Pricing Experimental Design: Causal Effect, Expected Revenue and Tail Risk

David Simchi-Levi [1 2 3 4] **Chonghuan Wang** [1 4 5]

Abstract

When launching a new product, historical sales data is often not available, leaving price as a crucial experimental instrument for sellers to gauge market response. When designing pricing experiments, there are three fundamental objectives: estimating the causal effect of price (i.e., price elasticity), maximizing the expected revenue through the experiment, and controlling the tail risk suffering from a very huge loss. In this paper, we reveal the relationship among such three objectives. Under a linear structural model, we investigate the trade-offs between causal inference and expected revenue maximization, as well as between expected revenue maximization and tail risk control. Furthermore, we propose an optimal pricing experimental design, which can flexibly adapt to different desired levels of trade-offs. Through the optimal design, we also explore the relationship between causal inference and tail risk control.

1. Introduction

The importance of understanding the market response to new products or services cannot be overstated for today's online platforms. In the absence of informative historical data, price always serves as the primary or even sole experimental instrument that sellers can use to gain insights into market demand. The seller can use adaptive pricing strategies during experimental periods such as a product launch. As a result, designing efficient and reliable pricing experiments is becoming increasingly crucial (Xu et al. 2019; Bastani et al. 2022). However, in real world, the objectives of the pricing experiments could be very different.

Without a doubt, like most experiments on online platforms (see, e.g., Xiong et al. 2019; Bojinov et al. 2022), one of the primary objectives of pricing experiments is to estimate the casual effect of price, which in our context is also known as price elasticity. This objective has been a central focus in both experimental design literature and fields such as marketing (see, e.g., Tellis 1988; Bijmolt et al. 2005) and operations management (see, e.g., Chintagunta et al. 2002; Kocabıyıkoğlu & Popescu 2011) for the past decades. Estimating price elasticity not only enhances understanding of the market, but also provides valuable *long-term rewards* for the seller. For example, it can inform decisions such as settling on a fixed and permanent price for the future, or adjusting pricing in response to inventory backlogs or upcoming promotions, or even designing new products or services that align with consumer preferences.

Recently, there has been growing attention paid to the in-experiment performance, which refers to the revenue that the seller can collect during the experimental period. This *short-term reward* is also essential and needs to be carefully considered in many cases, particularly for products with relatively short life cycles (Ferreira et al. 2016), limited inventory (Xu et al. 2019) or in the online service economy (Scott 2015). Recent advances in dynamic pricing have been found to be effective in earning short-term rewards, particularly when the price-dependent demand models are not fully known. To achieve optimal in-experiment performance, sellers must strike a delicate balance between exploring (learning about the demand) and exploiting (charging estimated optimal prices) (see, e.g., Keskin & Zeevi 2014; Qiang & Bayati 2016; Ban & Keskin 2021).

There is a trade-off between the long-term and the short-term rewards. Gaining a very accurate understanding of price elasticity through experimentation could disrupt the balance between exploration and exploitation that is necessary for achieving optimal short-term revenue, because it may require more investment in learning demand models. As a result, the pursuit of long-term rewards may come at the expense of short-term rewards. Additionally, the decisions that optimize short-term rewards may not necessarily be the best for long-term rewards. This can result in a "short-sighted" policy, particularly if after-experiment revenues are given more weight than in-experiment revenues. Furthermore, having a high-quality estimator of price elasticity is beneficial for both long-term and short-term rewards. A

[1] Laboratory for Information & Decision Systems, MIT [2] Institute for Data, Systems, and Society, MIT [3] Operations Research Center, MIT [4] Department of Civil and Environmental Engineering, MIT [5] Center for Computational Science and Engineering, MIT. Correspondence to: David Simchi-Levi <dslevi@mit.edu>, Chonghuan Wang <chwang9@mit.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

high-quality estimator is one that can provide accurate and precise estimates, reducing the randomness in estimating the treatment effect. Such a reduction in turn can help reduce uncertainty in the in-experiment rewards the seller can expect. Therefore, one of the central questions in this research is *how to systematically and statistically understand the relationship between long-term rewards after experiments and short-term rewards within experiments*.

For the after-experiment learning, we will focus on the quality of the price elasticity estimator which is measured by the commonly-used squared error between the estimated and real values. For in-experiment learning, the primary metric of importance is the expected revenue loss, also known as the expected regret. This is the expected difference between the optimal pricing decisions made by a clairvoyant policy and the policy we design. In many cases, it may not be possible to repeat experiments multiple times for online platforms due to resource constraints or other factors, and the experiment can be run only once. In such cases, the expected value of the regret may not be informative enough to make a good decision. Therefore, we will also consider the distribution of the regret, specifically the tail of the distribution, which captures the risk of large revenue losses. In summary, the fundamental objectives of pricing experiments that we consider include estimating the causal effect of price, managing the expected within-experiment revenue, and controlling the tail risk. This research aims to reveal the close relationship among these objectives and how to design optimal pricing experiments accordingly.

1.1. Main results and contribution

The main contribution of this work lies in the statistical and systematical understanding of the fundamental relationship among causal effect, expected revenue and tail risk in pricing experimental designs. We also propose an optimal pricing experimental design that can flexibly adapt to different objectives. To our best knowledge, this work is the first to jointly consider the in-experiment and after-experiment revenues in the dynamic pricing literature. We next summarize our main results and highlight our contributions.

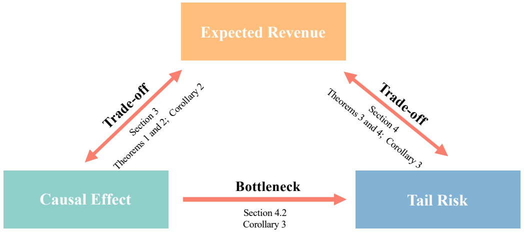

Figure 1. An overview of the main results.

Between causal inference and expected revenue management. We first statistically establish the trade-off between the accuracy of causal estimators and the expected regret in revenue. More specifically, for any given admissible pricing policy and causal estimator, we derive a lower bound describing that the product of the squared error of the estimator and the expected regret can always be lower bounded by a constant independent of the horizon T in the worst case. This means that there is always a trade-off between the two objectives and it is not possible to achieve high accuracy and low regret at the same time. To the best of our knowledge, this is the first result in current literature that captures this trade-off between causal inference and expected revenue management. In addition to this, we have also proposed an optimal experimental design, `RSD`, which adopts the idea of random shock pricing from (Nambiar et al., 2019). Our results show that `RSD` can achieve the lower bound that we derived, indicating that the lower bound is achievable and that `RSD` is optimal in this sense. Furthermore, our `RSD` design is flexible and can be adapted to different desired levels of trade-off between the two objectives, providing a useful tool for practitioners in the field.

Between expected revenue management and tail risk control. We also study the relationship between the expected regret and the tail probability that the realized empirical regret is very large. We discover that it is possible to reduce the tail probability of `RSD` by sacrificing some of the expected regret, when the horizon T is sufficiently large. In other words, while the expected value of the regret may increase, the tail of the distribution of the regret becomes lighter. Furthermore, we derive an information-theoretical lower bound that shows that the tail probability of `RSD` is the best achievable in terms of the dependence on T, up to logarithm factors. This implies that our `RSD` design is not only optimal in terms of the trade-off between accuracy and regret, but also in terms of tail risk. To the best of our knowledge, we are the first to study this important aspect of tail risk in revenue management literature and to explicitly reveal the essential trade-off between revenue management and risk control in dynamic pricing literature.

Between causal inference and tail risk control. Under the structural model where the demand is linear with the price and exogenous features, we disclose that the tail risk mainly comes from the uncertainty in the estimation of the price elasticity within the experiment. This means that the bottleneck of reducing tail risk is the quality of the causal inference of the price, rather than the estimation of the influence of exogenous features. Intuitively, if the estimation of the causal effect is more accurate and closer to the real value, the distribution of the regret will also be more likely to be concentrated around its expectation, resulting in lighter tail risk. This illustrates that if we can design a strong causal estimator through the experiment, the tail risk can be further reduced. As far as we know, we are the first to reveal the relationship between the causal inference and risk control in the dynamic pricing literature.

