameters and the optimization of price and promotion (under the learned demand parameters). Specifically, we have used the discrete choice model of consumer demand as a causal world model. This helps to handle the high-dimensional action space and coupled rewards by using the mapping from (true) demand parameters to optimal (retail) prices and promotions. For exploration, we used the d-optimal design of the Fisher information matrix to select prices and promotions from candidate proposals, which maximize the learning of the demand parameters. By alternating d-optimal exploration and exploitation (i.e., price and promotion optimization conditional on demand parameters), we showed that we can achieve more sample efficient price and promotion optimization compared to "greedy" or "Thompson sampling" baselines.

1 Introduction

We consider the problem of setting optimal retail prices and promotions for a multi-product category when demand information needs to be explored. Retailers typically sell a large number of products in a given category. They need to make two key decisions: (1) setting retail prices (continus variables) and (2) assigning promotions (binary assignment.). For example, in Figure 1, Safeway has 146 coca cola products (even more if we consider other cola products such as Pepsi or store brands) for which the retailer has to decide on retail prices and promotion assignment.

If the retailer has perfect information on the underlying demand model, it can solve a joint optimization problem for optimal prices and promotion to maximize category profits. However, retail stores lack such demand information. The two standard ways to learn demand information are: (1) use observational data to estimate demand models or (2) run A/B tests. The former can lead to misleading demand parameter estimates, especially when the price movements are limited and concentrated in a few products. There are also confounder problems. The latter approach is very costly and does not scale with a large number of products. Thus, we investigate an algorithm based on the causal demand model-based reinforcement learning (RL), which can balance sample efficient exploration with exploitation (i.e., price and promotion optimization under learned demand parameters).

2 Related Work

The proposed work relates to three broad streams of literature - (1) the adaptive pricing work in operations research and computer science literature (also referred as dynamic pricing), (2) discrete

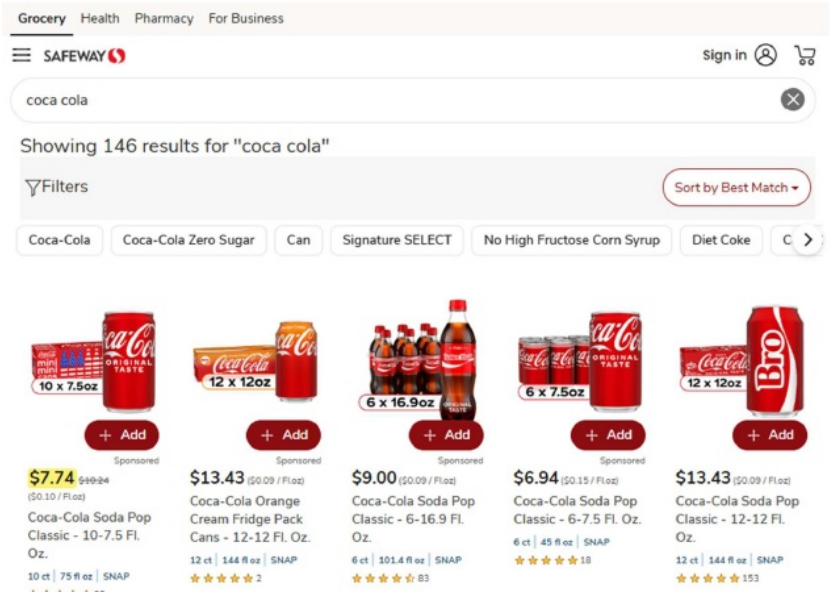

Figure 1: An example page of an online retailer selling cola

choice model of consumer demand literature in economics and marketing, and (3) model-based RL literature in computer science.

Adaptive Pricing: The work on adaptive pricing is quite extensive in operation research literature. However, most of studies follow a very simplified setup: one retailer/seller who sells a single product for a fixed number of rounds and chooses a price and then observe an associated demand[5]. However, multi-products setting with as many as 5000 products are very common in practice, which significantly limits practical applicability of these models.
In addition, many of the works in the pricing literature employ "forced/random exploration", which is not managerially feasible. The only exception is model-based exploration by Lalit et al[5]. Both parametric and non-parametric models are explored in the literature. It is shown that parametric approach is more appropriate given the limitation in the data[5].

