HARVARD | BUSINESS | SCHOOL

Online Network Revenue Management using Thompson Sampling

Kris Johnson Ferreira
David Simchi-Levi
He Wang

Working Paper 16-031

Online Network Revenue Management using Thompson Sampling

Kris Johnson Ferreira
Harvard Business School

David Simchi-Levi
Massachusetts Institute of Technology

He Wang
Massachusetts Institute of Technology

Working Paper 16-031

Online Network Revenue Management using Thompson Sampling

Kris Johnson Ferreira[*] David Simchi-Levi[†] He Wang[‡]

April 21, 2015

Abstract

We consider a network revenue management problem where an online retailer aims to maximize revenue from multiple products with limited inventory. As common in practice, the retailer does not know the expected demand at each price and must learn the demand information from sales data. We propose an efficient and effective dynamic pricing algorithm, which builds upon the Thompson sampling algorithm used for multi-armed bandit problems by incorporating inventory constraints into the pricing decisions. Our algorithm proves to have both strong theoretical performance guarantees as well as promising numerical performance results when compared to other algorithms developed for the same setting. More broadly, our paper contributes to the literature on the multi-armed bandit problem with resource constraints, since our algorithm applies directly to this setting when the inventory constraints are interpreted as general resource constraints.

Keywords: revenue management, dynamic pricing, demand learning, multi-armed bandit, Thompson sampling

1 Introduction

1.1 Motivation and Setting

The online retail industry has experienced approximately 10% annual growth over the last 5 years in the United States, reaching nearly $300B in revenue in 2014; see industry report by Lerman (2014).[1] Businesses in this large and growing industry have additional information available to them as compared to traditional brick-and-mortar retailers. For example, online retailers have the ability to view real-time consumer purchase decisions (e.g. buy vs. no buy), whereas brick-and-mortar retailers typically do not have this capability. In addition, some business constraints that brick-and-mortar retailers face are not prevalent in the online setting. For example, many brick-and-mortar retailers find it costly and impractical to change prices of each product on a frequent basis, whereas online retailers may have the ability to change prices frequently at negligible cost. In this paper, we address how an online retailer can use such information and capabilities to make tactical pricing decisions.

Our work focuses on a revenue management problem common to many online retailers: given an initial inventory of products and finite selling season, the retailer must choose a price to maximize revenue over the course of the season. Inventory decisions are fixed prior to the selling season, and inventory cannot be replenished throughout the season. The retailer has the ability to observe consumer purchase decisions

[*]Technology and Operations Management Unit, Harvard Business School, kferreira@hbs.edu

[†]Engineering Systems Division, Department of Civil and Environmental Engineering, and the Operations Research Center, Massachusetts Institute of Technology, dslevi@mit.edu

[‡]Operations Research Center, Massachusetts Institute of Technology, wanghe@mit.edu

[1]These numbers exclude online sales of brick-and-mortar stores.

in real-time and can dynamically adjust the price at negligible cost. We refer the readers to the books by Talluri and van Ryzin (2005) and Özer and Phillips (2012) for examples of different applications of this revenue management problem. More generally, our work applies to the *network revenue management* problem, where the retailer must price several unique products, each of which may consume common resources with limited inventory.

The network revenue management problem has been well-studied in the academic literature under the additional assumption that the mean demand rate associated with each price is known to the retailer prior to the selling season (see seminal paper by Gallego and Van Ryzin (1997)). In practice, many retailers do not know the exact mean demand rates; thus, we focus on the network revenue management problem with unknown demand.

Given unknown mean demand rates, the retailer faces a tradeoff commonly referred to as the *exploration-exploitation tradeoff*. Towards the beginning of the selling season, the retailer may offer several different prices to try to learn and estimate the mean demand rate at each price ("exploration" objective). Over time, the retailer can use these mean demand rate estimates to set a price that maximizes revenue throughout the remainder of the selling season ("exploitation" objective). In our setting, the retailer is constrained by limited inventory and thus faces an additional tradeoff. Specifically, pursuing the exploration objective comes at the cost of diminishing valuable inventory. Simply put, if inventory is depleted while exploring different prices, there is no inventory left to exploit the knowledge gained.

We develop an algorithm for the network revenue management setting with unknown mean demand rates which balances the exploration-exploitation tradeoff while also incorporating inventory constraints. In the following section, we outline the academic literature that has addressed similar revenue management problems and describe how our work fits in this space. Then in Section 1.3 we provide an overview of the main contributions of our paper to this body of literature and to practice.

1.2 Literature Review

Due to the increased availability of real-time demand data, there is a fast growing literature on dynamic pricing problems with a demand learning approach. Review papers by Aviv and Vulcano (2012) and den Boer (2014) provide up-to-date surveys of this area. Our review below is focused on existing literature that considers inventory constraints. In Table 1, we list research papers that address limited inventory constraints and classify their models based on the number of products, allowable price sets being discrete or continuous, and whether the demand model is parametric or non-parametric.

As described earlier, our work generalizes to the network revenue management setting, thus allowing for multiple products, whereas much of the literature is for the single product setting. We assume the retailer chooses prices from a finite set of possible price vectors because discrete price sets are widely used in practice (Talluri and van Ryzin, 2005). Finally, we choose not to restrict ourselves to a particular parametric demand model since in practice it is hard to know the family of possible demand functions a priori.

The papers included in Table 1 propose pricing algorithms that can be roughly divided into three groups. The first group consider joint learning and pricing problem using dynamic programing (DP) (Aviv and Pazgal, 2005a,b; Bertsimas and Perakis, 2006; Araman and Caldentey, 2009; Farias and Van Roy, 2010). The resulting DP is usually intractable due to high dimensions, so heuristics are often used in these papers. Since the additional inventory constraints add to the dimension of DPs, papers in this group almost exclusively consider single product settings in order to limit the complexity of the DPs.

The second group applies a simple strategy that separates the time horizon into a demand learning phase (exploration objective) and a revenue maximization phase (exploitation objective) (Besbes and Zeevi, 2009, 2012; Chen et al., 2014). Recently, Wang et al. (2014), and Lei et al. (2014) show that when

	# products		set of prices		demand model	
	single	multiple	discrete	continuous	parametric	non-parametric
Aviv and Pazgal (2005a)	X			X	X	
Aviv and Pazgal (2005b)	X			X	X	
Bertsimas and Perakis (2006)	X			X	X	
Araman and Caldentey (2009)	X			X	X	
Besbes and Zeevi (2009)	X			X	X	X
Farias and Van Roy (2010)	X			X	X	
Besbes and Zeevi (2012)		X	X	X		X
Badanidiyuru et al. (2013)		X	X	X		X
Wang et al. (2014)	X			X		X
Lei et al. (2014)	X			X		X
Chen et al. (2014)		X		X	X	X
This paper		X	X			X

Table 1: Literature on Dynamic Learning and Pricing with Limited Inventory

there is a single product, pricing algorithms can be improved by mixing the exploration and exploitation phases. However, their methods cannot be generalized beyond the single-product/continuous-price setting.[2]

The third group of papers builds on the classical multi-armed bandit algorithms (Badanidiyuru et al. (2013) and this paper) or the stochastic approximation methods (Besbes and Zeevi, 2009; Lei et al., 2014). The multi-armed bandit problem is often used to model the exploration-exploitation tradeoff in the dynamic learning and pricing model *without* limited inventory constraints; see Gittins et al. (2011), and Bubeck and Cesa-Bianchi (2012) for an overview of this problem. Thompson sampling is a powerful algorithm used for the multi-armed bandit problem, and it is a key building block of the algorithm that we propose.

Thompson Sampling In one of the earliest papers on the multi-armed bandit problem, Thompson (1933) proposed a randomized Bayesian algorithm, which was later referred to as *Thompson sampling*. The basic idea of Thompson sampling is that at each time period, random numbers are sampled according to the posterior probability distributions of the reward of each arm, and then the arm with the highest sampled reward is chosen. (A formal description of the algorithm can be found in Appendix C). The algorithm is also known as *probability matching* since the probability of an arm being chosen matches the posterior probability that this arm has the highest expected reward. Compared to other popular multi-armed bandit algorithms such as those in the Upper Confidence Bound (UCB) family (Lai and Robbins, 1985; Auer et al., 2002a; Garivier and Cappé, 2011), Thompson sampling enjoys comparable theoretical performance guarantees (Agrawal and Goyal, 2011; Kaufmann et al., 2012; Agrawal and Goyal, 2013) and better empirical performance (Chapelle and Li, 2011). In addition, the Thompson sampling algorithm shows great flexibility and has been adapted to various model settings (Russo and Van Roy, 2014). We believe that a salient feature of this algorithm's success is its ability to update mean demand estimation in every period and then exploit this knowledge in subsequent periods. As you will see, we take advantage of this property in our development of a new algorithm for the network revenue management problem when demand distribution is unknown.

[2]Lei et al. (2014) also consider a special case of multi-product setting where there is no cross-elasticity in product demands.

1.3 Overview of Main Contributions

The main contribution of our work is the design and development of an algorithm for the network revenue management setting with unknown demand distribution which balances the exploration-exploitation tradeoff while also incorporating inventory constraints. Our algorithm is based on the Thompson sampling algorithm for the multi-armed bandit problem, where we add a linear program (LP) subroutine to incorporate inventory constraints. Recall that the original Thompson sampling algorithm *without* inventory constraints samples data from the posterior demand distribution at each time period and chooses the price that has the highest revenue under sampled data. In our algorithm, after sampling demand, the retailer solves an LP that maximizes revenue subject to inventory and time constraints and then selects a price vector according to the solution to the LP.

The proposed algorithm is easy to implement and also has strong performance guarantees. In the theoretical analysis, we consider a widely used scaling regime where inventory scales linearly with the time horizon, and we measure the algorithm's performance by *regret*, i.e., the revenue loss compared to the case where mean demand is known a priori. We show that our proposed algorithm has a regret of $O(\sqrt{T}\log T \log\log T)$, where T is the scale of both inventory and length of the time horizon.

Besbes and Zeevi (2012) and Badanidiyuru et al. (2013) consider the same model as in this paper, and propose different pricing algorithms.[3] Besbes and Zeevi (2012) design a pricing algorithm that separates demand learning and revenue maximization, which has regret $O(T^{2/3}(\log T)^{1/2})$. Badanidiyuru et al. (2013) propose two algorithms inspired by the UCB and Hedge algorithms. Using the same scaling regime in this paper, the first algorithm has regret $O(\sqrt{T}(\log T)^2)$ but is inefficient to implement. The second algorithm has regret $O(\sqrt{T}(\log T)^{1/2})$. In Section 5 we present numerical experiments which show that our algorithm has better empirical performance than Besbes and Zeevi (2012) and Badanidiyuru et al. (2013) in several examples.

More broadly, our algorithm can be viewed as a way to solve multi-armed bandit problems with resource constraints. Such problems have wide applications including dynamic pricing, dynamic online advertising, and crowdsourcing (see more examples in Badanidiyuru et al., 2013). The flexibility of Thompson sampling allows our algorithm to be generalized for this broad setting.

The remainder of this paper is outlined as follows. The next section provides a more detailed description of the network revenue management problem in an unknown demand setting. In Section 3, we propose a dynamic pricing algorithm and present variants of this algorithm as well as a generalization to other applications. Section 4 includes theoretical performance analysis of our proposed algorithm. In Section 5, we compare our algorithm with the algorithms presented in Besbes and Zeevi (2012) and Badanidiyuru et al. (2013) via simulations to illustrate our algorithm's expected performance in practice. In Section 6, we consider an extension of the model to include contextual information. Finally, we conclude in Section 7 with a few observations.