1.2. Literature Review

Dynamic pricing with unknown demand model. There is a rich body of literature on dynamic pricing with unknown demand models, with the main objective being to maximize in-experiment revenues. This area includes works without contextual information (see, e.g., Broder & Rusmevichientong 2012; Keskin & Zeevi 2014; Besbes & Zeevi 2009; 2012; Wang et al. 2014; Miao & Wang 2021; Wang et al. 2021b) and more recently, literature on context-based dynamic pricing (e.g., Javanmard & Nazerzadeh 2019; Cohen et al. 2020; Wang et al. 2021a; Chen & Gallego 2021; Keskin et al. 2022; Miao et al. 2022; Xu & Wang 2021; Fan et al. 2021). We refer to (Ban & Keskin, 2021) for a careful review. Relevant to our work, (Qiang & Bayati, 2016) consider the same linear demand model as us, but in an incumbent-price setting. (Cheung et al., 2017) design pricing experiment to identify the real demand model from a finite function class. (Nambiar et al., 2019) study a partially linear structure demand model with model misspecification. Our design also follows the idea of the random price shock algorithm proposed by (Nambiar et al., 2019) which consider maximizing expected in-experiment revenue but do not cover two other objectives. (Bastani et al., 2022) recently propose a novel transfer learning approach to learn across different pricing experiments.

Adaptive experimental design. The benefits of experimentation for online platforms have been widely recognized by academia (e.g., Athey et al. 2018; Xiong et al. 2019; Wager & Xu 2021b; Johari et al. 2022; Farias et al. 2022b; Bojinov et al. 2022) and industry (see, e.g., Bakshy et al. 2014; Azevedo et al. 2020; Kohavi et al. 2020). Our pricing experimental design is most relevant to adaptive experimental design, which often uses the Multi-armed bandit (MAB) tool for its efficiency (Lai et al. 1985). Relevant works include those by (Wager & Xu, 2021a) on diffusion-asymptotic analysis for sequentially randomized experiments, (Adusumilli, 2021) on asymptotic Bayes and minimax risk for bandit experiments, (Kasy & Sautmann, 2021) on experiments to identify the best arm with batched data, (Dimakopoulou et al., 2021) on adaptive inference on the true mean of each arm at each step, (Offer-Westort et al., 2021) on adaptive experiments in political science, (Farias et al., 2022a) on the combination of synthetic control and MAB, (Qin & Russo, 2022) on bandit experiments with potentially nonstationary contexts, and (Simchi-Levi & Wang, 2022) on the trade-off between with-experiment rewards and after-experiment inference. There are also a large number of works that do not integrate MAB, such as (Hahn et al., 2011) on two-stage design for estimating average treatment effect, (Kato et al., 2020) on estimators constructed from dependent samples, (Glynn et al., 2020) on optimal experimental design for temporal interference, and (Bhat et al., 2020) on optimal allocation of test subjects for precision of treatment effect estimation in A/B test. These works mostly focus on discrete treatments rather than continuous treatments like price.

After-experiment inference with adaptively collected data. There is a body of literature on after-experiment inference from adaptively collected data. One area of interest is evaluating new policies using observational data that cannot be treated as i.i.d. samples (Dudík et al. 2014; Swaminathan & Joachims 2015; Li et al. 2015; Wang et al. 2017; Farajtabar et al. 2018; Zhan et al. 2021). Researchers also begin to optimize decision rules based on non-i.i.d. data (Kallus & Zhou 2018; Athey & Wager 2021; Zhou et al. 2022; Jin et al. 2022). More relevant works include (Bareinboim et al., 2015) on the unobserved confounding issue, (Hadad et al., 2021) on constructing confidence intervals for after-experiment evaluations, (Dimakopoulou et al., 2017; 2019) on estimating heterogeneous treatment effect, and (Chen et al., 2022) on a bootstrap mechanism for debiasing sample means without knowledge of the reward distribution.

Risk-averse decision making. Recently, there is a stream of works studying the risk-averse MAB problem that is related to our work (Sani et al. 2012; Galichet et al. 2013; Zimin et al. 2014; Cassel et al. 2018; Prashanth et al. 2020; Zhu & Tan 2020; Baudry et al. 2021; Khajonchotpanya et al. 2021; Chang & Tan 2022). These works refine the optimal arm based on metrics such as conditional value-at-risk and mean-variance criteria, instead of expected regret. Different from them, our work adopts the risk measure in (Fan & Glynn, 2021) and (Simchi-Levi et al., 2022). Risk has also been attracting attention in revenue management and dynamic pricing, with early works by (Feng & Xiao, 1999; Levin et al., 2008) and more recent works by (Gönsch et al., 2018; Schur et al., 2019). We refer to (Gönsch, 2017) for a survey.

Throughout this paper, we define $a \wedge b \triangleq \min\{a, b\}$ and $a \vee b \triangleq \max\{a, b\}$ for $a, b \in \mathbb{R}$, and use $[n]$ to denote the set $\{1, 2, \cdots, n\}$ for any positive integer n. Throughout this paper, the notations $\mathcal{O}(\cdot)$, $\Omega(\cdot)$ and $\Theta(\cdot)$ are used to hide constant factors, and $\widetilde{\mathcal{O}}(\cdot)$, $\widetilde{\Omega}(\cdot)$ and $\widetilde{\Theta}(\cdot)$ are used to hide both constant and logarithmic factors.

Finally, we remark that the full version of this paper (containing additional theoretical results, computational experiments, and missing proofs) is available at `https://ssrn.com/abstract=4357543`.

2. Formulation

In this paper, we study the pricing experimental design for a seller trying to figure out the price elasticity (i.e., the treatment effect of price) of new products without access to informative historical data. The expected earned revenue and the risk of suffering from huge loss throughout the experiment are also of great interest. Formally, a seller is allowed to conduct pricing experiments on an online platform over a horizon of T periods. At the beginning of each period t, the seller observes an exogenous context vector $x_t \in [0,1] \in \mathbb{R}^d$ encoding the product's characteristics and some other confounding factors in period t, e.g., economic indicator, weather, competitors' prices. We assume that $\{x_t\}_{t\geq 1}$ are independently and identically distributed (i.i.d.) random variables (r.v.'s) drawn from some unknown distribution $\mathcal{P}$. Given the context vector x_t, customer demand D_t as a function of price p is generated by the following linear structural function:

$$D_t(p) = bp + \theta^\top x_t + \varepsilon_t, \qquad (1)$$

where for simplicity, we assume the first dimension of x_t is fixed to be 1. b and θ are both unknown to the seller and are what need to be learned. $\{\varepsilon_t\}_{t\geq 1}$ are i.i.d. mean-zero σ^2-sub-Gaussian r.v.'s, i.e., $\mathbb{E}[e^{\lambda\varepsilon_t}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$ for any $\lambda \in \mathbb{R}$. The linear structural model (1) is seen as one of the most fundamental models in different fields, such as revenue management (see, e.g., Qiang & Bayati 2016) and econometrics (see, e.g., Aizer & Doyle Jr 2015). In causal inference, b can always be interpreted as the treatment effect of the continuous treatment (i.e., price in this work). b is also referred to as price elasticity in operations management. Moreover, we assume the boundedness of parameters and the positive definite information matrix.

Assumption 1. (i) $\underline{b} \leq b \leq \overline{b} < 0$ and $\theta \in [-1,1]^d$.

(ii) The minimum eigenvalue of $M := \mathbb{E}[xx^\top]$ is lower bounded by a constant c_0, i.e., $\lambda_{\min}(M) \geq c_0 > 0$.