Consumer Demand Model: Academics in Marketing Science and Economics proposed pricing and promotion planning models based on an econometric modeling of consumer demand. [8] Econometric promotion decision support models are based on either (1) log-log model or (2) discrete choice models. In the log-log model, log of the unit sales (i.e., volume) is used as response variable and log of (own) prices and other control variables are used as features. This model specification is motivated from the observations that the multiplicative model specification shows better predictive accuracy in point-of-sale (POS) data. In the multiplicative model specification, the impact of each feature is multiplied to each other. Taking log of multiplicative model gives an additive model in log-log space, which can be estimated with the linear regression methods or hierarchical Bayes method. In addition, the use of log(price) allows to use the coefficient of log(price) variable as the estimate of price elasticity, which is defined as:

$$E_d = \frac{dQ}{dp} \cdot \frac{p}{Q} = \frac{log(Q)}{log(p)}$$

where Q is quantity sold, and p is a price of the good.

For model estimation, an ordinary least square (OLS) linear regression is used if we can assume the same price elasticity across products and location, without having too many parameters. This is restrictive assumption, thus mixed effect model, which allows different price elasticities across products and stores are used to address the limitation of the constant price elasticity model [2]. The log-log model suffers the curse of dimensionality with respect to cross price elasticity estimation since cross price elasticy scales with the square of number of products. An aggregate

or individual-level discrete choice models were proposed to address this curse of dimensionality challenge [1]. Discrete choice models impose economic theory-based structure on the model specification to handle the limitation in the data [1]. This can be viewed as a causal world model of demand in which a microeconomic theory of consumer demand complements the data gaps.

For machine learning models, which focused on demand forecasting tasks, it has been shown that tree-based models such as boosted tree models (e.g., XGBoost, LightGBM) or random forests perform the best in predictive accuracy for tabular panel data. However, it is unclear whether these models can provide good counter-factual simulation results, which match with the economic theory. As an example, if prices and unit sales have been always higher for products with unobserved high quality, the ML model may learn the correlation between high prices and high sales without recognizing confounding unobserved product quality.

Therefore, I will use the discrete choice model of consumer demand as a base world model for model-based RL approach in this study

Model-based RL: In the realm of model-based reinforcement learning (MBRL), DreamerV3, developed by DeepMind, stands out as a significant advancement. This algorithm demonstrates remarkable generalization across over 150 diverse tasks using a single configuration [3].

DreamerV3's architecture comprises three primary components[3]:
(1) World Model: Learns a latent representation of the environment's dynamics, enabling the agent to predict future states and rewards without direct interaction.
(2) Actor: Determines optimal actions based on the latent representations provided by the world model.
(3) Critic: Evaluates the value of predicted future states to guide the actor's learning process.

The world model utilizes a Recurrent State-Space Model (RSSM) with discrete latent variables, allowing it to capture complex temporal dependencies in the environment. This model is trained using a combination of reconstruction loss, reward prediction loss, and a Kullback-Leibler (KL) divergence term to ensure consistency between predicted and actual latent states [3].

Drawing inspiration from DreamerV3, I propose to apply a model-based RL approach to pricing and promotion planning. However, instead of employing neural networks for the world model, I propose to utilize a discrete choice model of consumer demand, which is causal and more explainable. It has additional benefits of reduced problem space due to the mapping from low-dimensional demand parameters to high-dimensional optimal prices and promotions based on economic theory. Discrete choice models are particularly effective in capturing underlying demand patterns, especially when dealing with limited data, as they model individual decision-making processes and can incorporate various factors influencing consumer choices. We do not need actor or critic models since optimal prices and promotions are defined as long as we can recover true demand model. Therefore, the main problem to solve in our case is (1) how to recover true demand parameters in sample-efficient exploration and (2) how to balance exploration and exploitation.

The comprehensive surveys and research from the Berkeley Artificial Intelligence Research (BAIR) lab, provided valuable insights, which helped me to design and implement model-based RL algorithms [6, 4].

3 Method

Setup

We consider a retailer (e.g., Amazon, Safeway) that sells multiple products in a given product category. A retailer can be either e-commerce site or physical store. In each period, the retailer decides how to price each product and whether to promote them. These promotions can be feature or display promotions such as (1) highlighting a product on a search result page of a e-commerce website, (2) placing products on end-of-the-aisle displays in a physical supermarket, or (3) promoting products in mailers/email sent to customers. The retailer does not know the true demand parameters but tries to maximize category profits over a time horizon by choosing prices and promotions.