2 Dynamic Pricing Model

In this section, we describe our model in more detail and discuss some potential limitations.

2.1 Formulation

We consider a retailer who sells N products, indexed by $i = 1, \ldots, N$, over a finite selling season. These products consume M limited resources, indexed by $j = 1, \ldots, M$. Specifically, product i consumes a_{ij} units of resource j, which has I_j units of initial inventory. There is no replenishment during the selling

[3] To be precise, Besbes and Zeevi (2012) assumes a slightly different continuous-time setting with Poisson arrivals, but their results can easily be extended to the discrete-time setting in this paper.

season. The selling season is divided into T periods, and the following sequence of events occurs in each period $t = 1, \ldots, T$:

1. The retailer offers a price for each product from a finite set of admissible price vectors. We denote this set as $\{p_1, p_2, \ldots, p_K\}$, where each p_k $(\forall k = 1, \ldots, K)$ is a vector of length N specifying the price of each product. More specifically, we have $p_k = (p_{1k}, \ldots, p_{Nk})$, where p_{ik} is the price of product i, for all $i = 1, \ldots, N$. Suppose the retailer chooses price $p_{\hat{k}}$; we then denote the price vector in this period as $p(t) = p_{\hat{k}}$.

2. Customers observe the price vector chosen, $p(t)$, and make purchase decisions. For each product $i = 1, \ldots, N$, we assume that the demand of product i in each period is a Bernoulli random variable z_i.[4] (In general, we show in Section 3.1 that any bounded demand distribution can be transformed to the Bernoulli distribution setting in our proposed algorithm.) The mean of each z_i, denoted $d_i(p)$, represents the purchase probability, or "mean demand", of item i when it is offered at price p. We denote $z(t) = (z_1(t), \ldots, z_N(t))$ as the realization in period t, where $z_i(t) = 1$ if product i is purchased and $z_i(t) = 0$ otherwise.

   (a) If there is inventory available to satisfy the customers' purchase requests, then the retailer receives revenue $\sum_{i=1}^{N} z_i(t)p_i(t)$. Inventory is depleted by $\sum_{i=1}^{N} z_i(t)a_{ij}$ for each resource $j = 1, \ldots, M$.

   (b) If there is not enough inventory available to satisfy the customers' purchase decisions, the selling season ends and no future sales can be made.

We assume that the probability of each product being purchased is determined exclusively by the price vector $p(t)$. Most importantly, this exogenous purchase probability is *unknown to the retailer*. We assume that demands in different time periods are independent. However, demands for different products in the same time period can be correlated. The retailer's goal is to choose price vector $p(t)$ sequentially to maximize revenue over the course of the selling season.

Relationship to the Multi-Armed Bandit Problem Our formulation is similar to the common multi-armed bandit problem – where each price is an "arm" and revenue is the "reward" – except for two main deviations. First, our formulation allows for the network setting where multiple products consuming common resources are sold. Second, there are inventory constraints present in our setting, whereas there are no such constraints in the multi-armed bandit problem.

The presence of inventory constraints significantly complicates the problem, even for the special case of a single product. In a multi-armed bandit setting, if mean revenue of each price is known, the optimal strategy is to choose the price with the highest mean revenue. But in the presence of limited inventory, a mixed strategy that chooses multiple prices over the selling season may achieve much higher revenue than any single price strategy. Therefore, good pricing algorithms should converge not to the optimal single price, but to the optimal distribution of (possibly) multiple prices. Another challenging task is to estimate the time when inventory runs out and the selling season ends early, which is itself a random variable depending on the pricing algorithms. Such estimation is necessary for computing the retailer's expected revenue. This is opposed to classical multi-armed bandit problems for which algorithms always end at a fixed period.

In line with standard terminology in the multi-armed bandit literature, we will refer to "pulling arm k" as the retailer "offering price p_k", and "arm k" will be used interchangeably with "price p_k".

[4]Here, we have $z_i = 1$ if one unit of product i is sold and $z_i = 0$ if either there is no customer arrival or the customers did not buy product i in this period.

2.2 Model Discussion

There are a few important assumptions that we have made in formulating our problem that we would like to briefly discuss.

First, we do not consider strategic consumers who may postpone their purchase decisions in hopes to be offered a discounted price later in the selling season. However, many online retailers have the ability to tie a price to a customer identification code and thus not offer the same customer different prices at different times. Even if the retailer is not able to do this, since our algorithm does not necessarily implement a markdown pricing strategy, consumers may actually be worse off if they postpone their purchase decisions.

Second, in the model defined above, we do not incorporate competitors' pricing effects and assume the customer purchase probabilities do not change over time. In Section 6, however, we consider an extension of the model to address the issue of changing demand and competitors' pricing effects. Even without this extension, the retailer may choose to limit the set of feasible prices to a range that it feels is appropriate given competitors' pricing.

Third, the model assumes that the selling season ends when inventory of *any* resource runs out. In reality, the selling process would continue until the end of the season, and the retailer would fulfill a customer order whenever the necessary resources were available. Since the revenue gained in the latter case is always no less than the revenue gained in the former case with the same algorithm, our theoretical results in Section 4 also hold for the latter case.

The model described above is particularly suitable for the online retail setting. Since online retailers can adjust price in real time, they can choose the appropriate length of each time period in order to control the frequency of price changes. As a special case, if an online retailer is able to track purchase decisions of all incoming customers, then we can model each customer arrival as one time period. In the extension presented in Section 6, we show it is even possible to incorporate each customer's personal information into the pricing model. Although this "customized pricing" approach is still not common among many online retailers, it has been widely used in insurance and loaning, business-to-business selling, and online advertising. Nevertheless, the model described above can also be applied to brick-and-mortar retailers. In this case, demand in each period may not be Bernoulli distributed, so we consider extensions in Section 3.1 where demand either follows a general distribution with bounded support or follows a Poisson distribution.

3 Thompson Sampling Algorithm with Limited Inventory

In this section, we propose an algorithm called "Thompson Sampling with Inventory" that builds off the original Thompson sampling algorithm (Thompson, 1933) to incorporate inventory constraints. Let $N_k(t-1)$ be the number of time periods that the retailer has offered price p_k in the first $t-1$ periods, and let $W_{ik}(t-1)$ be the number of periods that product i is purchased under price p_k during these periods. Define $I_j(t-1)$ as the remaining inventory of resource j at the beginning of the t^{th} time period. Define constants $c_j = I_j/T$ for resource j, where $I_j = I_j(0)$ is the initial inventory level.

We show the complete Thompson Sampling with Inventory algorithm in Algorithm 1. Steps 1 and 4 are based on the Thompson sampling algorithm for the classical multi-armed bandit setting. In particular, in step 1, the retailer randomly samples product demands according to prior distributions of demand. In step 4, upon observing customer purchase decisions, the retailer updates the posterior distributions of product demands under the chosen price. We use Beta posterior distributions in the algorithm—a common choice in Thompson sampling—because the Beta distribution is conjugate to Bernoulli random variables. The posterior distributions of the $K-1$ unchosen price vectors are not changed.

Algorithm 1 Thompson Sampling with Inventory (TS-fixed)

Repeat the following steps for all $t = 1, ..., T$:

1. Sample Demand: For each price k and each product i, sample $\theta_{ik}(t)$ from a $Beta(W_{ik}(t-1)+1, N_k(t-1) - W_{ik}(t-1) + 1)$ distribution.
2. Optimize: Solve the following linear program, $OPT(\theta)$

$$\begin{aligned} f(\theta) = \max_{x_k} & \sum_{k=1}^{K}(\sum_{i=1}^{N} p_{ik}\theta_{ik}(t))x_k \\ \text{subject to } & \sum_{k=1}^{K}(\sum_{i=1}^{N} a_{ij}\theta_{ik}(t))x_k \le c_j \quad \text{for all } j = 1, \dots, M \\ & \sum_{k=1}^{K} x_k \le 1 \\ & x_k \ge 0, \text{ for all } k = 1, \dots, K \end{aligned}$$

   Let $(x_1(t), \dots, x_K(t))$ be the optimal solution to $OPT(\theta)$ at time t.
3. Offer Price: Retailer chooses price vector $p(t) = p_k$ with probability $x_k(t)/\sum_{k=1}^{K} x_k(t)$.
4. Update: Customer's purchase decisions, $z_i(t)$, are revealed to the retailer. The retailer sets $N_k(t) = N_k(t-1) + 1$, $W_{ik}(t) = W_{ik}(t-1) + z_i(t)$ for all $i = 1, \dots, N$, and $I_j(t) = I_j(t-1) - \sum_{i=1}^{N} z_i(t)a_{ij}$ for all $j = 1, \dots, M$.
5. Check Inventory Level: If $I_j(t) \le 0$ for any resource j, the algorithm terminates.

The algorithm differs from the ordinary Thompson sampling algorithm in steps 2 and 3. In step 2, instead of choosing the price with the highest reward using sampled demand, the retailer first solves a linear program (LP) which identifies the optimal mixed price strategy that maximizes expected revenue given sampled demand. The first constraint specifies that the average resource consumption per time period cannot exceed the initial inventory divided by length of the time horizon. The second constraint specifies that the sum of the probabilities of choosing all price vectors cannot exceed one. In step 3, the retailer randomly offers one of the K price vectors according to probabilities specified by the LP's optimal solution. Note that if all resources have positive inventory levels, we have $\sum_{k=1}^{K} x_k(t) > 0$ in the optimal solution to the LP, so the probabilities are well-defined.

In the remainder of the paper, we use TS-fixed as an abbreviation for Algorithm 1. The term "fixed" refers to the fact that the algorithm uses a fixed inventory to time ratio $c_j = I_j/T$ in the LP for every period.

3.1 Variants of the Algorithm

Inventory Updating Intuitively, improvements can be made to the TS-fixed algorithm by incorporating the real time inventory information. In particular, we can change $c_j = I_j/T$ to $c_j(t) = I_j(t-1)/(T-t+1)$ in the LP in step 2. This change is shown in Algorithm 2. We refer to this modified algorithm as the "Thompson Sampling with Inventory Rate Updating algorithm" (TS-update, for short).

In the revenue management literature, the idea of using updated inventory rates, $c_j(t)$, has been studied under the assumption that the demand distribution is known (Secomandi, 2008; Jasin and Kumar, 2012). Recently, Chen et al. (2014) consider a pricing policy using updated inventory rates in the unknown demand setting. In practice, using updated inventory rates usually improves revenue compared to using

Algorithm 2 Thompson Sampling with Inventory Rate Updating (TS-update)

Repeat the following steps for all $t = 1, ..., T$:

- Perform step 1 in Algorithm 1.
- Optimize: Solve the following linear program, $OPT(\theta)$

$$\begin{aligned} f(\theta) = \max_{x_k} & \sum_{k=1}^{K} (\sum_{i=1}^{N} p_{ik}\theta_{ik}(t))x_k \\ \text{subject to } & \sum_{k=1}^{K} (\sum_{i=1}^{N} a_{ij}\theta_{ik}(t))x_k \le c_j(t) \quad \text{for all } j = 1, \ldots, M \\ & \sum_{k=1}^{K} x_k \le 1 \\ & x_k \ge 0, \text{ for all } k = 1, \ldots, K \end{aligned}$$

- Perform steps 3–5 in Algorithm 1.