These two assumptions are relatively mild and commonly used in current literature (see, e.g., Qiang & Bayati 2016 and Nambiar et al. 2019). We assume the seller's awareness of $\underline{b}$ and $\overline{b}$, which can be seen as some prior information based on the seller's past experience. We only ask for the existence of c_0, but not the exact value of c_0. The expected revenue under price p given context x_t is $p(bp + \theta^\top x_t)$. The optimal price for period t can be defined to be $p_t^* := \arg\max_{p\in[\underline{p},\overline{p}]} p(bp + \theta^\top x_t) = -\frac{\theta^\top x_t}{2b}$. Following the literature, e.g., (Chen & Gallego, 2021) and (Ban & Keskin, 2021), we assume that the optimal price p_t^* at each time t always falls into $[\underline{p}, \overline{p}]$.

Assumption 2. $p_t^* = -\frac{\theta^\top x_t}{2b} \in [\underline{p}, \overline{p}]$ for any $b \in [\underline{b}, \overline{b}]$ and $x_t \in [0,1]^d$.

All the instances ν under the structural equation (1) satisfying Assumptions 1 and 2 constitute an instance class denoted as $\mathcal{E}_0$. Given the context x_t, the seller charges a price p_t. After observing the demand realization D_t, the seller collects the revenue $p_t D_t$. An *admissible* pricing experimental design π is defined as a sequence of functions $\{\pi_t\}_{t\geq 1}$, where each $\pi_t(\cdot)$ maps the historical information observed up to the beginning of period t, denoted by vector $H_t = (x_1, p_1, D_1, \ldots, x_{t-1}, p_{t-1}, D_{t-1}, x_t)$, and possibly some external randomness to a feasible price $p_t \in [\underline{p}, \overline{p}]$. In addition, an admissible inference of the treatment effect b is defined as a sequence of estimators $\{\hat{b}_t\}_{t\geq 1}$, mapping H_t to $[\underline{b}, \overline{b}]$. Now, we are going to define the performance metrics for a pricing experimental design π and $\hat{b}_t$.

Causality. For any estimator $\hat{b}_t$, we adopt the expected squared error $e_\nu^2(\hat{b}_t) := \mathbb{E}[(\hat{b}_t - b)^2]$ to measure the performance of causal inference. Note that, for an unbiased estimator, $e_\nu^2(\hat{b}_t)$ is equal to its variance $\mathbb{V}(\hat{b}_t)$. If the estimator is biased, $\mathbb{V}(\hat{b}_t)$ is a strict lower bound of $e_\nu^2(\hat{b}_t)$.

Revenue. In order to measure the revenue loss of a design π, we define the empirical *regret* $R_\nu^\pi(T)$ to be the difference between the T-period revenue generated by the clairvoyant optimal policy and the revenue collected by π. That is

$$R_\nu^\pi(T) = \sum_{t=1}^{T} p_t^* D_t(p_t^*) - p_t D_t(p_t). \qquad (2)$$

Because $\mathbb{E}[\sum_{t=1}^T p_t^* D_t(p_t^*)]$ is a universal constant, our investigation of the expected revenue during experiment can be directly transformed to the conventional expected regret in online learning literature i.e., $\mathcal{R}_\nu^\pi(T) := \mathbb{E}[R_\nu^\pi(T)]$. Intuitively, $\mathcal{R}_\nu^\pi(T)$ is the expectation of the revenue loss a seller will suffer by conducting the experiment π. If one wants to maximize the total expected revenue during experiment, it is equivalent to minimizing the expected regret.

Risk. Following the definition of tail risk of (Fan & Glynn, 2021) and (Simchi-Levi et al., 2022), a policy has β-tailed risk where $\beta > 0$, if for any constant $c > 0$, there exists a constant $C > 0$ and constant k such that

$$\limsup_{T\to+\infty} \frac{\ln \sup_{\nu\in\mathcal{E}_0} \mathbb{P}(R_\nu^\pi(T) \geq cT)}{T^\beta \log(T)^k} \leq -C, \qquad (3)$$

where we allow C to be dependent on c. It is also sufficient to show that for any $c > 0$, such that

$$\sup_{\nu\in\mathcal{E}_0} \mathbb{P}(R_\nu^\pi(T) \geq cT) = \exp(-\widetilde{\Omega}(T^\beta)). \qquad (4)$$

Eq. (4) describes that the risk that the pricing experiment may suffer from a large linear regret decays exponentially fast. A larger β implies a lighter tail of the empirical regret.

Remark. For revenue maximization and risk control, both metrics are closely related to the regret. Expected regret is the commonly used metric to measure the performance of a policy π in dynamic pricing literature (see, Keskin & Zeevi 2014 and Qiang & Bayati 2016) and bandit literature (see, Slivkins 2011). β-tailed risk, on the other hand, focuses on the distribution of the empirical regret. In some situations, such as when an experiment can only run once or a few times, the expectation of the empirical regret may not be sufficient to secure the risks of the experiment. In (Simchi-Levi et al., 2022), the regret satisfying Eq. (3) or Eq. (4) is called light-tailed. Our focus here is on the value of β, rather than just whether it is light-tailed or not.

3. Between Causal Inference And Expected Revenue Management

In this section, we are going to explore the relationship between causal inference and expected revenue maximization. Specifically, we will investigate the correlation between e_ν^2 and $\mathcal{R}_\nu^\pi(T)$. First, in Section 3.1, we will establish a lower bound that statistically captures the trade-off between these two metrics. Then, in Section 3.2, we will introduce an optimal pricing experimental design that can be adjusted to meet different requirements of the trade-off between causality inference and expected revenue maximization.

3.1. A Lower Bound

The trade-off between exploration and exploitation is a central focus in online learning literature (see e.g., Lattimore & Szepesvári 2020). When considering both causal inference and expected revenue maximization, this trade-off becomes even more crucial. The goal of exploration is typically to gain more information about the underlying structural model, which is also beneficial for causal inference. However, excessive exploration can result in large regret. On the other hand, any policy with a small expected regret will sacrifice some exploration opportunities in order to exploit the current information and earn revenue, which may harm the potential for more accurate inference of price elasticity. We derive the following theorem that states this trade-off.

Theorem 1. *For any admissible design π and causal estimator $\hat{b}_T$, there always exists a hard instance $\nu \in \mathcal{E}_0$ that $e_\nu^2(\hat{b}_T)\mathcal{R}_\nu^\pi(T)$ is no less than a constant order, i.e.,*

$$\inf_{\pi, \hat{b}_T} \max_{\nu \in \mathcal{E}_0} \left[e_\nu^2(\hat{b}_T) \mathcal{R}_\nu^\pi(T) \right] = \Omega(1).$$

Theorem 1 states that for any admissible π and $\hat{b}_T$, there always exists an instance ν such that $e_\nu^2(\hat{b}_T)\mathcal{R}_\nu^\pi(T) \gtrsim T^p$ for some $p \geq 0$. Intuitively, Theorem 1 implies that it is not possible for both $e_\nu^2(\hat{b}^T)$ and $\mathcal{R}_\nu^\pi(T)$ to be small simultaneously. For any π that can ensure $\mathcal{R}_\nu^\pi(T) = \mathcal{O}(T^k)$ for all $\nu \in \mathcal{E}_0$ where $k \geq 0$, there must exist an instance $\tilde{\nu}$ such that $e_{\tilde{\nu}}^2(\hat{b}_T) \gtrsim \frac{1}{T^k}$. In other words, no one can expect a $\hat{b}_T$ that can uniformly guarantee $e_\nu^2(\hat{b}_T) < \Theta(\frac{1}{T^k})$ among all ν. Based on the results of (Nambiar et al., 2019) and (Bu et al., 2022), the worst-case optimal expected regret under the structural model (1) is $\widetilde{\Theta}(\sqrt{T})$. This means that if revenue maximization is the only objective, there exists π such that $\mathcal{R}_\nu^\pi(T) = \widetilde{\mathcal{O}}(\sqrt{T})$ for any ν, and such a $\sqrt{T}$ bound is unimprovable. Therefore, any design that can achieve the optimal expected regret can only expect $e_\nu^2(\hat{b}_T) = \Omega(\frac{1}{\sqrt{T}})$ in the worst case. In addition, the following proposition formally presents the strong statistical power of random control trails (RCTs) on inferring the price elasticity.