Mathematical Problem definition

Suppose that a retailer has a set of K products (in a given category) that it offers to a consumer in each of T periods.[1] In each time period $t,\ 1 \le t \le T$, the retailer must choose a price and a promotion for each of the K products in the category. Let the price vector be

$$\mathbf{p}_t \ = \ (p_{1t}, \ldots, p_{Kt}) \ \in \ \mathcal{P} := [\ell, u]^K,$$

where $\ell, u \in \mathbb{R}_{\ge 0}$ are, respectively, lower and upper bounds on the admissible prices. The promotion allocation is

$$\mathbf{x}_t \ = \ (x_{1t}, \ldots, x_{Kt}) \ \in \ \mathcal{X} \subset \mathbb{R}^K,$$

where $\mathcal{X}$ denotes the feasible set of such allocations. We can assume that $\mathcal{X}$ is finite.)

After observing $(\mathbf{p}_t, \mathbf{x}_t)$, the consumer chooses a product

$$I_t \ \in \ \{0, 1, \ldots, K\},$$

where $I_t = 0$ corresponds to the no-purchase option. Each item i has a marginal cost $m_i \ge 0$ for the retailer. Therefore, in round t the seller's reward (profit) is

$$r_t \ = \ \begin{cases} p_{I_t} - m_{I_t}, & \text{if } I_t \in [K], \\ 0, & \text{if } I_t = 0. \end{cases}$$

The retailer's policy is *adaptive*: the choice of $(\mathbf{p}_t, \mathbf{x}_t)$ may depend on the full history up to round t,

$$\big\{(\mathbf{p}_s, \mathbf{x}_s, I_s, r_s)\big\}_{s=1}^{t-1}.$$

Following econometric demand estimation literature, we assume that the probability the consumer selects product i is given by a random-utility model (i.e., discrete choice model of consumer demand.) Specifically, there exists a parameter vector

$$\theta \ = \ \big[\alpha_1, \ldots, \alpha_K, \ \beta_1, \ldots, \beta_K, \ \gamma_1, \ldots, \gamma_K\big] \ \in \ \mathbb{R}^K \times \mathbb{R}^K_{>0} \times \mathbb{R}^K_{>0}$$

such that the utility the consumer receives from each alternative is modeled accordingly.

The consumer's utility from purchasing product i at price p_{it} with promotion variable x_{it} is

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

where ε_{it} follows a type-1 extreme-value distribution. We allow both the price sensitivity β_i and the promotion response γ_i to differ by product.[2]
The probability that the buyer selects product i in period t, conditional on $(\mathbf{p}_t, \mathbf{x}_t)$ and the past history $\{(\mathbf{p}_s, \mathbf{x}_s, I_s, r_s)\}_{s=1}^{t-1}$, is given by the multi-nomial logit formula

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

where $\theta = [\alpha_1, \ldots, \alpha_K, \beta_1, \ldots, \beta_K, \gamma_1, \ldots, \gamma_K]$ and the associated probability and expectation are denoted $\Pr_\theta$ and $\mathbb{E}_\theta$, respectively.

Therefore, the retailer's expected profit in period t is

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

Throughout we assume bounded parameters,

$$\alpha_i \in [-M, M], \qquad \beta_i \in [0, M], \qquad \gamma_i \in [0, M],$$

[1] Throughout, K and T are positive integers.

[2] Although theory might suggest homogeneous price and promotion effects across items, empirical studies (e.g., **?**) document substantial heterogeneity.

for some known constant $M > 0$. These restrictions $\beta_i \geq 0$ enforces diminishing utility in price, while $\gamma_i \geq 0$ captures the (weakly) positive impact of promotion on utility.

Retailer's Goal

The aim of the retailer is to develop a policy that *simultaneously* learns the profit maximizing price vector and the promotion vector for the K products while balancing exploration and exploitation. For any fixed demand parameter θ, let the profit-maximizing pair be

$$(\mathbf{p}^\star, \mathbf{x}^\star) \;:=\; {}_{\mathbf{p}\in\mathcal{P},\,\mathbf{x}\in\mathcal{X}} R_\theta(\mathbf{p}, \mathbf{x}), \tag{4}$$

where $R_\theta(\mathbf{p}, \mathbf{x})$ is the expected profit defined in (3).