Algorithm 3 Thompson Sampling with Poisson Demand

Repeat the following steps for all $t = 1, ..., T$:

- Sample Demand: For each price k and each product i, sample $\theta_{ik}(t)$ from a $Gamma(W_{ik}(t-1) + 1, N_k(t-1) + 1)$ distribution.
- Optimize: Solve the linear program, $OPT(\theta)$, used in either Algorithm 1 or Algorithm 2.
- Perform steps 3-5 in Algorithm 1.

fixed inventory rates c_j (Talluri and van Ryzin, 2005). We also show in Section 5 that TS-update outperforms TS-fixed in simulations. Unfortunately, since TS-update involves a randomized demand sampling step, theoretical performance analysis of this algorithm appears to be much more challenging than TS-fixed and other algorithms in the existing literature.

General Demand Distributions with Bounded Support If $d_i(p_k)$ is randomly distributed with bounded support $[\underline{d}_{ik}, \bar{d}_{ik}]$, we can reduce the general distribution to a two-point distribution, and update it with Beta priors as in Algorithm 1. Suppose the retailer observes random demand $z_i(t) \in [\underline{d}_{ik}, \bar{d}_{ik}]$. We then re-sample a new random number, which equals to $\underline{d}_{ik}$ with probability $(\bar{d}_{ik} - z_i(t))/(\bar{d}_{ik} - \underline{d}_{ik})$ and equals to $\bar{d}_{ik}$ with probability $(z_i(t) - \underline{d}_{ik})/(\bar{d}_{ik} - \underline{d}_{ik})$. It is easily verifiable that the re-sampled demand has the same mean as the original demand.[5] By using re-sampling, the theoretical results in Section 4 also hold for the bounded demand setting.

As a special case, if demands have multinomial distributions for all $i = 1, \ldots, N$ and $k = 1, \ldots, K$, we can use Beta parameters similarly as in Algorithm 1 without resorting to re-sampling. This is because the Beta distribution is conjugate to the multinomial distribution.

Poisson Demand Distribution If the demand follows a Poisson distribution, we can use the Gamma distribution as the conjugate prior; see Algorithm 3. We use $Gamma(\alpha, \lambda)$ to represent a Gamma distribution with shape parameter α and rate parameter λ.

Note that Poisson demand cannot be reduced to the Bernoulli demand setting by the re-sampling method described above because Poisson demand has unbounded support. Note that Besbes and Zeevi

[5] This re-sampling trick is also mentioned by Agrawal and Goyal (2013).

(2012) assume that customers arrive according to a Poisson distribution, so when we compare our algorithm with theirs in Section 5, we use this variant of our algorithm.

3.2 Generalized Algorithm: Multi-armed Bandit with Global Constraints

The TS-fixed algorithm estimates the mean demand of each product under each price/arm, and the estimated demand is used to predict expected reward and consumption rate. We propose another algorithm that directly estimates the reward and the resource consumption rate for each arm. This algorithm applies to a very general setting of the multi-armed bandit problem with resource constraints. The problem is also named "bandits with knapsacks" in Badanidiyuru et al. (2013).

The problem is the following: there are multiple arms ($k = 1, \ldots, K$) and multiple resources ($j = 1, \ldots, M$). At each time period $t = 1, \ldots, T$, the decision maker chooses to pull one of the arms. If arm k is pulled, it consumes either 0 or $\bar{b}_{jk}$ units of resource j (for all $j = 1, \ldots, M$); in addition, the decision maker receives reward of either 0 or $\bar{r}_k$. More generally, if resource consumptions and rewards in each period are not binary but have bounded probability distributions, we can reduce the problem to the one with binary distributions using the re-sampling trick described in Section 3.1. We assume that these distributions are fixed over time, independent for different time periods, but unknown to the decision maker. At any given time, if there exists $j = 1, \ldots, M$ such that the total consumption of resource j up to this time is greater than a fixed constant I_j, the whole process stops and no future rewards will be received.

We adapt the Thompson sampling algorithm to this problem. Let $c_j = I_j/T$ for all $j = 1, \ldots, M$. Let $R_{jk}(t-1)$ be the number of periods that resource j is consumed under arm k before period t. Let $V_k(t-1)$ be the number of periods that the reward is nonzero under arm k before period t. Algorithm 4 outlines our Generalized Thompson Sampling with Global Constraints algorithm (TS-general, for short). Compared to TS-fixed, the only change is that demand information is updated at the resource level instead of the product level. We describe one application of TS-general below—the display advertising problem for online ad platforms.

Display Advertising Example Consider an online ad platform (e.g. Google) that uses the pay per click system. For each user logging on to a third-party website, Google may display a banner ad on the website. If the user clicks the ad, the advertiser sponsoring the ad pays some amount of money that is split between Google and the website hosting the banner ad (known as the *publisher*). If the user does not click the ad, no payment is made. Assuming that click rates for ads are unknown, Google faces the problem of allocating ads to publishers to maximize its revenue, while satisfying advertisers' budgets.

This problem fits into the limited resource multi-armed bandit model as follows: each arm is an ad, and each resource corresponds to an advertiser's budget. If ad k is clicked, the advertiser j sponsoring ad k pays $\bar{b}_{jk}$ units from its budget, of which Google gets a split of $\bar{r}_k$.

Note that this model is only a simplified version of the real problem. In practice, Google also obtains some data about the user and the website (e.g. the user's location or the website's keywords) and is able to use this information to improve its ad display strategy. We consider such an extension with contextual information in Section 6.

4 Theoretical Analysis

In this section, we present a theoretical analysis of the proposed algorithms. We consider a scaling regime where the initial inventory I_j increases linearly with the time horizon T for all resources $j = 1, \ldots, M$.

Algorithm 4 Generalized Thompson Sampling with Global Constraints (TS-general)

Perform the following steps for all $t = 1, ..., T$:

1. Sample data: For each arm k and each resource j, sample $\eta_{jk}(t)$ from a $Beta(R_{jk}(t-1)+1, N_k(t-1) - R_{jk}(t-1)+1)$ distribution. For each arm k, sample $\nu_k(t)$ from a $Beta(V_k(t-1)+1, N_k(t-1)-V_k(t-1)+1)$ distribution.
2. Optimize: Solve the following linear program, $OPT(\eta, \nu)$

$$\begin{aligned} f(\eta,\nu) = \max_{x_k} & \sum_{k=1}^{K} \bar{r}_k \nu_k(t) x_k \\ \text{subject to } & \sum_{k=1}^{K} \bar{b}_{jk} \eta_{jk}(t) x_k \le c_j \quad \text{for all } j = 1, \ldots, M \\ & \sum_{k=1}^{K} x_k \le 1 \\ & x_k \ge 0, \text{ for all } k = 1, \ldots, K \end{aligned}$$

   Let $(x_1(t), \ldots, x_K(t))$ be the optimal solution to the LP.
3. Pull Arm: Choose arm k with probability $x_k(t)/\sum_{k=1}^{K} x_k(t)$.
4. Update: Let $z_0(t)$ be the indicator function that revenue is positive, and $z_j(t)$ be the indicator function that resource j is consumed. Set $N_k(t) = N_k(t-1)+1$, $V_k(t) = V_k(t-1)+z_0(t)$, $R_{jk}(t) = R_{jk}(t-1)+z_j(t)$ for all $j = 1, \ldots, M$, and $I_j(t) = I_j(t-1) - \bar{b}_{jk} z_j(t)$ for all $j = 1, \ldots, M$.
5. Check Constraint Feasibility: If $I_j(t) \le 0$ for any resource j, the algorithm terminates.

Under this scaling regime, the average inventory per time period $c_j = I_j/T$ remains constant. This scaling regime is widely used in revenue management literature and also assumed by most papers reviewed in Section 1.2.

4.1 Benchmark and Linear Programming Relaxation

To evaluate the retailer's strategy, we compare the retailer's revenue with a benchmark where the true demand distribution is known a priori. We define the retailer's regret as

$$Regret(T) = E[R^*(T)] - E[R(T)],$$

where $R^*(T)$ is the optimal revenue if the demand distribution is known a priori, and $R(T)$ is the revenue when the demand distribution is unknown. In words, the regret is a non-negative quantity measuring the retailer's revenue loss due to not knowing the latent demand.

Because evaluating the expected optimal revenue with known demand requires solving a dynamic programming problem, it is difficult to compute the optimal revenue exactly even for moderate problem sizes. Gallego and Van Ryzin (1997) show that the expected optimal revenue can be approximated by the following upper bound. Let x_k be the fraction of periods that the retailer chooses price p_k for $k = 1, \ldots, K$. The upper bound is given by the following deterministic LP, denoted by OPT_{UB}:

$$\begin{aligned} f^* = \max & \sum_{k=1}^{K} (\sum_{i=1}^{N} p_{ik} d_i(p_k)) x_k \\ \text{subject to } & \sum_{k=1}^{K} (\sum_{i=1}^{N} a_{ij} d_i(p_k)) x_k \le c_j \quad \text{for all } j = 1, \ldots, M \end{aligned}$$

$$\sum_{k=1}^{K} x_k \leq 1$$
$$x_k \geq 0, \text{ for all } k = 1, \ldots, K.$$

Recall that $d_i(p_k)$ is the expected demand of product i under price p_k. Problem OPT_{UB} is almost identical to the LP used in step 2 of TS-fixed, except that it uses the true mean demand instead of sampled demand from posterior distributions. We denote the optimal value of OPT_{UB} as f^* and the optimal solution as $(x_1^*, \ldots, x_n^*)$. A well-known result (cf. Gallego and Van Ryzin, 1997) shows that

$$E[R^*(T)] \leq f^* T.$$

4.2 Analysis of Thompson Sampling with Inventory Algorithm

We now prove the following regret bound for TS-fixed and TS-general.

Theorem 1. *Suppose that the optimal solution(s) of OPT_{UB} are non-degenerate. If the demand distribution has bounded support, the regret of either* TS-fixed *or* TS-general *is bounded by*

$$Regret(T) \leq O(\sqrt{T} \log T \log \log T).$$

Proof Sketch. The complete proof of Theorem 1 can be found in Appendix A. We start with the case where there is a unique optimal solution, and the proof has three main parts. Suppose X^* is the optimal basis of OPT_{UB}. The first part of the proof shows that TS-fixed chooses arms that do not belong to X^* for no more than $O(\sqrt{T} \log T)$ times. Then in the second part, assuming there is unlimited inventory, we bound the revenue when only arms in X^* are chosen. In particular, we take advantage of the fact that TS-fixed performs continuous exploration and exploitation, so the LP solution of the algorithm converges to the optimal solution of OPT_{UB}. In the third part, we bound the expected lost sales due to having limited inventory, which should be subtracted from the revenue calculated in the second part. The case of multiple optimal solutions is proved along the same line. Finally, we prove the regret bound for TS-general by reducing it to a special case of TS-fixed. □

Note that the non-degeneracy assumption of Theorem 1 only applies to the LP with the true mean demand, OPT_{UB}. The theorem does not require that optimal solutions to $OPT(\theta)$, the LP that the retailer solves at each step, to be non-degenerate. Moreover, the simulation experiments in Section 5 show that TS-fixed performs well even if the non-degeneracy assumption does not hold.