Proposition 1. *At any time $t \in [T]$, the seller charges $\underline{p}$ or $\overline{p}$ with equal probability independently from the observed history. Let $\hat{b}_t^{RCT} := \frac{\sum_{s=1}^t (p_s - (\overline{p}+\underline{p})/2) d_s}{\sum_{s=1}^t (p_s - (\overline{p}+\underline{p})/2)^2}$. Then, $e_\nu^2(\hat{b}_t^{RCT}) = \mathcal{O}(\frac{1}{t})$ for all $\nu \in \mathcal{E}_0$ and $t \in [T]$.*

Thus, by designing the estimator based on the data generated by RCTs, we can obtain a fast decaying speed of the estimator of price elasticity, i.e., $e_\nu^2(\hat{b}_T^{\text{RCT}}) = \mathcal{O}(\frac{1}{T})$ for all ν. From the traditional statistics (see, e.g., Wainwright 2019), the $\frac{1}{T}$ speed is almost the fastest speed one can expect. Under such a case, Theorem 1 tells that the linear expected regret of RCTs is generally unavoidable, which is saying $\mathcal{R}_\nu^{\pi^{\text{RCT}}}(T) = \Omega(T)$ in the worst case. Furthermore, the lower bound is restricting that any design $(\pi, \hat{b}_T)$ can that guarantee $e_\nu^2(\hat{b}_T) = \mathcal{O}(\frac{1}{T})$ will suffer from linear expected regret in the worst case. Thus, in this sense, it is almost impossible to find other designs that are strictly better than RCTs if the causal inference is the only objective of interest. In practice, (Fisher et al., 2018) applies RCTs in a field experiment to gauge the demand model of an online retailer.

3.2. An optimal design

In the preceding, we have shown that RCTs are optimal when causal inference is the only objective. A natural question is how to provide a flexible pricing experimental design that can be adjusted to different desired levels of trade-off between causal inference and expected revenue.

If the only objective of the pricing experiment is to maximize expected revenue, there already exist several pioneer works in dynamic pricing working on the structural model (1) (Qiang & Bayati 2016; Bu et al. 2022). (Nambiar et al., 2019) design a dynamic pricing strategy that can achieve the optimal $\widetilde{\Theta}(\sqrt{T})$ regret rate based on an elegant idea of "random shock". Our random shock design (RSD for short) presented in Algorithm 1 follows from such an idea. Specifically, in each period t, the seller first computes a greedy price p_t^g by plugging in the estimators for $\theta^\top x$ and b, since the seller wants to exploit the past information to maximize revenue. Then, it charges a price p_t by adding an independent random shock Δ_t, which takes the value of δ_t or $-\delta_t$ with equal probability, to the greedy price p_t^g. Note that Δ_t is a series of instrument variables. Thus, it is eligible to apply a two-stage least squares regression, i.e., regressing d_t directly against the random shock Δ_t, and we get an unbiased estimate of b. That is $\frac{\sum_{s=1}^t \Delta_s d_s}{\sum_{s=1}^t \Delta_s^2} = \frac{\sum_{s=1}^t \Delta_s (b p_s + \theta^\top x_s + \varepsilon_s)}{\sum_{s=1}^t \delta_s^2} = b + \frac{\sum_{s=1}^t \Delta_s (b p_s^g + \theta^\top x_s + \varepsilon_s)}{\sum_{s=1}^t \delta_s^2}$. Note

Algorithm 1 Random Shock Design (RSD)

```
1: Input: price range $[\underline{p}, \overline{p}]$, bounds on the price coefficient $\underline{b}$ and $\overline{b}$, trade-off control parameter $\alpha \in [0, \frac{1}{4}]$
2: Initialization: $\hat{b}_1^{\mathrm{RSD}} = \frac{\underline{b}+\overline{b}}{2}$, $\hat{\theta}_1 = 0$
3: for $t = 1, 2, \cdots, T$ do
4:     Set $\delta_t \leftarrow t^{-\alpha}$ and observe $x_t$;
5:     Set unconstrained greedy price: $p_t^u \leftarrow -\frac{\hat{\theta}_t x_t}{2\hat{b}_t^{\mathrm{RSD}}}$;
6:     Project greedy price: $p_t^g \leftarrow \mathrm{Proj}(p_t^u, [\underline{p}+\delta_t, \overline{p}-\delta_t])$;
7:     Generate an independent random variable $\Delta_t \leftarrow \delta_t$ w.p. $\frac{1}{2}$ and $\Delta_t \leftarrow -\delta_t$ w.p. $\frac{1}{2}$;
8:     Set price $p_t \leftarrow p_t^g + \Delta_t$;
9:     Observe realized demand $d_t$;
10:    Update and output $\hat{b}_{t+1}^{\mathrm{RSD}} \leftarrow \mathrm{Proj}(\frac{\sum_{s=1}^t \Delta_s d_s}{\sum_{s=1}^t (\Delta_s)^2}, [\underline{b}, \overline{b}])$;
11:    Set $\hat{\theta}_{t+1} \leftarrow \arg\min \sum s = 1^t (d_s - \hat{b}_{t+1}^{\mathrm{RSD}} p_s - \hat{\theta}_t x_s)^2$;
12: end for
```

that since for each $1 \leq s \leq t$, Δ_s is independent of $bp_s^g + \theta^\top x + \varepsilon_s$ and has zero mean, $\frac{\sum_{s=1}^t \Delta_s d_s}{\sum_{s=1}^t \Delta_s^2}$ is an unbiased estimate of b. The final estimator that RSD adopts has an extra projection of $\frac{\sum_{s=1}^t \Delta_s d_s}{\sum_{s=1}^t \Delta_s^2}$ to $[\underline{b}, \overline{b}]$ (line 10). Such a projection may introduce some undesirable bias to $\hat{b}_t$. However, it can reduce the variance of the estimator, and it is also beneficial for the regret analysis. If there exists some situation where the unbiasedness is crucial, the algorithm can be modified as using $\hat{b}_t^{\mathrm{RSD}} = \mathrm{Proj}(\frac{\sum_{s=1}^t \Delta_s d_s}{\sum_{s=1}^t (\Delta_s)^2}, [\underline{b}, \overline{b}])$ to calculate the unconstrained greedy price in line 5 and outputting $\frac{\sum_{s=1}^t \Delta_s d_s}{\sum_{s=1}^t (\Delta_s)^2}$ as the inference for b. Such an idea of random shock is also adopted by (Goyal & Perivier, 2021; Miao et al., 2022). Additionally, another idea one may have is to run short regression, which is to regress d_t directly on p_t. Such an idea will bring in omitted variable bias because of the correlation between x_s and p_s. The long regression (i.e., regressing d_t against p_t and x_t) also has some key challenges that need to be solved, for example, the incomplete learning issue mentioned by (Keskin & Zeevi, 2018).

Our RSD algorithm bears some similarity to the policy presented in (Nambiar et al., 2019), but with an added trade-off control parameter $\alpha \in [0, \frac{1}{4}]$. In (Nambiar et al., 2019), α is carefully chosen as $\frac{1}{4}$ to achieve the optimal $\sqrt{T}$-regret rate. However, in this section, since we are taking into account both causal inference and expected revenue maximization, we will demonstrate the different roles of α in these two tasks. Additionally, we will provide further theoretical results in the following section, despite the similarity in design. Theorem 2 illustrates the behavior of RSD on causal inference and expected revenue maximization.

Theorem 2. *With the input of $\alpha \in [0, \frac{1}{4}]$, under the linear demand model in Eq. (1), RSD satisfies*

(i) there exists a constant c_1 independent of T, such that for all $t \in [T]$ and any $\nu \in \mathcal{E}_0$, $e_\nu^2(\hat{b}_t^{RSD}) \leq c_1 t^{2\alpha-1}$;

(ii) $\mathcal{R}_\nu^{RSD}(T) = \mathcal{O}(T^{1-2\alpha})$ for any $\nu \in \mathcal{E}_0$.