We assume $\mathbf{p}^\star \in \mathcal{P} = [\ell, u]^K$ Taking a Bayesian viewpoint, let the unknown parameter $\theta \sim \Pi_0$, where the prior Π_0 may be uninformative or encode the firm's domain knowledge.
The retailer's goal is to regret minimization: minimize the total loss in profits due to the exploration of the pricing and promotion. The time T cumulative regret is defined as:

$$\begin{aligned} Reg_T^{\mathrm{B}} \;&:=\; \mathbb{E}_{\theta\sim\Pi_0}\Big[\sum_{t=1}^{T}\big(R_\theta(\mathbf{p}^\star, \mathbf{x}^\star) - R_\theta(\mathbf{p}_t, \mathbf{x}_t)\big)\Big] \\ &=\; \mathbb{E}_{\theta\sim\Pi_0}\Big[T\, R_\theta(\mathbf{p}^\star, \mathbf{x}^\star) - \sum_{t=1}^{T} R_\theta(\mathbf{p}_t, \mathbf{x}_t)\Big]. \end{aligned} \tag{5}$$

We refer to Reg_T^{B} simply as the *regret*.

The *simple regret* at period t is

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

Recap of Problem Statement. Based on the problem formulation until now, we recap our problem statement:
Play a sequence $\{(\mathbf{p}_t, \mathbf{x}_t)\}_{t=1}^T$ that minimizes cumulative regret (5).

Change in Exploration Space

The use of demand model allows us to do exploration in low-dimensional demand parameter space instead of high-dimensional prices and promotion spaces. I will show how optimal prices and promotion can be calculated conditional on a true demand model.

Optimal price vector for a fixed promotion. For a given $\mathbf{x} \in \mathcal{X}$, there exists a unique global maximizer

$$\mathbf{p}^\star(\mathbf{x}) \;:=\; {}_{\mathbf{p}\in\mathbb{R}^K_{\geq 0}} R_\theta\big(\mathbf{p}, \mathbf{x}\big),$$

which can be characterized by a fixed-point scalar equation.

Fix $\mathbf{x} \in \mathcal{X}$. The optimal price vector $\mathbf{p}^\star(\mathbf{x}) = (p_1^\star, \ldots, p_K^\star)$ satisfies

$$p_i^\star \;=\; \frac{1}{\beta_i} \;+\; R \;+\; m_i, \qquad i = 1, \ldots, K, \tag{7}$$

where the scalar R solves

$$R \;=\; \sum_{i=1}^{K} \frac{1}{\beta_i} \exp\big[-\big(1 + \beta_i R + \beta_i m_i\big) + \alpha_i + \gamma_i x_i\big]. \tag{8}$$

Optimal promotion - Outer Loop. Lalit et al. show that only a *finite* collection of promotion vectors can be optimal, i.e. we may assume $\mathcal{X}$ is finite, which makes the optimization of fixed promotion tractable. [5]. For details, please refer to Lalit et al.[5]

4 Experimental Setup

4.1 Setup: Experiment with Stylized Setting

We follow Lalit et. al. for the design of numerical experiment in stylized setting. [5]. The key differences are (1) smaller number of purchase decisions and replications due to time constraint: 5,000 purchase decision steps and 10 replications, and (2) exploration of the impact of different batch sizes (10 vs. 500)

We consider a retailer in a market with three products ($K = 3$) that must decide how to set prices and promotions for these products. In each period t the retailer may choose prices $p_{it} \in [0, \$30.00]$ for $i = 1, \ldots, K$ and a binary promotion variable x_{it} that satisfies the simplex constraint

$$\sum_{i=1}^{3} x_{it} = 1, \qquad x_{it} \in \{0, 1\}, \qquad \forall\, t \geq 1.$$

In other words, the firm can promote *at most* one product per period (or promote none), consistent with the result of Theorem 1.

Demand for each product follows the choice model described in the previous section. The parameter values are reported in Table 1, and marginal costs are normalized to zero for simplicity. The table also lists the optimal price $\mathbf{p}_\star$ and promotion $\mathbf{x}_\star$ for each product, along with the corresponding optimal revenue.

Table 1: Parameters and outcomes at the optimal pricing/promotion for a three-product case. The optimal revenue is $R(\mathbf{p}_\star, \mathbf{x}_\star) = \$10.5$.