It is useful to compare the results in Theorem 1 to the regret bounds in Besbes and Zeevi (2012) and Badanidiyuru et al. (2013), since our model settings are essentially the same. Our regret bound improves upon the $O(T^{\frac{2}{3}})$ bound proved in Besbes and Zeevi (2012) and matches (omitting log factors) the $O(\sqrt{T})$ bound in Badanidiyuru et al. (2013). We believe that the reason why our algorithm and the one proposed in Badanidiyuru et al. (2013) have stronger regret bounds is that they both perform continuous exploration and exploitation, whereas the algorithm in Besbes and Zeevi (2012) separates periods of exploration and exploitation.

However, we should note that the regret bound in Theorem 1 is *problem-dependent*, whereas the bounds in Besbes and Zeevi (2012) and Badanidiyuru et al. (2013) are *problem-independent*. More specifically, we show that TS-fixed or TS-update guarantees $Regret(T) \leq C\sqrt{T} \log T \log \log T$, where the constant C is a function of the demand data. As a result, the retailer cannot compute the constant C a priori since the mean demand is unknown. In contrast, the bounds proved in Besbes and Zeevi (2012) and Badanidiyuru et al. (2013) are independent of the demand data and only depend on parameters such as the number of price vectors or the number of resource constraints, which are known to the retailer.

Moreover, the regret bounds in Besbes and Zeevi (2012) and Badanidiyuru et al. (2013) do not require the non-degeneracy assumption.

It is well-known that the *problem-independent lower bound* for the multi-armed bandit problem is $Regret(T) \geq \Omega(\sqrt{T})$ (Auer et al., 2002b). Since the multi-armed bandit problem can be viewed as a special case of our setting where inventory is unlimited, the algorithm in Badanidiyuru et al. (2013) has the best possible problem-independent bound (omitting log factors). The $\Omega(\sqrt{T})$ lower bound is also proved separately by Besbes and Zeevi (2012) and Badanidiyuru et al. (2013).[6] On the other hand, it is not clear what the *problem-dependent lower bound* is for our setting, so we do not know for sure if our $O(\sqrt{T})$ problem-dependent bound (omitting log factors) can be improved.

5 Numerical Results

In this section, we first provide an illustration of the TS-fixed and TS-update algorithms for the setting where a single product is sold throughout the course of the selling season, and we compare these results to other proposed algorithms in the literature. Then we present results for a multi-product example; for consistency, the example we chose to use is identical to the one presented in Section 3.4 of Besbes and Zeevi (2012).

5.1 Single Product Example

Consider a retailer who sells a single product ($N = 1$) throughout a finite selling season. Without loss of generality, we can assume that the product is itself the resource ($M = 1$) which has limited inventory. The set of feasible prices is $\{\$29.90, \$34.90, \$39.90, \$44.90\}$, and the mean demand is given by $d(\$29.90) = 0.8, d(\$34.90) = 0.6, d(\$39.90) = 0.3$, and $d(\$44.90) = 0.1$. As aligned with our theoretical results, we show numerical results when inventory is scaled linearly with time, i.e. initial inventory $I = \alpha T$, for $\alpha = 0.25$ and 0.5.

We evaluate and compare the performance of the following five dynamic pricing algorithms which have been proposed for our setting:

- TS-update: intuitively, this algorithm outperforms TS-fixed and is thus what we suggest for retailers to use in practice.
- TS-fixed: this is the algorithm we have proposed with strong regret bounds as shown in Section 4.
- The algorithm proposed in Besbes and Zeevi (2012): we implemented the algorithm with $\tau = T^{2/3}$ as suggested in their paper, and we used the actual remaining inventory after the exploration phase as an input to the optimization problem which sets prices for the exploitation phase.
- The PD-BwK algorithm proposed in Badanidiyuru et al. (2013): this algorithm is based on the primal and the dual of OPT_{UB}. For each period, it estimates upper bounds of revenue, lower bounds of resource consumption, and the dual price of each resource, and then selects the arm with the highest revenue-to-resource-cost ratio.
- Thompson sampling (TS): this is the algorithm described in Thompson (1933) which has been proposed for use as a dynamic pricing algorithm but does *not* consider inventory constraints.

We measure performance as the average percent of "optimal revenue" achieved over 500 simulations. By "optimal revenue", we are referring to the upper bound on optimal revenue where the retailer knows the mean demand at each price prior to the selling season; this upper bound is the solution to OPT_{UB},

[6] Badanidiyuru et al. (2013) proves a more general lower bound where the initial inventory is not required to scale linearly with time T. However, one can show that their bound becomes $\Omega(\sqrt{T})$ under the additional assumption that inventory is linear in T.

f^*T, described in Section 4.1. Thus, the percent of optimal revenue achieved is at least as high as the numbers shown. Figure 1 shows performance results for the five algorithms outlined above.

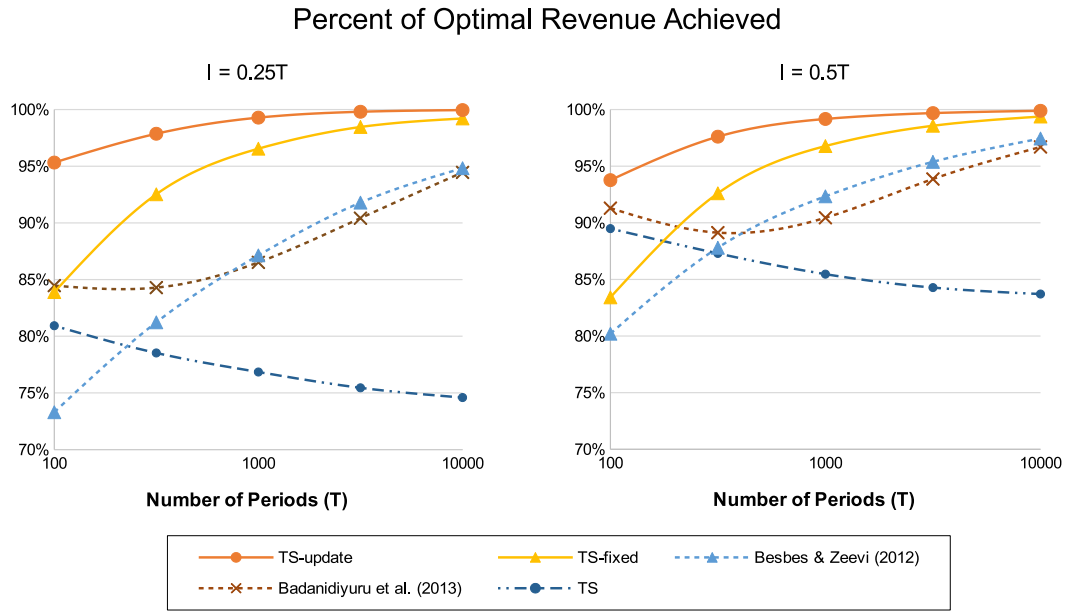

Figure 1: Performance Comparison of Dynamic Pricing Algorithms – Single Product Example

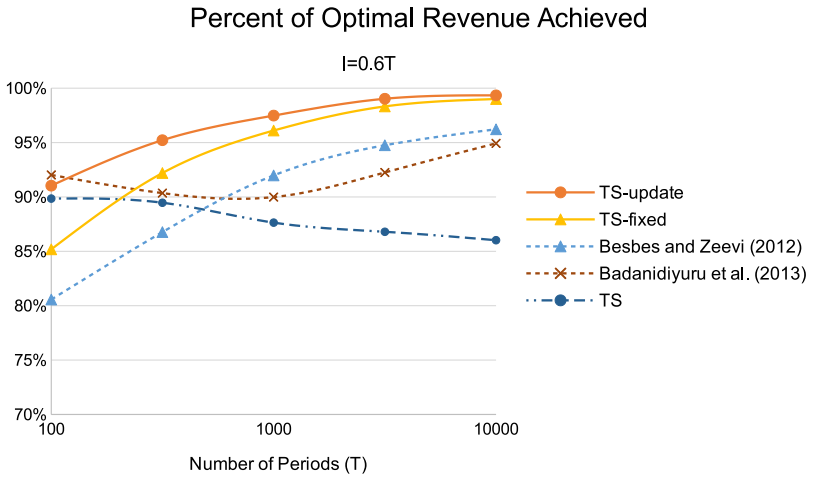

Figure 2: Performance Comparison of Dynamic Pricing Algorithms – Single Product with Degenerate Optimal Solution

The first thing to notice is that all four algorithms that incorporate inventory constraints converge to the optimal revenue as the length of the selling season increases. The TS algorithm, which does not incorporate inventory constraints, does not converge to the optimal revenue. This is because in each of the examples shown, the optimal pricing strategy is a mixed strategy where two prices are offered throughout the selling season as opposed to a single price being offered to all customers. The optimal strategy when $I = 0.25T$ is to offer the product at \$39.90 to $\frac{3}{4}$ of the customers and \$44.90 to the remaining $\frac{1}{4}$ of the customers. The optimal strategy when $I = 0.5T$ is to offer the product at \$34.90 to $\frac{2}{3}$ of the customers

and \$39.90 to the remaining $\frac{1}{3}$ of the customers. In both cases, TS converges to the suboptimal price \$29.90 offered to all the customers since this is the price that maximizes expected revenue given unlimited inventory. This really highlights the necessity of incorporating inventory constraints when developing dynamic pricing algorithms.

Overall, TS-update outperforms all of the other algorithms in both examples. Interestingly, when considering only those algorithms that incorporate inventory constraints, the gap between TS-update and the others generally shrinks when (i) the length of the selling season increases, and (ii) the ratio I/T increases. This is consistent with many other examples that we have tested and suggests that our algorithm is particularly powerful (as compared to the others) when inventory is very limited and the selling season is short. In other words, our algorithm is able to more quickly learn mean demand and identify the optimal pricing strategy, which is particularly useful for low inventory settings.

We then perform another experiment when the optimal solution to OPT_{UB} is degenerate. We assume the initial inventory is $I = 0.6T$, so the degenerate optimal solution is to offer the product at \$34.90 to all customers. Note that the degenerate case is not covered in the result of Theorem 1. Despite the lack of theoretical support, Figure 2 shows that TS-fixed and TS-update still perform well.

5.2 Multi-Product Example

Now we consider the example presented in Section 3.4 of Besbes and Zeevi (2012) where a retailer sells two products ($N = 2$) using three resources ($M = 3$). Selling one unit of product $i = 1$ consumes 1 unit of resource $j = 1$, 3 units of resource $j = 2$, and no units of resource $j = 3$. Selling one unit of product $i = 2$ consumes 1 unit of resource 1, 1 unit of resource 2, and 5 units of resource 3. The set of feasible prices is $(p_1, p_2) \in \{(1, 1.5), (1, 2), (2, 3), (4, 4), (4, 6.5)\}$. Besbes and Zeevi (2012) assume customers arrive according to a multivariate Poisson process. We would like to compare performance using a variety of potential underlying functions that specify mean demand, so we consider the following three possibilities for mean demand of each product as a function of the price vector:

1. *Linear:* $\mu(p_1, p_2) = (8 - 1.5p_1, 9 - 3p_2)$,
2. *Exponential:* $\mu(p_1, p_2) = (5e^{-0.5p_1}, 9e^{-p_2})$, and
3. *Logit:* $\mu(p_1, p_2) = \left(\frac{10e^{-p_1}}{1+e^{-p_1}+e^{-p_2}}, \frac{10e^{-p_2}}{1+e^{-p_1}+e^{-p_2}}\right)$.

Since customers arrive according to a Poisson process, we must evaluate the variant of TS-fixed and TS-update described in Section 3.1 that allows for such arrivals. Since the PD-BwK algorithm proposed in Badanidiyuru et al. (2013) does not apply to the setting where customers arrive according to a Poisson process, we cannot include this algorithm in our comparison.