The expected regret bound decreases with an increase in α, while the expected error bound increases. This shows the impact of α on the balance between the two objectives. When $\alpha = \frac{1}{4}$, Theorem 2 recovers Lemma 1 and Theorem 1 of (Nambiar et al., 2019). This implies that no other designs can strictly outperform RSD with $\alpha = \frac{1}{4}$ on the expected regret in terms of the dependence on T. If $\alpha = 0$, $e_\nu^2(\hat{b}_t^{\mathrm{RSD}}) \leq \frac{c_1}{t}$ and $\mathcal{R}_\nu^{\mathrm{RSD}}(T) = \mathcal{O}(T)$, which are similar to RCTs discussed in Proposition 1. Thus, our RSD can successfully cover the two extreme cases which have been well established in current literature. We also want to point out that if we run RSD with $\alpha \in (\frac{1}{4}, \frac{1}{2}]$, the first result in Theorem 2 still holds (i.e., $e_\nu^2(\hat{b}_t^{\mathrm{RSD}}) \leq c_1 t^{2\alpha-1}$). However, the expected regret of RSD will become $\mathcal{O}(T^{2\alpha})$ instead of $\mathcal{O}(T^{1-2\alpha})$. This indicates that when $\alpha \geq \frac{1}{4}$, increasing α will worsen both the expected regret and the expected error, and thus there is no incentive to consider $\alpha > \frac{1}{4}$. Moreover, we can have an immediate corollary.

Corollary 1. *For any $\nu \in \mathcal{E}_0$, $e_\nu^2(\hat{b}_T^{RSD})\mathcal{R}_\nu^{RSD}(T) = \mathcal{O}(1)$.*

Corollary 1 matches with the lower bound we have established in Theorem 1 up to a constant level. This indicates that our RSD is rate optimal on the metric expected error times expected regret. Thus, we derive a tight description of the relationship between causal inference and expected revenue in Corollary 2. Intuitively, $e_\nu^2(\hat{b}_T) \simeq \frac{1}{\mathcal{R}_\nu^\pi(T)}$ is roughly the smallest error one can expect from the best estimator based on a given policy π. In turn, given an estimator $\hat{b}_T$ and the desired error bound $e_\nu^2(\hat{b}_T)$, the best policy π can be expected to achieve roughly $\frac{1}{e_\nu^2(\hat{b}_T)}$ regret.

Corollary 2. $\inf_{\pi, \hat{b}_T} \max_{\nu \in \mathcal{E}_0} [e_\nu^2(\hat{b}_T)\mathcal{R}_\nu^\pi(T)] = \Theta(1)$.

4. Between Expected Revenue Management and Tail Risk Control

In this section, we study the connection between expected revenue management and risk control. We begin by analyzing the risk tail of RSD in Section 4.1. By the analysis, we uncover the relationship between causal inference and tail risk in Section 4.2. In Section 4.3, we present a lower bound demonstrating the trade-off between expected regret and tail risk. Although RSD follows some ideas from (Nambiar et al., 2019), the results in this section are completely novel.

4.1. Tail Risk of RSD

Here, we analyze the tail risk of RSD. To the best of our knowledge, this work is the first to examine the tail behavior of any dynamic pricing algorithm. The main theorem of this section states that RSD has $(1-2\alpha)$-tailed risk. Formally,

Theorem 3. *With the input parameter $\alpha \in (0, \frac{1}{4}]$, under the linear demand model in Eq. (1), RSD has $(1-2\alpha)$-tailed risk. Particularly, for any fixed $c > 0$ and $\nu \in \mathcal{E}_0$,*

$$\mathbb{P}(R_\nu^{RSD}(T) \geq cT) = \exp\left(-\Omega\left(\frac{T^{1-2\alpha}}{\log\log(T)}\right)\right). \quad (5)$$

Since all the parameters are assumed to be bounded in Assumption 1, Eq. (5) holds trivially when c is large. Eq. (5) becomes informative when c is small. Theorem 3 does not cover $\alpha = 0$. The reason is that when $\alpha = 0$, from Theorem 2, the expected regret $\mathcal{R}_\nu^{\mathrm{RSD}}(T)$ is in the order of T in the worst case. Therefore, it is unreasonable to expect the probability that the empirical regret is no smaller than order T can decay with T. Moreover, the RHS of Eq. (5) increases in terms of T as α decreases. This indicates that increasing α will reduce the tail of empirical regret when T is sufficiently large. Note that $T^{1-2\alpha}$ in the RHS of Eq. (5) is exactly the order of the expected regret. In other words, $\ln \mathbb{P}(R_\nu^{\mathrm{RSD}}(T) \geq cT)$ decays almost at the same speed as the expected regret grows. Specifically,

$$\max_{\nu \in \mathcal{E}_0} \ln \mathbb{P}(R_\nu^{\mathrm{RSD}}(T) \geq cT) \simeq -T^{1-2\alpha} \simeq -\max_{\nu \in \mathcal{E}_0} \mathcal{R}_\nu^{\mathrm{RSD}}(T),$$

which explicitly establishes a bridge between the expected revenue management and risk control. When T is sufficiently large, we ignore the $\log\log(T)$ term and constants following the traditions in online learning literature.

Remark. With the established upper bound of the expected regret in Theorem 2, a straightforward way to derive the tail bound is by simply applying Markov's inequality. This will provide us with an upper bound of $\mathcal{O}(T^{-2\alpha})$, which is much larger than the exponentially decaying one in Eq. (5).

Recently, a similar trade-off between the expected regret and the risk tail has also been observed in the traditional multi-armed bandit setting by (Simchi-Levi et al., 2023).

4.1.1. Proof sketch to Theorem 3.

In this subsection, we provide a sketched proof whose intermediate results may also be of independent interest. We identify two sources of risk: the possibility that the prices set by the seller are not optimal, and the inherent noise present in each time period. As a result, we decompose the empirical regret, $R_\nu^{\mathrm{RSD}}(T)$, into two components: $\tilde{R}_\nu^{\mathrm{RSD}}(T) := \sum_{t=1}^T (p_t^*(bp_t^* + \theta x_t) - p_t(bp_t + \theta x_t))$ and $M_T := \sum_{t=1}^T (p_t^* - p_t)\varepsilon_t$. $\tilde{R}_\nu^{\mathrm{RSD}}(T)$ is usually referred to as the pseudo empirical regret in the literature. Then, by the union bound, we can have $\mathbb{P}(R_\nu^{\mathrm{RSD}}(T) \geq cT) \leq \mathbb{P}\left(M_T \geq \frac{c}{2}T\right) + \mathbb{P}(\tilde{R}_\nu^{\mathrm{RSD}}(T) \geq \frac{c}{2}T)$, where the first term is related to the independent sub-Gaussian noise. We have the following lemma to bound the tail of the martingale M_T.