Parameters	Product 1	Product 2	Product 3
α	1	1	1
β	0.1	0.2	0.3
γ	0.8	0.3	0.5
$\mathbf{p}_\star$	\$20.50	\$15.50	\$13.83
$\mathbf{x}_\star$	1	0	0

So far, I have implemented two baseline policies: (1) Greedy and (2) Thompson Sampling (TS). The TS algorithm is based on Lalit. et al. [5]

In the experiments, We compare two alternative policies the retailer might adopt under different batch setting (10 vs. 500). Note that the prices and promotion are maintained at the same level within the same batch:

1. **Greedy.** A purely myopic benchmark in which, at each period, the firm *(i)* estimates the demand parameters by maximum likelihood (i.e., MLE) from all data observed so far (prices, promotions, and realized market shares) and *(ii)* chooses the price-promotion pair that maximizes the resulting *in-sample* revenue prediction. Thus, Greedy relies exclusively on exploitation and performs no deliberate exploration.

2. **D-optimal Exploration (DO).** This is the proposed algorithm based on sample efficient D-optimal exploration. We have alternated exploration and exploitation rounds. In the exploration round, We used D-optimal design with Fisher information matrix to select price and promotion combinations, which can maximize demand parameter learning. Demand parameters for a multi-nomial logit demand model are estimated by using maximum likelihood (i.e., MLE.) In the exploitation round, we used the learned demand model to find and set optimal prices and promotions.

3. **Thompson Sampling (TS).** This algorithm follows the Bayesian / posterior-sampling approach outlined in Lalit. et. al. to balance exploration and exploitation through randomized draws from the posterior over demand parameters. [5]

We conduct a Monte Carlo experiment with 10 independent replications, each consisting of 5,000 purchase decisions with a bath size of 10. Note that prices are promotions are maintained within the same batch across multiple purchase occasions. Results are averaged across replications and plotted in Figure 1.

5 Results

The results of the experiments are averaged across replications and plotted in Figures 1. The optimal revenue at time t that the firm would have earned playing the optimal price and promotion is \$10.5 × Tİn Figure 1, the cumulative regrets (i.e., optimal revenue - realized revenue) are plotted when the batch size is 10. As the reference paper ([5]) claimed, Thompson sampling (TS-Laplace) resulted in lower cumulative regrets compared to the "Greedy" case, since the "Greedy" case can stick in inferior local optimum and stops exploration. Our proposed D-optimal algorithm (Optimal-DO) achieved the lowest cumulative regret, as expected. This shows that a sample efficient D-optimal exploration based on Fisher information matrix can be very effective compared to the baseline and other approach based on Thompson sampling of demand parameters.