We again measure performance as the average percent of "optimal revenue" achieved, where optimal revenue refers to the upper bound on optimal revenue when the retailer knows the mean demand at each price prior to the selling season, f^*T. Thus, the percent of optimal revenue achieved is at least as high as the numbers shown. Figure 3 shows average performance results over 500 simulations for each of the three underlying demand functions; we show results when inventory is scaled linearly with time, i.e. initial inventory $I = \alpha T$, for $\alpha = (3, 5, 7)$ and $\alpha = (15, 12, 30)$.

As in the single product example, each algorithm converges to the optimal revenue as the length of the selling season increases. In most cases, the TS-update algorithm outperforms the algorithm proposed in Besbes and Zeevi (2012). The TS-fixed algorithm has slightly worse performance than TS-update as expected, but in several cases the difference between the two algorithms is almost indistinguishable. For each set of parameters and when $T = 10,000$, TS-update and TS-fixed achieve 99–100% of the optimal revenue whereas the Besbes and Zeevi (2012) algorithm achieves 92–98% of the optimal revenue. As we saw in the single product example, TS-update performs particularly well when inventory is very limited ($I = (3, 5, 7)T$); it is able to more quickly learn mean demand and identify the optimal pricing strategy.

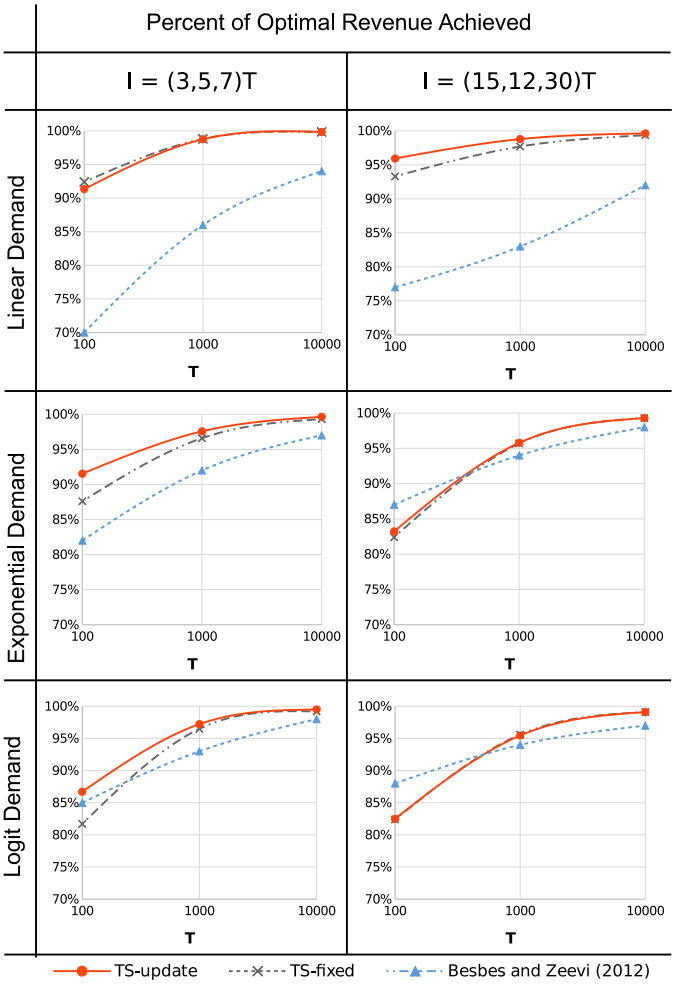

Figure 3: Performance Comparison of Dynamic Pricing Algorithms – Multi-Product Example

TS-update and TS-fixed also seem to perform particularly well when mean demand is linear. Finally, note that the algorithm's performance appears to be fairly consistent across the three demand models tested; this suggests that the retailer can confidently use our algorithm even when the underlying demand function is unknown.

6 Extension: Contextual Information

In the model described in Section 2.1, we assume that the mean demand rates are unknown to the retailer but fixed over time. However, we mentioned several factors that may cause the demand function to change over time in Section 2.2 , including seasonality, markdowns of competitors' prices, and shifts of customer preferences. In this section, we adapt the TS-fixed and TS-update algorithms to changing demands.

Several papers in the revenue management literature have studied dynamic pricing problems with unknown and changing demand (Aviv and Pazgal, 2005b; Besbes and Zeevi, 2011; den Boer, 2013; Keskin and Zeevi, 2013; Besbes and Sauré, 2014). These papers usually assume that only historical sales data can be used to detect demand changes. Unlike these papers, we assume that at each period,

the retailer receives some *contextual information* (also known as *context*, *feature*, or *side information*) that can be used to predict demand changes. For example, the contextual information may contain competitors' prices or seasonality factors predicted from previous selling seasons. As another example, in the customized pricing/ad-display problem, each period models one customer arrival, so the contextual information may include personal features of arriving customers (Chen et al., 2015). For more examples on the contextual multi-armed bandit problem, we refer the readers to Chapter 4 of Bubeck and Cesa-Bianchi (2012).

6.1 Model and Algorithm

Suppose that at the beginning of each period t, the retailer observes some context $\xi(t)$, where $\xi(t) \in \mathcal{X} \subset \mathbb{R}^d$ is a d-dimensional vector. Given context $\xi(t)$, the mean demand of product i under price p_k is a Bernoulli variable with mean $d_i(\xi(t), p_k)$, where $d_i(\cdot, p_k) : \mathcal{X} \to [0, 1]$ is a fixed function for all $i = 1, \ldots, N$ and $k = 1, \ldots, K$. Similar to the original model, we assume that function $d_i(\cdot, p_k)$ is unknown to the retailer.

For this model, we propose a modification of TS-fixed and TS-update to incorporate the contextual information (see Algorithm 5). We name the modified algorithm Thompson Sampling with Contextual Information (TS-contextual, for short). The main changes are in steps 1, 2 and 5. In step 1, given contextual information, $\xi(t)$, the retailer predicts the mean demand for each product i and price k, denoted by $h_i(\xi(t), p_k)$. Then the retailer samples demand from Beta distributions in step 2. Note that the predicted demand, $h_i(\xi(t), p_k)$, replaces the simple average of historical demand, $W_{ik}(t-1)/N_k(t-1)$, used in step 1 of TS-fixed and TS-update. Steps 3 and 4 remain the same. In step 5, the retailer observes customer purchase decisions and updates its predictions for the chosen price, p_k. The update requires a regression over all pairs of contextual information and purchase decisions in the historical data for which price p_k is offered:

$$\{(\xi(s), z_i(s)) \mid 1 \le s \le t, p(s) = p_k\}.$$

For example, if the retailer uses logistic regression, the predicted mean demand has the following form:

$$h_i(\xi, p_k) = \frac{1}{1 + e^{-\beta_0(p_k) - \beta_i^T(p_k)\xi}},$$

where parameters $\beta_0(p_k) \in \mathbb{R}, \beta_i(p_k) \in \mathbb{R}^d$, for all $i = 1, \ldots, N$ and $k = 1, \ldots, K$, are the estimated coefficients of the logistic function. The retailer has the freedom of choosing other prediction methods in TS-contextual.

6.2 Upper Bound Benchmark

In the case where the space of contextual information $\mathcal{X}$ is finite and context ξ are generated i.i.d., we can upper bound the expected optimal revenue when the demand functions and the distribution of ξ are known. The upper bound is a deterministic linear programming relaxation, similar to the one presented in Section 4.1. Suppose that $q(\xi)$ is the probability mass function of context ξ. The upper bound is given by

$$\begin{aligned} f^* = \max & \sum_{\xi \in \mathcal{X}} \sum_{k=1}^{K} (\sum_{i=1}^{N} p_{ik} d_i(\xi, p_k)) x_k(\xi) q(\xi) \\ \text{subject to } & \sum_{\xi \in \mathcal{X}} \sum_{k=1}^{K} (\sum_{i=1}^{N} a_{ij} d_i(\xi, p_k)) x_k(\xi) q(\xi) \le c_j \quad \text{for all } j = 1, \ldots, M \end{aligned}$$

Algorithm 5 Thompson Sampling with Contextual Information (TS-contextual)

Suppose the retailer starts with $h_i(\cdot, p_k)$, an estimation of the true demand function $d_i(\cdot, p_k)$, for all $i = 1, \ldots, N$ and $k = 1, \ldots, K$.
Repeat the following steps for all $t = 1, ..., T$:

1. Observe Contextual Information $\xi(t)$: Compute demand prediction $h_i(\xi(t), p_k)$ for all $i = 1, \ldots, N$ and $k = 1, \ldots, K$.
2. Sample Demand: For each price k and each product i, sample $\theta_{ik}(t)$ from a $Beta(h_i(\xi(t), p_k)N_k(t-1)+1, N_k(t) - h_i(\xi(t), p_k)N_k(t-1)+1)$ distribution.
3. Optimize: Solve the following linear program, $OPT(\theta)$

$$\begin{aligned} f(\theta) = \max_{x_k} & \sum_{k=1}^{K}(\sum_{i=1}^{N} p_{ik}\theta_{ik}(t))x_k \\ \text{subject to } & \sum_{k=1}^{K}(\sum_{i=1}^{N} a_{ij}\theta_{ik}(t))x_k \le c_j \text{ (or } c_j(t)) \quad \text{for all } j = 1, \ldots, M \\ & \sum_{k=1}^{K} x_k \le 1 \\ & x_k \ge 0, \text{ for all } k = 1, \ldots, K \end{aligned}$$

   Let $(x_1(t), \ldots, x_K(t))$ be the optimal solution to the LP.
4. Offer Price: Retailer chooses price vector $p(t) = p_k$ with probability $x_k(t)/\sum_{k=1}^{K} x_k(t)$.
5. Update: Customer purchase decisions, $z_i(t)$, are revealed to the retailer. For each product $i = 1, \ldots, N$, the retailer performs a regression (e.g. logistic regression) on the historical data:

$$\{(\xi(s), z_i(s)) \mid 1 \le s \le t, p(s) = p_k\},$$

   and updates $h_i(\cdot, p_k)$—the predicted demand function associated with product i and price p_k. The retailer also sets $N_k(t) = N_k(t-1) + 1$ and $I_j(t) = I_j(t-1) - \sum_{i=1}^{N} z_i(t)a_{ij}$ for all $j = 1, \ldots, M$.
6. Check Inventory Level: If $I_j(t+1) \le 0$ for any resource j, the algorithm terminates.

$$\begin{aligned} & \sum_{k=1}^{K} x_k(\xi) \le 1 \quad \text{for all } \xi \in \mathcal{X} \\ & x_k(\xi) \ge 0, \text{ for all } k = 1, \ldots, K, \xi \in \mathcal{X}. \end{aligned}$$

To prove the upper bound, we can view different context as different "products". In particular, we can consider a new model with $N \times |\mathcal{X}|$ "products" without contextual information, and show that the new model is equivalent to the contextual model described in Section 6.1. Suppose demand for product $(i, \xi) \in N \times |\mathcal{X}|$ has mean $d_i(\xi, p_k)q(\xi)$. Product (i, ξ) in the new model has the same feasible price set and resource consumption rate as product i in the contextual model. Then, there is an obvious equivalence between the retailer's pricing strategy in the contextual model, determined by $x_k(\xi)$, and the pricing strategy of product (i, ξ) for all $i = 1, \ldots, N$ in the new model. Due to this equivalence, the upper bound above is an immediate result of the upper bound in Section 4.1.