Lemma 1. *For any $\epsilon > 0$, $\mathbb{P}\left(M_T \geq \epsilon\right) \leq e^{-\frac{\epsilon^2}{2\overline{p}^2\sigma^2 T}}$.*

We then know that $\mathbb{P}\left(M_T \geq \frac{c}{2}T\right) = \exp(-\Omega(T))$. Therefore, $\mathbb{P}(\tilde{R}_\nu^{\mathrm{RSD}}(T) \geq \frac{c}{2}T)$ is the main focus of the remaining proof. The pseudo empirical regret can be further decomposed into three terms as follows.

$$\tilde{R}_\nu^{\mathrm{RSD}}(T) \leq \frac{3|b|\|\theta\|^2}{2\overline{b}^4} \sum_{t=1}^T (b - \hat{b}_t^{\mathrm{RSD}})^2 + \frac{3|b|}{2\overline{b}^2} \sum_{t=1}^T \left(\theta^\top x_t - \hat{\theta}_t^\top x_t\right)^2 + 6|b| \sum_{t=1}^T \delta_t^2, \quad (6)$$

where the first term captures the regret incurred by the inaccuracy of causal inference of the pricing effect, the second term stems from the error of estimating θ, and the third term is generated by the random shocks. Again, with the union bound, what we need to bound is the probability that each of the three terms on the RHS of Eq. (6) is no smaller than $\frac{c}{6}T$. Note that since $\sum_{t=1}^T \delta_t^2 = \Theta(T^{1-2\alpha})$, when T is sufficiently large, the probability that the third term of Eq. (6) is no less than $\frac{c}{6}T$ is zero. Therefore, we only need to control the first two terms of Eq. (6). Intuitively, by how RSD obtains $\hat{\theta}_t$, the second term of Eq. (6) is decided by how well $\hat{b}_t$ can perform and the behavior of the least square estimator. We have the following inequality,

$$\mathbb{P}\left(\frac{3|b|}{2\overline{b}^2} \sum_{t=1}^T \left(\theta^\top x_t - \hat{\theta}_t^\top x_t\right)^2 \geq \frac{cT}{6}\right) \leq \mathbb{P}(\sum_{t=1}^T (b - \hat{b}_t^{\mathrm{RSD}})^2 \geq c_{r1}T) + \exp(-\Omega(T^{1-2\alpha})), \quad (7)$$

where c_{r1} is a universal constant whose value will be specified in the proof. Eq. (7) decomposes the tail risk incurred by the estimating error of θ into the tail of inferring b and the convergence of the least square estimator. The final remaining block is how to bound $\mathbb{P}(\sum_{t=1}^T (b - \hat{b}_t^{\mathrm{RSD}})^2 \geq c_2 T)$ for some fixed $c_2 > 0$. We first have the following lemma on the tail bound of a single step estimator $\hat{b}_t^{\mathrm{RSD}}$.

Lemma 2. *For any $t \in [T]$,*

$$\mathbb{P}\left((b - \hat{b}_t^{RSD})^2 \geq \epsilon\right) \leq 2\exp(-\epsilon c_{b1} t^{1-2\alpha}), \quad (8)$$

where $c_{b1} = 1/(4(\sigma^2 + 2b^2\overline{p}^2 + 2\|\theta\|^2))$.

The last gap we need to bridge is from $\mathbb{P}((b - \hat{b}_t^{\mathrm{RSD}})^2 \geq \epsilon)$ to $\mathbb{P}(\sum_{t=1}^T (b - \hat{b}_t^{\mathrm{RSD}})^2 \geq c_2 T)$. The challenges come from the fact that $\{\hat{b}_t^{\mathrm{RSD}}\}_{t\geq 1}$ are highly correlated. Specifically, since intuitively $(b - \hat{b}_t^{\mathrm{RSD}})$ becomes smaller and smaller as t grows, how to allocate the total c_2T "budget" to each time t to make the union bound provide a desired result is challenging. We have the following Lemma 3.

Lemma 3. *For any fixed $c_2 > 0$,*

$$\mathbb{P}\left(\sum_{t=1}^{T}(b-\hat{b}_t)^2 \ge c_2 T\right) = \exp\left(-\Omega\left(\frac{T^{1-2\alpha}}{\log\log(T)}\right)\right).$$

Combining Eq. (6), Eq. (7) and Lemmas 1 and 3, we have $\mathbb{P}(\tilde{R}^{\mathrm{RSD}}_{\nu}(T) \ge \frac{c}{2}T) = \exp(-\Omega(\frac{T^{1-2\alpha}}{\log\log(T)}))$. Since $\mathbb{P}\left(M_T \ge \frac{c}{2}T\right) = \exp(-\Omega(T))$, we finish the proof. □

Remark. Lemma 2 can be used to establish the confidence interval of our estimator $\hat{b}^{\mathrm{RSD}}_t$. Also, note that Lemma 2 holds for any $t \in [T]$. This means that we can conduct an adaptive inference through the whole experimental period unlike the traditional after-experiment inference where one can only get a reliable inference at the end of the experiment.

4.2. Between Causal Inference And Tail Risk Control

Based on Eq. (6) and Eq. (7), we can figure out that the main bottleneck of the tailed risk becomes the tail behavior of the squared error of estimated b. We now can draw the following conclusion on the relationship between the causal inference and risk control under the linear structure model.

Corollary 3. *Under the linear structural model* (1)*, the tail risk of* `RSD` *is mainly caused by the uncertainty of causal inference (i.e., the error of estimating b).*

Intuitively, if $\hat{b}^{\mathrm{RSD}}_t$ can well concentrate around b, then the risk that the seller suffers from a large empirical regret can also be reduced. This can also interpret why decreasing α can reduce the risk tail (discussed in Theorem 3). When increasing α, from Theorem 2 and Lemma 2, both the squared error of $\hat{b}^{\mathrm{RSD}}_t$ (i.e., $e^2_{\nu}(\hat{b}^{\mathrm{RSD}}_t)$) and the tail of $\hat{b}^{\mathrm{RSD}}_t$ decrease. Moreover, roughly speaking,

$$\max_{\nu\in\mathcal{E}_0} \ln \mathbb{P}(R^{\mathrm{RSD}}_{\nu}(T) \ge cT) \simeq -T^{1-2\alpha} \simeq -1/e^2_{\nu}(\hat{b}^{\mathrm{RSD}}_T).$$

Note that here we just establish the quantitative relationship between $\max_{\nu\in\mathcal{E}_0} \ln \mathbb{P}(R^{\mathrm{RSD}}_{\nu}(T) \ge cT)$ and $e^2_{\nu}(\hat{b}^{\mathrm{RSD}}_T)$ for `RSD`. We do not mean that only the estimation error at the last time period T influences the tail behavior of the empirical regret. There exists complex dependence between $R^{\mathrm{RSD}}_{\nu}(T)$ and $e^2_{\nu}(\hat{b}^{\mathrm{RSD}}_T)$ through the history H_T, which allows us to establish such an explicit numerical relationship. Another important message that Corollary 3 conveys is about the different roles of b and θ in Eq. (1). Although b and θ are both coefficients of the linear model, they are heterogeneous in influencing the performances of `RSD`. The reason is that all x_t are i.i.d. generated, but the prices p_t are endogenous and adaptively selected. Therefore, estimating b can become the bottleneck of the linear structural model.

4.3. A Lower Bound

In this section, we are going to show that the $(1-2\alpha)$-tailed risk is almost the lightest tail that can be achieved by any design whose expected regret can be uniformly bounded by $\mathcal{O}(T^{1-2\alpha})$. Formally, we define the policy class $\Pi_{1-2\alpha}$ as

$$\Pi_{1-2\alpha} := \{\pi : \exists c_0 > 0, \forall \nu \in \mathcal{E}_0, \mathcal{R}^{\pi}_{\nu}(T) \le c_0 T^{1-2\alpha}\},$$

which includes all the policies π with the expected regret upper bound $\mathcal{O}(T^{1-2\alpha})$. It is straightforwards to verify that `RSD` with input α belongs to $\Pi_{1-2\alpha}$ based on Theorem 2. Our main theorem in this section is as follows.

Theorem 4. *Specify the environment class $\mathcal{E}_0$ as $p \in [\frac{3}{4}, 2]$ and $b \in [-1, -\frac{1}{6}]$. For any $\pi \in \Pi_{1-2\alpha}$, there exists a hard instance ν which makes the risk no lighter than a $(1-2\alpha)$-tailed one, i.e., for any $0 < c \le \frac{1}{768}$,*

$$\inf_{\pi\in\Pi_{1-2\alpha}} \max_{\nu\in\mathcal{E}_0} \mathbb{P}(R^{\pi}_{\nu}(T) \ge cT) = \exp(-\mathcal{O}(T^{1-2\alpha})). \quad (9)$$

Like Theorem 3, when c is large, Eq. (9) holds trivially. The main challenge arises when c is close to 0. Therefore, in Theorem 4, though $0 < c \le \frac{1}{768}$ may not be ideal, it has already revealed the essence of the problem and is sufficient for our purpose. The lower bound aligns with the upper bound we derived in Theorem 3. This allows us to safely conclude that `RSD` also possesses optimality in terms of tail risk, in addition to the established optimality in causal inference and revenue management in Section 3.2. The lower bound established in Theorem 4, as far as we know, is the first lower bound of the risk in online learning with continuous action space and in dynamic pricing literature. Together, Theorems 3 and 4 demonstrate the optimal achievable risk for $\Pi_{1-2\alpha}$ is $(1-2\alpha)$-tailed.