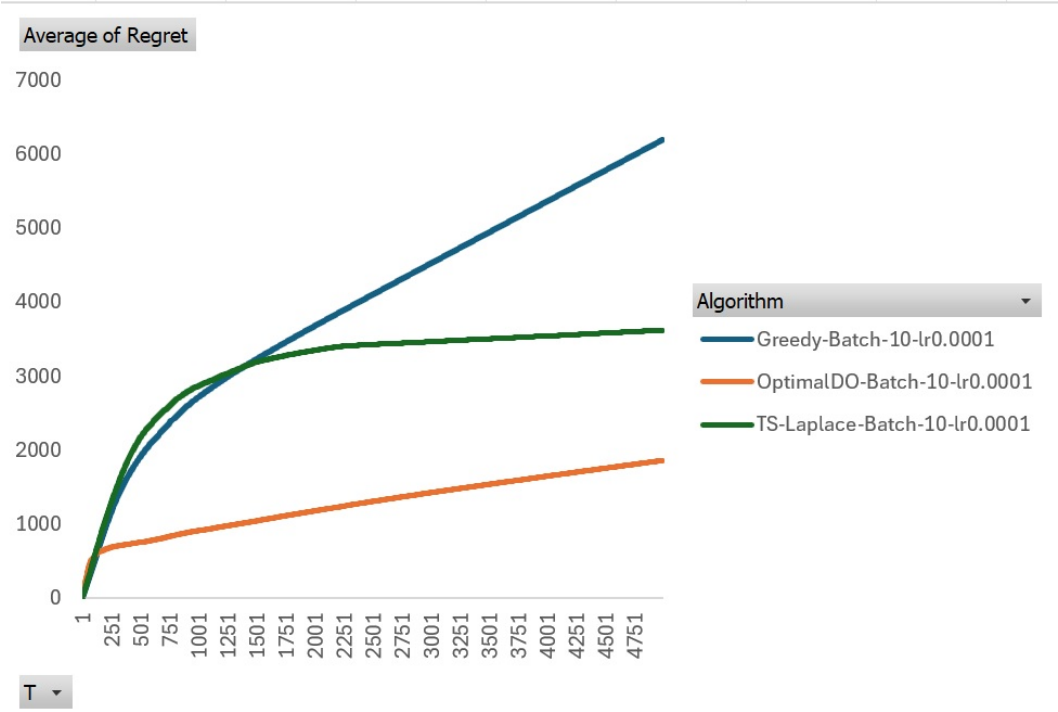

Figure 2: Cumulative Regret with Batch of 10

5.1 Quantitative Evaluation

Table 2 summarizes both (1) cumulative regret and (2) % of optimal profit to compare three algorithms. It clearly shows that our proposed D-optimal exploration method results in the least cumulative regret (1861) and largest % optimal profit (96.5 %). Thompson sampling (TS) is second best, followed by the baseline of "Greedy" algorithm.

5.2 Qualitative Analysis

When we investigated promotion histories, the proposed D-optimal exploration algorithm properly found that it is optimal to promote Product 1. In contrast, a "Greedy" algorithm often stuck in local optimum and select to promote either Product 2 or Product 3. Thompson sampling algorithm (TS) was also able to properly found that it is optimal to promote Product 1.

Table 2: Performance Comparison

Method	Cumulative Regret	% of Optimal Profit
Baseline (Greedy)	6201	88.2%
Our Approach (DO)	1861	96.5%
Thompson Sampling (TS)	3621	93.1%

6 Discussion

In the greedy setting, we found that the lack of strategic exploration in prices makes it stuck in sub-optimal prices and promotions. As a result, the data it gathers are not very informative. In contrast, in our proposed d-optimal exploration algorithm, more sample efficient exploration helps to do more informative price and promotion exploration, which helps to recover true demand parameters. Compared to Thompson sampling method (TS), the proposed D-optimal algorithm is shown to be more sample efficient.

7 Conclusion

In conclusion, this study provides an effective approach to adaptive pricing and promotions with a discrete choice model as a causal world model. Using a D-optimal exploration approach based on Fisher information matrix, we develop a regret minimizing, (or profit maximizing) algorithm for a retailer. Using simulations based on stylized settings, we show that the proposed method significantly outperforms existing baseline approaches. This approach can be used to optimize prices and promotions adaptively while simultaneously learning the demand model.

8 Team Contributions

- **Minha Hwang:** This is a solo project: 100% contribution

Changes from Proposal Not applicable since this is a solo project.

References

[1] Patrick Bajari, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang. Machine learning methods for demand estimation. *American Economic Review*, 105(5):481–85, May 2015.

[2] Robert C. Blattberg and Edward I. George. Shrinkage estimation of price and promotion elasticities: Seemingly unrelated equations. *Journal of American Statistical Association*, 86(414):304–315, Jun 1991.

[3] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2024.

[4] Michael Janner. Model-based reinforcement learning: Theory and practice. *Blog*, Dec 2019.

[5] Erfan Loghmani Blake Mason Hema Yoganarasimhan Lalit Jain, Zhaoqi Li. Effective adaptive exploration of prices and promotions in choice-based demand models. *Marketing Science*, 43(5):925–1151, Sep 2024.

[6] Fan-Ming Luo, Tian Xu, Hang Lai, Xiong-Hui Chen, Weinan Zhang, and Yang Yu. A survey on model-based reinforcement learning, 2022.

[7] Abdul Wahab Shaikh Minha Hwang, Ryan Murphy. How analytics can drive growth in consumer packaged-goods trade promotions. 2019.

[8] Martin Natter, Thomas Reutterer, Andreas Mild, and Alfred Taudes. Practice prize report an assortment-wide decision support system for dynamic pricing and promotion planning in diy retailing. *Marketing Science*, 26(4):576–583, 2007.

A Implementation Details

The proposed algorithm is implemented in Python and PyTorch. To reduce overall run-time, we have used Ray for parallelization. In the algorithm, We have alternated exploration and exploitation rounds. In the exploration round, We used D-optimal design with Fisher Information Matrix to select price and promotion combinations, which can maximize demand parameter learning. Demand parameters for a multi-nomial logit demand model are estimated by using maximum likelihood (i.e., MLE.) In the exploitation round, we used the learned demand model to find and set optimal prices and promotions. Based on stylized simulated dataset, we have assessed the performance of the proposed algorithm compared to the baseline of greedy algorithm.