6.3 Numerical Example

Consider an example where the retailer sells a single product (i.e. $N = M = 1$) with initial inventory $I = 0.6T$. At each period, the retailer observes an exogenous context $\xi \in [0, 1]$ and chooses between two

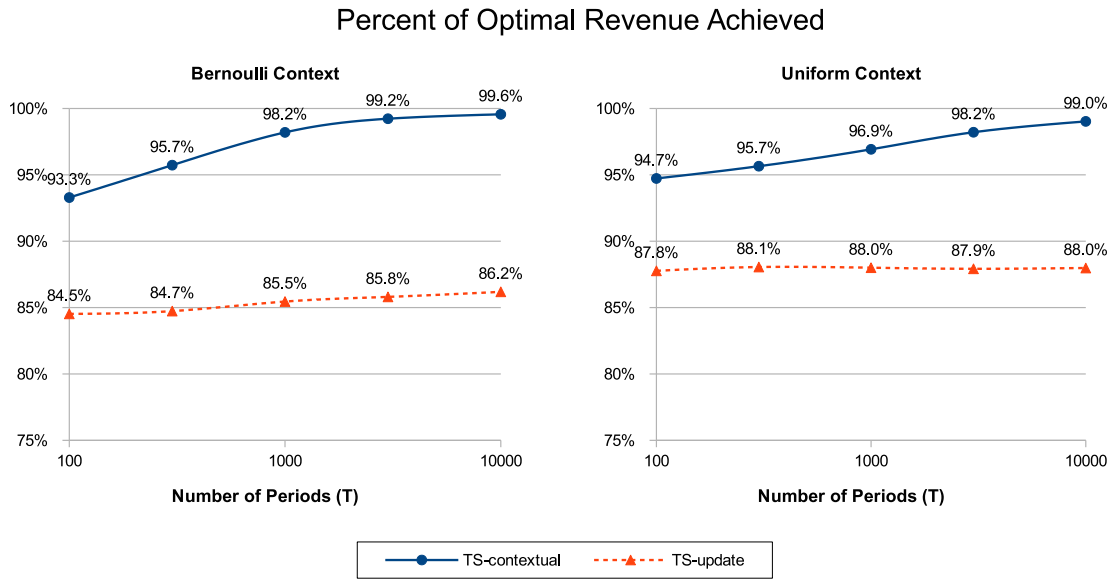

Figure 4: Performance Comparison of Dynamic Pricing Algorithms – Contextual Example

prices $\{\$9.99, \$19.99\}$. We assume that ξ represents the competitor's pricing effect normalized between 0 and 1; small ξ means the competitor offers a higher price, while larger ξ means the competitor offers a lower price. We tested two scenarios where the context ξ is either discrete or continuous: 1) ξ is generated i.i.d. from a Uniform$(0, 1)$ distribution, and 2) ξ is generated i.i.d. from a Bernoulli(0.5) distribution.

The mean demand (unknown to the retailer) as a function of ξ is

$$d(\xi, \$9.99) = 0.7e^{-0.2\xi}, \quad d(\xi, \$19.99) = 0.5e^{-1.0\xi}.$$

In particular, we assume the mean demand at the higher price ($19.99) decreases faster as ξ increases, because customer demand at a high price may be more sensitive when the competitor offers a price discount.

We compare the performance of the following two pricing algorithms:

- TS-contextual: we use logistic regression $h(\xi, \cdot) = 1/(1+e^{-\beta_0-\beta_1\xi})$ to estimate the demand function $d(\xi, \cdot)$. We also use the updated inventory rate $c_j(t)$ in the LP (similar to TS-update) instead of the fixed inventory rate c_j.
- TS-update: the retailer ignores the contextual information. Because we assume that ξ is i.i.d., ignoring the contextual information reduces the problem to the simple case where demands are i.i.d. at each price, so TS-update can be applied to this example.

Note that in both algorithms, the retailer knows neither the demand functions nor the distribution of ξ. We compare the two algorithms to the LP upper bound when the demand functions and the distribution of ξ are known. (In the case where $\xi \sim$ Uniform$(0, 1)$, the upper bound is calculated by approximating the uniform distribution with a discrete uniform distribution.) Figure 4 shows average performance results over 500 simulations for each distribution of ξ. Clearly, the revenue is significantly improved when the retailer incorporates contextual information in the pricing strategy. When $T = 10,000$, TS-contextual achieves 99% of the optimal revenue (technically, the LP upper bound) whereas TS-update achieves only 85%–88% of the optimal revenue. Over all instances, TS-contextual increases revenue by 8%–16% compared to TS-update.

Figure 4 also shows that TS-contextual converges faster to optimality when $\xi \sim$ Bernoulli(0.5) compared to when $\xi \sim$ Uniform$(0, 1)$. We suspect that the faster convergence can be explained by two reasons. First, it requires fewer data points to learn demand with Bernoulli context, since there are only two context types. Second, our algorithm incorrectly specifies the demand functions as logistic functions, while the true demand functions are exponential functions of the contextual information. Misspecification may hurt the algorithm's performance for the uniform context, but not for the Bernoulli context. In the latter case, TS-contextual only needs demand predictions for $\xi = 0$ and $\xi = 1$, so misspecification for $\xi \in (0, 1)$ does not matter. Since the retailer has freedom to choose the regression method in TS-contextual, other regression methods may be used to reduce misspecification error.

Finally, we note that even without model misspecification, the expected revenue of TS-contextual may not converge to the LP upper bound. This is because in TS-contextual, we define an LP constraint that bounds the resource consumption rate for *any* given context, while the upper bound LP bounds the resource consumption rate averaged over *all* context instances. Of course, since the constraint in TS-contextual is more conservative, the algorithm can be applied to cases where the contextual information is not i.i.d., whereas the LP upper bound requires the i.i.d. assumption.

7 Conclusion

We focus on the finite-horizon network revenue management problem in which an online retailer aims to maximize revenue from multiple products with limited inventory. As common in practice, the retailer does not know the expected demand at each price. The main contribution of our work is the development of an efficient and effective algorithm which learns mean demand and dynamically adjusts prices to maximize revenue. Our algorithm builds upon the Thompson sampling algorithm used for multi-armed bandit problems by incorporating inventory constraints into the pricing strategy. Similar to Thompson sampling, in every time period, our algorithm updates its estimation of the mean demand at the price which was offered to the customer and then uses this new information to help make all future pricing decisions.

Our algorithm proves to have both strong theoretical performance guarantees as well as promising numerical performance results when compared to other algorithms developed for the same setting. We also show the algorithm can be extended to the setting with contextual information. More broadly, our paper contributes to the literature on the multi-armed bandit problem with resource constraints, since our algorithm applies directly to this setting when the inventory constraints are interpreted as general resource constraints.

There are several directions for future work that we think would be valuable. One direction involves developing and conducting field experiments in partnership with an online retailer(s) to understand the effectiveness of our algorithm in practice. Another direction of future work is developing a non-parametric algorithm that incorporates correlations between mean demand at different prices. For example, the retailer may assume a priori that demand is decreasing in price. Although such demand correlation is allowed in the current model, our algorithm does not exploit such correlations to accelerate the demand learning process.

References

Agrawal, S. and Goyal, N. (2011). Analysis of thompson sampling for the multi-armed bandit problem. *arXiv preprint arXiv:1111.1797*.

Agrawal, S. and Goyal, N. (2013). Further optimal regret bounds for thompson sampling. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 99–107.

Araman, V. F. and Caldentey, R. (2009). Dynamic pricing for nonperishable products with demand learning. *Operations Research*, 57(5):1169–1188.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002a). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (2002b). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

Aviv, Y. and Pazgal, A. (2005a). Dynamic pricing of short life-cycle products through active learning. working paper, Olin School Business, Washington University, St. Louis, MO.

Aviv, Y. and Pazgal, A. (2005b). A partially observed markov decision process for dynamic pricing. *Management Science*, 51(9):1400–1416.

Aviv, Y. and Vulcano, G. (2012). Dynamic list pricing. In Özer, Ö. and Phillips, R., editors, *The Oxford Handbook of Pricing Management*, pages 522–584. Oxford University Press, Oxford.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. (2013). Bandits with knapsacks. In *IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 207–216. IEEE.

Bertsimas, D. and Perakis, G. (2006). Dynamic pricing: A learning approach. In Lawphongpanich, S., Hearn, D. W., and Smith, M. J., editors, *Mathematical and Computational Models for Congestion Charging*, volume 101 of *Applied Optimization*, pages 45–79. Springer US.

Besbes, O. and Sauré, D. (2014). Dynamic pricing strategies in the presence of demand shifts. *Manufacturing & Service Operations Management*, 16(4):513–528.

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

Besbes, O. and Zeevi, A. (2011). On the minimax complexity of pricing in a changing environment. *Operations Research*, 59(1):66–79.

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

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122.

Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems*, pages 2249–2257.

Chen, Q., Jasin, S., and Duenyas, I. (2014). Adaptive parametric and nonparametric multi-product pricing via self-adjusting controls. Working paper, Ross School of Business, University of Michigan.

Chen, X., Owen, Z., Pixton, C., and Simchi-Levi, D. (2015). A statistical learning approach to personalization in revenue management. *Available at SSRN: http://ssrn.com/abstract=2579462*.

den Boer, A. (2014). Dynamic pricing and learning: Historical origins, current research, and new directions. Working paper, University of Twente, Netherland.

den Boer, A. V. (2013). Tracking the market: Dynamic pricing and learning in a changing environment. Department of Applied Mathematics, University of Twente.

Farias, V. and Van Roy, B. (2010). Dynamic pricing with a prior on market response. *Operations Research*, 58(1):16–29.

Gallego, G. and Van Ryzin, G. (1994). Optimal dynamic pricing of inventories with stochastic demand over finite horizons. *Management Science*, 40(8):999–1020.

Gallego, G. and Van Ryzin, G. (1997). A multiproduct dynamic pricing problem and its applications to network yield management. *Operations Research*, 45(1):24–41.

Garivier, A. and Cappé, O. (2011). The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Conference on Learning Theory (COLT)*, pages 359–376.

Gittins, J., Glazebrook, K., and Weber, R. (2011). *Multi-armed bandit allocation indices*. John Wiley & Sons.

Jasin, S. and Kumar, S. (2012). A re-solving heuristic with bounded revenue loss for network revenue management with customer choice. *Mathematics of Operations Research*, 37(2):313–345.

Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory*, pages 199–213. Springer.

Keskin, N. B. and Zeevi, A. (2013). Chasing demand: Learning and earning in a changing environment. working paper, Booth School Business, University of Chicago, Chicago, IL.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22.

Lei, M. Y., Jasin, S., and Sinha, A. (2014). Near-optimal bisection search for nonparametric dynamic pricing with inventory constraint. Working paper, Ross School of Business, University of Michigan.

Lerman, S. (2014). E-commerce and online auctions in the us. Technical report, IBISWorld Industry Report 45411a.

Özer, Ö. and Phillips, R. (2012). *The Oxford Handbook of Pricing Management*. Oxford University Press.

Russo, D. and Van Roy, B. (2014). Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243.

Secomandi, N. (2008). An analysis of the control-algorithm re-solving issue in inventory and revenue management. *Manufacturing & Service Operations Management*, 10(3):468–483.

Talluri, K. T. and van Ryzin, G. J. (2005). *Theory and Practice of Revenue Management*. Springer-Verlag.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pages 285–294.