Corollary 4. *For any fixed $\mathcal{E}_0$, there exists a universal constant $c_3 > 0$, such that for any $0 < c \le c_3$,*

$$\inf_{\pi\in\Pi_{1-2\alpha}} \max_{\nu\in\mathcal{E}_0} \mathbb{P}(R^{\pi}_{\nu}(T) \ge cT) = \exp(-\widetilde{\Theta}(T^{1-2\alpha})).$$

Corollary 4 indicates that if a seller wishes to implement a policy with a risk tail that is strictly lighter than a $(1-2\alpha)$-tailed one, they must expand the class of policies under consideration from $\Pi_{1-2\alpha}$ to $\Pi_{1-2\alpha+\epsilon}$ for some $\epsilon > 0$. The seller must sacrifice expected regret in order to achieve a better risk tail. The trade-off between expected revenue and tail risk is further highlighted by this result.

5. Discussion

In this section, we discuss the assumptions, limitations, and future directions of this work. One important assumption is the prior knowledge of the bounds of the parameters (i.e., $\underline{b}$, $\bar{b}$, and $\theta \in [0,1]^d$) outlined in Assumption 1. This assumption is commonly made in the dynamic pricing literature (see, e.g., Keskin & Zeevi 2014; Besbes & Zeevi 2015; Miao et al. 2022), but admittedly it is a strong one. (Bijmolt et al., 2005) observe that the distribution of the elasticities is strongly peaked across a set of 1851 price elasticities based on 81 different studies. 50 percent of the observations are between -1 and -3 and 81 percent between -4 and 0. (Nambiar et al., 2019) has a similar finding that most of the b for different categories of fashion items lie in the range $[-1, -0.1]$, at Oracle Retail. Although the seller may not have informative data on the new products, the data from other products can be used to provide the bound. Another strong assumption of this work is the linear structure model. When the structure model is nonparametric with the context x, (Bu et al., 2022) have designed an algorithm which can achieve the optimal expected regret. The risk behavior may be different under a nonparametric function, as the main source of risk may come from the difficulty in identifying the nonparametric function instead of estimating b. Thus, the claim in Corollary 3 may not hold under a semi-parametric model. The linear treatment effect of price is also strong. Future work includes extending the results to other parametric forms of treatment effect and non-parametric forms.

6. Conclusion

In this paper, we statistically investigate the fundamental relationship among estimating the price elasticity, earning the revenue through the experiment, and controlling the tail risk. We first establish the trade-off between the estimating error and the expected regret, which is no pair of policy and estimator can make $\max_{\nu \in \mathcal{E}_0}[e^2_\nu(\hat{b}_T)\mathcal{R}^\pi_\nu(T)]$ strictly smaller that a constant. We also reveal the trade-off between the expected regret and the risk tail, which is the risk tail is lower bounded by an exponential term which is decreasing with the expected regret. We propose an optimal design `RSD`, which can match with both the lower bounds and are optimal. For `RSD`, under the linear structural model, the risk tail is mainly because of the uncertainty of estimating price elasticity during the experiment.

References

Adusumilli, K. Risk and optimal policies in bandit experiments. *arXiv preprint arXiv:2112.06363*, 2021.

Aizer, A. and Doyle Jr, J. J. Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *The Quarterly Journal of Economics*, 130 (2):759–803, 2015.

Athey, S. and Wager, S. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

Athey, S., Eckles, D., and Imbens, G. W. Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240, 2018.

Azevedo, E. M., Deng, A., Montiel Olea, J. L., Rao, J., and Weyl, E. G. A/b testing with fat tails. *Journal of Political Economy*, 128(12):4614–000, 2020.

Bakshy, E., Eckles, D., and Bernstein, M. S. Designing and deploying online field experiments. In *Proceedings of the 23rd international conference on World wide web*, pp. 283–292, 2014.

Ban, G.-Y. and Keskin, N. B. Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science*, 67(9): 5549–5568, 2021.

Bareinboim, E., Forney, A., and Pearl, J. Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, 28, 2015.

Bastani, H., Simchi-Levi, D., and Zhu, R. Meta dynamic pricing: Transfer learning across experiments. *Management Science*, 68(3):1865–1881, 2022.

Baudry, D., Gautron, R., Kaufmann, E., and Maillard, O. Optimal thompson sampling strategies for support-aware cvar bandits. In *International Conference on Machine Learning*, pp. 716–726. PMLR, 2021.

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

Besbes, O. and Zeevi, A. Blind network revenue management. *Operations research*, 60(6):1537–1550, 2012.

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

Bhat, N., Farias, V. F., Moallemi, C. C., and Sinha, D. Near-optimal ab testing. *Management Science*, 66(10): 4477–4495, 2020.

Bijmolt, T. H., Van Heerde, H. J., and Pieters, R. G. New empirical generalizations on the determinants of price elasticity. *Journal of marketing research*, 42(2):141–156, 2005.

Bojinov, I., Simchi-Levi, D., and Zhao, J. Design and analysis of switchback experiments. *Management Science*, 2022.

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

Bu, J., Simchi-Levi, D., and Wang, C. Context-based dynamic pricing with separable demand models. *Available at SSRN*, 2022.

Cassel, A., Mannor, S., and Zeevi, A. A general approach to multi-armed bandits under risk criteria. In *Conference On Learning Theory*, pp. 1295–1306. PMLR, 2018.

Chang, J. Q. and Tan, V. Y. A unifying theory of thompson sampling for continuous risk-averse bandits. In *Proc. of the 36th AAAI Conference on Artificial Intelligence. AAAI Press*, 2022.

Chen, N. and Gallego, G. Nonparametric pricing analytics with customer covariates. *Operations Research*, 69(3): 974–984, 2021.

Chen, N., Gao, X., and Xiong, Y. Debiasing samples from online learning using bootstrap. In *International Conference on Artificial Intelligence and Statistics*, pp. 8514–8533. PMLR, 2022.

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

Chintagunta, P. K., Bonfrer, A., and Song, I. Investigating the effects of store-brand introduction on retailer demand and pricing behavior. *Management Science*, 48(10):1242–1267, 2002.

Cohen, M. C., Lobel, I., and Paes Leme, R. Feature-based dynamic pricing. *Management Science*, 66(11):4921–4943, 2020.

Dimakopoulou, M., Zhou, Z., Athey, S., and Imbens, G. Estimation considerations in contextual bandits. *arXiv preprint arXiv:1711.07077*, 2017.

Dimakopoulou, M., Zhou, Z., Athey, S., and Imbens, G. Balanced linear contextual bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3445–3453, 2019.

Dimakopoulou, M., Ren, Z., and Zhou, Z. Online multi-armed bandits with adaptive inference. *Advances in Neural Information Processing Systems*, 34, 2021.

Dudík, M., Erhan, D., Langford, J., and Li, L. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Fan, J., Guo, Y., and Yu, M. Policy optimization using semiparametric models for dynamic pricing. *Available at SSRN 3922825*, 2021.

Fan, L. and Glynn, P. W. The fragility of optimized bandit algorithms. *arXiv preprint arXiv:2109.13595*, 2021.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pp. 1447–1456. PMLR, 2018.

Farias, V., Moallemi, C., Peng, T., and Zheng, A. Synthetically controlled bandits. *arXiv preprint arXiv:2202.07079*, 2022a.

Farias, V. F., Li, A. A., Peng, T., and Zheng, A. T. Markovian interference in experiments. *arXiv preprint arXiv:2206.02371*, 2022b.

Feng, Y. and Xiao, B. Maximizing revenues of perishable assets with a risk factor. *Operations Research*, 47(2): 337–341, 1999.

Ferreira, K. J., Lee, B. H. A., and Simchi-Levi, D. Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & service operations management*, 18(1):69–88, 2016.

Fisher, M., Gallino, S., and Li, J. Competition-based dynamic pricing in online retailing: A methodology validated with field experiments. *Management science*, 64 (6):2496–2514, 2018.

Galichet, N., Sebag, M., and Teytaud, O. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pp. 245–260. PMLR, 2013.

Glynn, P. W., Johari, R., and Rasouli, M. Adaptive experimental design with temporal interference: A maximum likelihood approach. *Advances in Neural Information Processing Systems*, 33:15054–15064, 2020.

Gönsch, J. A survey on risk-averse and robust revenue management. *European Journal of Operational Research*, 263(2):337–348, 2017.

Gönsch, J., Hassler, M., and Schur, R. Optimizing conditional value-at-risk in dynamic pricing. *OR Spectrum*, 40 (3):711–750, 2018.

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

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

Hahn, J., Hirano, K., and Karlan, D. Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics*, 29(1):96–108, 2011.

Javanmard, A. and Nazerzadeh, H. Dynamic pricing in high-dimensions. *The Journal of Machine Learning Research*, 20(1):315–363, 2019.

Jin, Y., Ren, Z., Yang, Z., and Wang, Z. Policy learning" without"overlap: Pessimism and generalized empirical bernstein's inequality. *arXiv preprint arXiv:2212.09900*, 2022.

Johari, R., Li, H., Liskovich, I., and Weintraub, G. Y. Experimental design in two-sided platforms: An analysis of bias. *Management Science*, 2022.

Kallus, N. and Zhou, A. Confounding-robust policy improvement. *Advances in neural information processing systems*, 31, 2018.

Kasy, M. and Sautmann, A. Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1): 113–132, 2021.

Kato, M., Ishihara, T., Honda, J., Narita, Y., et al. Efficient adaptive experimental design for average treatment effect estimation. *arXiv preprint arXiv:2002.05308*, 2020.

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

Keskin, N. B. and Zeevi, A. On incomplete learning and certainty-equivalence control. *Operations Research*, 66 (4):1136–1167, 2018.

Keskin, N. B., Li, Y., and Sunar, N. Data-driven clustering and feature-based retail electricity pricing with smart meters. *Available at SSRN 3686518*, 2022.

Khajonchotpanya, N., Xue, Y., and Rujeerapaiboon, N. A revised approach for risk-averse multi-armed bandits under cvar criterion. *Operations Research Letters*, 49(4): 465–472, 2021.

Kocabıyıkoğlu, A. and Popescu, I. An elasticity approach to the newsvendor with price-sensitive demand. *Operations research*, 59(2):301–312, 2011.

Kohavi, R., Tang, D., and Xu, Y. *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press, 2020.

Lai, T. L., Robbins, H., et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6 (1):4–22, 1985.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

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

Levin, Y., McGill, J., and Nediak, M. Risk in revenue management and dynamic pricing. *Operations Research*, 56(2):326–343, 2008.

Li, L., Munos, R., and Szepesvári, C. Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, pp. 608–616. PMLR, 2015.

Miao, S. and Wang, Y. Network revenue management with nonparametric demand learning:\sqrt {T}-regret and polynomial dimension dependency. *Available at SSRN 3948140*, 2021.

Miao, S., Chen, X., Chao, X., Liu, J., and Zhang, Y. Context-based dynamic pricing with online clustering. *Production and Operations Management*, 31(9):3559–3575, 2022.

Nambiar, M., Simchi-Levi, D., and Wang, H. Dynamic learning and pricing with model misspecification. *Management Science*, 65(11):4980–5000, 2019.

Offer-Westort, M., Coppock, A., and Green, D. P. Adaptive experimental design: Prospects and applications in political science. *American Journal of Political Science*, 65(4): 826–844, 2021.

Prashanth, L., Jagannathan, K., and Kolla, R. K. Concentration bounds for cvar estimation: The cases of light-tailed and heavy-tailed distributions. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5577–5586, 2020.

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

Qin, C. and Russo, D. Adaptivity and confounding in multi-armed bandit experiments. *arXiv preprint arXiv:2202.09036*, 2022.

Sani, A., Lazaric, A., and Munos, R. Risk-aversion in multi-armed bandits. *Advances in Neural Information Processing Systems*, 25, 2012.

Schur, R., Gönsch, J., and Hassler, M. Time-consistent, risk-averse dynamic pricing. *European Journal of Operational Research*, 277(2):587–603, 2019.

Scott, S. L. Multi-armed bandit experiments in the online service economy. *Applied Stochastic Models in Business and Industry*, 31(1):37–45, 2015.

Simchi-Levi, D. and Wang, C. Multi-armed bandit experimental design: Online decision-making and adaptive inference. *Available at SSRN 4224969*, 2022.

Simchi-Levi, D., Zheng, Z., and Zhu, F. A simple and optimal policy design with safety against heavy-tailed risk for multi-armed bandits. *arXiv preprint arXiv:2206.02969*, 2022.

Simchi-Levi, D., Zheng, Z., and Zhu, F. Regret distribution in stochastic bandits: Optimal trade-off between expectation and tail risk. *arXiv preprint arXiv:2304.04341*, 2023.

Slivkins, A. Contextual bandits with similarity information. In *Proceedings of the 24th annual Conference On Learning Theory*, pp. 679–702. JMLR Workshop and Conference Proceedings, 2011.

Swaminathan, A. and Joachims, T. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755, 2015.

Tellis, G. J. The price elasticity of selective demand: A meta-analysis of econometric models of sales. *Journal of marketing research*, 25(4):331–341, 1988.

Wager, S. and Xu, K. Diffusion asymptotics for sequential experiments. *arXiv preprint arXiv:2101.09855*, 2021a.

Wager, S. and Xu, K. Experimenting in equilibrium. *Management Science*, 67(11):6694–6715, 2021b.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, H., Talluri, K., and Li, X. On dynamic pricing with covariates. *arXiv preprint arXiv:2112.13254*, 2021a.

Wang, Y., Chen, B., and Simchi-Levi, D. Multimodal dynamic pricing. *Management Science*, 67(10):6136–6152, 2021b.

Wang, Y.-X., Agarwal, A., and Dudık, M. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pp. 3589–3597. PMLR, 2017.

Wang, Z., Deng, S., and Ye, Y. Close the gaps: A learning-while-doing algorithm for single-product revenue management problems. *Operations Research*, 62(2):318–331, 2014.

Xiong, R., Athey, S., Bayati, M., and Imbens, G. Optimal experimental design for staggered rollouts. *arXiv preprint arXiv:1911.03764*, 2019.

Xu, J. and Wang, Y.-X. Logarithmic regret in feature-based dynamic pricing. *Advances in Neural Information Processing Systems*, 34, 2021.

Xu, J. J., Fader, P. S., and Veeraraghavan, S. Designing and evaluating dynamic pricing policies for major league baseball tickets. *Manufacturing & Service Operations Management*, 21(1):121–138, 2019.

Zhan, R., Hadad, V., Hirshberg, D. A., and Athey, S. Off-policy evaluation via adaptive weighting with data from contextual bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2125–2135, 2021.

Zhou, Z., Athey, S., and Wager, S. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 2022.

Zhu, Q. and Tan, V. Thompson sampling algorithms for mean-variance bandits. In *International Conference on Machine Learning*, pp. 11599–11608. PMLR, 2020.

Zimin, A., Ibsen-Jensen, R., and Chatterjee, K. Generalized risk-aversion in stochastic multi-armed bandits. *arXiv preprint arXiv:1405.0833*, 2014.