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

A Proofs for Section 4

We first prove the result for TS-fixed (Algorithm 1). Then, we outline the steps to reduce the setting of TS-general (Algorithm 4) to a special case of TS-fixed.

A.1 Preliminaries

To prepare for the proof of the main result, we write OPT_{UB} in the standard form. For each arm $k = 1, \ldots, K$, we denote $r_k = \sum_{i=1}^{N} p_{ik} d_i(p_k)$ as the mean revenue and $b_{jk} = \sum_{i=1}^{N} a_{ij} d_i(p_k)$ as the expected consumption of resource j. We also add $M + 1$ slack variables. Then we have

$$\begin{aligned} f^* = \max_{x,s} & \sum_{k=1}^{K} r_k x_k \\ \text{subject to } & \sum_{k=1}^{K} b_{jk} x_k + s_j = c_j \quad \text{for all } j = 1, \ldots, M \\ & \sum_{k=1}^{K} x_k + s_0 = 1 \\ & x_k \geq 0, \text{ for all } k = 1, \ldots, K \\ & s_j \geq 0, \text{ for all } j = 0, 1 \ldots, M. \end{aligned}$$

Assumption 1: Without loss of generality, we assume that $\sum_{i=1}^{N} p_{ik} \leq 1$ for all $k = 1, \ldots, K$ and $\max_{i,k}(p_{ik}/a_{ij}) \leq 1$ for all $j = 1, \ldots, M$ throughout this proof. Otherwise, we can rescale the unit of price. Similarly, we can assume $\sum_{i=1}^{N} a_{ij} \leq 1$ for all $i = 1, \ldots, N$ by rescale the units of resources. Also, we assume $c_j \leq \sum_{i=1}^{N} a_{ij}$ for all $j = 1, \ldots, M$ because $c_j > \sum_{i=1}^{N} a_{ij}$ implies that resource j is sufficient even for the maximum possible demand, so we can remove resource j from the constraints.

Assumption 2: To simplify the proof, we assume for now that OPT_{UB} has a unique optimal solution (x^*, s^*), and let $X^* = \{k \mid x_k^* > 0\}$ be the active price vectors (i.e. support) in the optimal solution and $S^* = \{j \mid s_j^* > 0\}$ be the active slack variables in the optimal solution. Let $|X^*|$ denote the cardinality of X^*. We also assume the optimal solution is non-degenerate, so $|X^*| + |S^*| = M + 1$. In the final part of the proof, we will extend the result to multiple optimal solutions.

We denote $r_k(\theta) = \sum_{i=1}^{N} p_{ik} \theta_{ik}(t)$ and $b_{jk}(\theta) = \sum_{i=1}^{N} a_{ij} \theta_{ik}(t)$ as the revenue and resource consumption under sampled demand, respectively. Then, $OPT(\theta)$ defined in TS-fixed (Algorithm 1) can be formulated equivalently as

$$\begin{aligned} f(\theta) = \max_{x,s} & \sum_{k=1}^{K} r_k(\theta) x_k \\ \text{subject to } & \sum_{k=1}^{K} b_{jk}(\theta) x_k + s_j = c_j \text{ for all } j = 1, \ldots, M \\ & \sum_{k=1}^{K} x_k + s_0 = 1 \\ & x_k \geq 0, \text{ for all } k = 1, \ldots, K \\ & s_j \geq 0, \text{ for all } j = 0, 1 \ldots, M. \end{aligned}$$

Let $X \subset \{k \mid x_1, \ldots, x_K\}$ and $S \subset \{j \mid s_0, s_1 \ldots, s_M\}$. Define constant $\Delta > 0$ as follows:

$$\Delta = \min\Big\{X^* \subsetneq X \text{ or } S^* \subsetneq S \mid \max_y f^* - \sum_{j=1}^{M} c_j y_j - y_0$$
$$\text{subject to } \sum_{j=1}^{M} b_{jk} y_j + y_0 \geq r_k \text{ for all } k \in X, y_j \geq 0 \text{ for all } j \in S\Big\}.$$

And define

$$\epsilon = \min_{j=1,\ldots,M} \frac{c_j \Delta}{8 \sum_{i=1}^{N} a_{ij}}.$$

Because OPT_{UB} has a unique and non-degenerate optimal solution, any other basic solution would be strictly suboptimal. The constant Δ is the minimum difference between the optimal value, f^*, and the value of the objective function associated with other basic solutions.

We first present a few technical lemmas for the main proof.

Lemma 1. *Consider problem $OPT(\theta)$ where the decision variables are restricted to a subset X and S such that $X^* \subsetneq X$ or $S^* \subsetneq S$. If $|\theta_{ik}(t) - d_{ik}| \leq \epsilon$ for all $i = 1, \ldots, N$ and $k \in X$, the optimal value of $OPT(\theta)$ satisfies $f(\theta) \leq f^* - \frac{3\Delta}{4}$.*

This lemma states that for any X such that $X^* \subsetneq X$ or S such that $S^* \subsetneq S$, as long as the difference between the sampled demand and the true demand is small enough, the optimal value of $OPT(\theta)$ is bounded away from f^*.

Proof. We first show that the defined constant, Δ, is indeed positive. According to the definition of (X^*, S^*), for any subset X such that $X^* \subsetneq X$ or any subset S such that $S^* \subsetneq S$, the following LP is either infeasible or has an optimal value strictly less than f^*.

$$\begin{aligned}
\max_{x,s} \ & \sum_{k \in X} r_k x_k \\
& \sum_{k \in X} b_{jk} x_k + s_j = c_j, \text{ for all } j = 1, \ldots, M \\
& \sum_{k \in X} x_k + s_0 = 1 \\
& x_k \geq 0 \text{ for all } k \in X \\
& s_j \geq 0 \text{ for all } j = 0, \ldots, M; s_j = 0 \text{ for all } j \notin S.
\end{aligned}$$

Let y_j $(j = 1, \ldots, M)$ be the dual variables associated with the first set of constraints, and y_0 be the dual variable associated with the second constraint. The dual of the LP is

$$\begin{aligned}
\min_y \ & \sum_{j=1}^{M} c_j y_j + y_0 \qquad (1) \\
\text{subject to } & \sum_{j=1}^{M} b_{jk} y_j + y_0 \geq r_k \quad \text{for all } k \in X \\
& y_j \geq 0 \text{ for all } j \in S.
\end{aligned}$$

Since the dual is feasible, its optimal value is either negative infinity or strictly less than f^* by strong duality, and thus Δ is positive.

Now consider problem $OPT(\theta)$ where the decision variables are restricted to a subset (X, S). Suppose $|\theta_{ik}(t) - d_{ik}| \leq \epsilon$ for all $i = 1, \ldots, N$ and $k \in X$. Let $\hat{x}_k$ $(k \in X)$ be the optimal solution to problem $OPT(\theta)$ with restricted (X, S), and let $\hat{y}_j$ $(j = 0, \ldots, M)$ be the optimal solution to the dual problem

(1). We have

$$
\begin{aligned}
& f^* - f(\theta) \\
=& f^* - \sum_{k\in X} r_k(\theta)\hat{x}_k && (2)\\
\geq& f^* - \sum_{k\in X} r_k(\theta)\hat{x}_k + \sum_{j=1}^{M}(\sum_{k\in S} b_{jk}(\theta)\hat{x}_k - c_j)\hat{y}_j + (\sum_{k\in X}\hat{x}_k - 1)\hat{y}_0 && (3)\\
\geq& f^* - \sum_{k\in X} r_k\hat{x}_k + \sum_{j=1}^{M}(\sum_{k\in X} b_{jk}\hat{x}_k - c_j)\hat{y}_j + (\sum_{k\in X}\hat{x}_k - 1)\hat{y}_0 - (\sum_{k\in X}\hat{x}_k\sum_{i=1}^{N} p_{ik} + \sum_{k\in X}\hat{x}_k\sum_{j=1}^{M}(\hat{y}_j\sum_{i=1}^{N} a_{ij}))\epsilon && (4)\\
\geq& f^* - \sum_{k\in X} r_k\hat{x}_k + \sum_{j=1}^{M}(\sum_{k\in X} b_{jk}\hat{x}_k - c_j)\hat{y}_j + (\sum_{k\in X}\hat{x}_k - 1)\hat{y}_0 - (\sum_{k\in X}\hat{x}_k + \sum_{k\in X}\hat{x}_k\sum_{j=1}^{M}\hat{y}_j c_j)\frac{\Delta}{8} && (5)\\
\geq& \left(f^* - \sum_{j=1}^{M} c_j\hat{y}_j - \hat{y}_0\right) + \sum_{k\in X}\hat{x}_k(\sum_{j=1}^{M} b_{jk}\hat{y}_j + \hat{y}_0 - r_k) - (1+f^*)\frac{\Delta}{8} && (6)\\
\geq& \Delta + 0 - \frac{\Delta}{4} = \frac{3\Delta}{4}. && (7)
\end{aligned}
$$

Step (3) uses the fact that $\hat{y}_j \geq 0$ and $\sum_{k\in X} b_{jk}(\theta)\hat{x}_k \leq c_j$ for all $j \in S$, and $\sum_{k\in X} b_{jk}(\theta)\hat{x}_k = c_j$ for all $j \notin S$. Step (4) uses the assumption that $|\theta_{ik}(t) - d_{ik}| \leq \epsilon$ for all $i = 1, \ldots, N$ and $k \in X$; therefore $|r_k(\theta) - r_k| \leq \sum_{i=1}^{N} p_{ik}\epsilon$ and $|b_{jk}(\theta) - b_{jk}| \leq \sum_{i=1}^{N} a_{ij}\epsilon$. In step (5), we use the definition of ϵ and the fact that $c_j \leq \sum_{i=1}^{N} a_{ij}$ and $\sum_{i=1}^{N} p_{ik} \leq 1$. Steps (6) and (7) use the fact that $\sum_{k\in X}\hat{x}_k \leq 1$ and $\sum_{j=1}^{M} c_j\hat{y}_j + \hat{y}_0 \leq f^* - \Delta \leq 1 - \Delta < 1$. □

Lemma 2. *We define $\mathcal{F}_{t-1}$ as the history prior to period t. We have*

$$E\left[\frac{\mathbb{P}(\exists i, k \in X^* : \theta_{ik}(t) < d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})}{\mathbb{P}(\forall i, k \in X^* : \theta_{ik}(t) \geq d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})} \mid N_k(t-1)\ \forall k \in X^*\right] \leq O(\frac{1}{\Delta^{N|X^*|}}). \tag{8}$$

Furthermore, if $N_k(t-1) \geq l \geq 32/\epsilon$ for all $k \in X^*$, we have

$$E\left[\frac{\mathbb{P}(\exists i, k \in X^* : \theta_{ik}(t) < d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})}{\mathbb{P}(\forall i, k \in X^* : \theta_{ik}(t) \geq d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})} \mid N_k(t-1)\ \forall k \in X^*\right] \leq O(\frac{1}{l\Delta^{N|X^*|+1}}). \tag{9}$$

Proof. Let $N_k(t-1)$ be the number of times that arm k has been played prior to period t. We have

$$
\begin{aligned}
& E\left[\frac{\mathbb{P}(\exists i, k \in X^* : \theta_{ik}(t) < d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})}{\mathbb{P}(\forall i, k \in X^* : \theta_{ik}(t) \geq d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})}\right] \\
=& E\left[E[\frac{1}{\mathbb{P}(\forall i, k \in X^* : \theta_{ik}(t) \geq d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})} \mid N_k(t-1)\ \forall k \in X^*] - 1\right] \\
=& E\left[\prod_{k\in X^*}\prod_{i=1}^{N} E[\frac{1}{\mathbb{P}(\theta_{ik}(t) \geq d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})} \mid N_k(t-1)] - 1\right] \\
\leq& E\left[\prod_{k\in X^*}\prod_{i=1}^{N}(\frac{12}{\epsilon} + 1) - 1\right] = O\left(\frac{1}{\Delta^{N|X^*|}}\right).
\end{aligned}
$$

The first equality is due to the fact that $\mathbb{P}(\exists i, k \in X^* : \theta_{ik}(t) < d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1}) = 1 - \mathbb{P}(\forall i, k \in X^* : \theta_{ik}(t) \geq d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})$ and uses the Tower rule. The second step uses the fact that conditioned on $N_k(t-1)$, the random variables $\mathbb{P}(\theta_{ik}(t) \geq d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})$ are independent. The last inequality uses Fact B.2.

Furthermore, if $N_k(t-1) \geq l \geq 32/\epsilon$ for all $k \in X^*$, we have

$$\begin{aligned}
&E\left[\frac{\mathbb{P}(\exists i, k \in X^* : \theta_{ik}(t) < d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})}{\mathbb{P}(\forall i, k \in X^* : \theta_{ik}(t) \geq d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})}\right] \\
=&E\left[\prod_{k\in X^*}\prod_{i=1}^{N} E[\frac{1}{\mathbb{P}(\theta_{ik}(t) \geq d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})} \mid N_k(t-1)] - 1\right] \\
\leq&E\left[\prod_{k\in X^*}\prod_{i=1}^{N}\left(O(\frac{1}{l\epsilon^2}) + 1\right) - 1\right] \\
=&\left(O(\frac{1}{l\epsilon^2}) + 1\right)^{N|X^*|} - 1 \leq O\left(\frac{1}{l\Delta^{N|X^*|+1}}\right).
\end{aligned}$$

The inequality again uses Fact B.2. Since $l \geq 32/\epsilon$, we have $O(1/(l\epsilon^2)) \leq O(1/(32\epsilon))$.

$$\left(O(\frac{1}{l\epsilon^2}) + 1\right)^{N|X^*|} - 1 \leq O\left(\frac{1}{l\epsilon^2}(N|X^*|)(1 + \frac{1}{32\epsilon})^{N|X^*|-1}\right) = O\left(\frac{1}{l\epsilon^{N|X^*|+1}}\right) = O\left(\frac{1}{l\Delta^{N|X^*|+1}}\right).$$

The inequality uses the fact that if $0 < x \leq c$, $(1+x)^y - 1 = \int_0^x y(1+u)^{y-1}du \leq xy(1+c)^{y-1}$. □

Lemma 3. *If $X_\epsilon(t) = X^*$ and $S(t) = S^*$, then for all $k = 1, \ldots, K$, there exists some constant $L > 0$ such that*

$$\left|\frac{x_k^*}{\sum_{k=1}^K x_k^*} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)}\right| \leq L\epsilon.$$

Proof. Define resource vector $c = (c_1, \ldots, c_M, 1)^T$. Let matrix B contain the optimal basis of OPT_{UB}, i.e., B is the coefficient matrix of OPT_{UB} whose columns are restricted to X^* and S^*. So we have $(x^*, s^*) = B^{-1}c$. Likewise, let $B(\theta)$ be the coefficient matrix with the same columns where the mean demands are replaced with sampled demands. Since $X_\epsilon(t) = X^*$, and $S(t) = S^*$, we have $(x(t), s(t))^T = B(\theta)^{-1}c$. In addition, we have $B(\theta) = B + \epsilon E$, where E is a matrix with elements in $[-1, 1]$, so

$$B(\theta)^{-1}c = (B + \epsilon E)^{-1}c = (I + \epsilon B^{-1}E)^{-1}B^{-1}c = (I - \epsilon B^{-1}E + O(\epsilon^2))B^{-1}c.$$

The last step is a result from the calculus of inverse matrix.

For $k \in X^*$, let e_k be the unit vector of length $|X^*| + |S^*| = M + 1$, whose element corresponding to x_k is one. We have

$$|x_k^* - x_k(t)| = |e_k^T(B^{-1} - B(\theta)^{-1})c| = |e_k^T(-\epsilon B^{-1}E + O(\epsilon^2))c| = O(\epsilon),$$

and thus

$$\left|\frac{x_k^*}{\sum_{k=1}^K x_k^*} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)}\right| \leq \frac{|x_k^* - x_k(t)|}{\min\{\sum_{k=1}^K x_k^*, \sum_{k=1}^K x_k(t)\}} \leq \frac{O(\epsilon)}{\sum_{k=1}^K x_k^* - K \cdot O(\epsilon)} = O(\epsilon).$$

□

Lemma 4. *Let τ be the minimum time t such that $N_k(t-1) \geq n$ for all $k \in X^*$. Then, for some $\gamma > 0$, we have*

$$E[\sum_{t=1}^{\tau} I(X_\epsilon(t) = X^*, S(t) = S^*)] \leq \frac{|X^*|}{\gamma}n.$$

Proof. Because we assume the optimal solution x^* is non-degenerate, we have $x_k^* > 0$ for all $k = 1, \ldots, K$. By Lemma 3, if $X_\epsilon(t) = X^*$ and $S(t) = S^*$, then we have $|x_k^* - x_k(t)| = O(\epsilon)$ for all k. Suppose ϵ is small enough so that we have $x_k(t) \geq \gamma > 0$ for all k. (Note that we can reduce the value of ϵ defined in

§A.1 to make this hold.) That means every arm is played with probability at least γ. So the expected time to play all arms in X^* for at least n times is bounded by $|X^*|n/\gamma$. □

A.2 Proof of Theorem 1

We now prove Theorem 1. For each time period $t = 1, \ldots, T$, we define the (possibly empty) set $X_\epsilon(t) = \{k : x_k(t) > 0, |\theta_{ik}(t) - d_{ik}| \leq \epsilon \forall i = 1, \ldots, N\}$. This set includes all arms in the optimal solution of $OPT(\theta)$ whose sampled demand is close to the true demand. Similarly, we define set $S(t) = \{j : s_j(t) > 0\}$. Recall that $N_k(T)$ is the number of times that the retailer offers price vector p_k during the selling season, and $Z_i(t)$ is the sales quantity of product i at time t (for all $i = 1, \ldots, N$, $t = 1, \ldots, T$).

From Section 4.1, we have

$$Regret(T) \leq f^*T - E[R(T)].$$

We need to provide a lower bound on the retailer's expected revenue, $E[R(T)]$, in order to complete the proof. To this end, we consider a new end period of the selling season: $T' = \lfloor (\sum_{k=1}^K x_k^*)T \rfloor$ and only count the retailer's revenue obtained from period 1 to period T', which is denoted as $E[R(T')]$. Note that if $\sum_{k=1}^K x_k^* < 1$, the optimal pricing strategy results in the retailer consuming all of a resource prior to the end of the selling horizon. Thus, we are essentially only considering revenue earned in periods prior to T', when we expect this resource to be fully consumed.

The retailer's total revenue $E[R(T)]$ is lower bounded by

$$E[R(T)] \geq E[R(T')] \geq \sum_{k \in X^*} r_k E[N_k(T')] - E[\sum_{j=1}^M \left(\max_{\substack{i=1,\ldots,N \\ k=1,\ldots,K}} \frac{p_{ik}}{a_{ij}} \right) (\sum_{i=1}^N \sum_{t=1}^{T'} a_{ij} Z_i(t) - I_j)^+]. \tag{10}$$

The first term on the right hand side, $\sum_{k \in X^*} r_k E[N_k(T')]$, is the expected revenue *without* inventory constraints when the retailer chooses arms in X^*; revenue obtained from arms not in X^* is ignored. This term can be decomposed as

$$\begin{aligned}
\sum_{k \in X^*} r_k E[N_k(T')] &= \sum_{k \in X^*} r_k E\left[\sum_{t=1}^{T'} \frac{x_k(t)}{\sum_{k=1}^K x_k(t)} \right] \\
&= \sum_{k \in X^*} r_k E\left[\sum_{t=1}^{T'} \frac{x_k^*(t)}{\sum_{k=1}^K x_k^*(t)} \right] + \sum_{k \in X^*} r_k E\left[\sum_{t=1}^{T'} \left(\frac{x_k(t)}{\sum_{k=1}^K x_k(t)} - \frac{x_k^*(t)}{\sum_{k=1}^K x_k^*(t)} \right) \right] \\
&\geq f^* \frac{T'}{\sum_{k=1}^K x_k^*(t)} + \sum_{k \in X^*} r_k E\left[\sum_{t=1}^{T'} \left(\frac{x_k(t)}{\sum_{k=1}^K x_k(t)} - \frac{x_k^*(t)}{\sum_{k=1}^K x_k^*(t)} \right) \right] \\
&\geq (f^*T - 1) - (\max_{k \in X^*} r_k) E[\sum_{t=1}^{T'} I(X_\epsilon(t) \neq X^* \text{ or } S(t) \neq S^*)] \\
&\quad + \sum_{k \in X^*} r_k E\left[\sum_{t=1}^{T'} I(X_\epsilon(t) = X^*, S(t) = S^*) \left(\frac{x_k(t)}{\sum_{k=1}^K x_k(t)} - \frac{x_k^*(t)}{\sum_{k=1}^K x_k^*(t)} \right) \right].
\end{aligned}$$

Since no sales can be made when inventory runs out, the first term in Equation (10) overestimates the retailer's actual revenue. As a result, the second term in (10), $E[\sum_{j=1}^M \max_{i,k}(p_{ik}/a_{ij}) \cdot (\sum_{i=1}^N \sum_{t=1}^{T'} a_{ij} Z_i(t) - I_j)^+]$, subtracts revenue obtained after resource consumption exceeds the inventory limit. For all $j = 1, \ldots, M$, $(\sum_{i=1}^N \sum_{t=1}^{T'} a_{ij} Z_i(t) - I_j)^+$ is the consumption of resource j that exceeds inventory budget I_j, and coefficient $\max_{i,k}(p_{ik}/a_{ij})$ is the maximum revenue that can be gained by adding one unit of resource j. Because we have assumed that $\max_{i,k}(p_{ik}/a_{ij}) \leq 1$ in Section A.1 for

all $j = 1, \ldots, M$, we can simplify (10) as:

$$\begin{aligned}
E[R(T')] \geq & \sum_{k \in X^*} r_k E[N_k(T')] - E[\sum_{j=1}^{M} (\sum_{i=1}^{N} \sum_{t=1}^{T'} a_{ij} Z_i(t) - I_j)^+] \\
\geq & (f^* T - 1) - E[\sum_{t=1}^{T'} I(X_\epsilon(t) \neq X^* \text{ or } S(t) \neq S^*)] \\
& - \sum_{k \in X^*} r_k E\left[\sum_{t=1}^{T'} I(X_\epsilon(t) = X^*, S(t) = S^*) \left(\frac{x_k^*(t)}{\sum_{k=1}^K x_k^*(t)} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)} \right) \right] \\
& - E[\sum_{j=1}^{M} (\sum_{i=1}^{N} \sum_{t=1}^{T'} a_{ij} Z_i(t) - I_j)^+]
\end{aligned}$$

Because $r_k \leq 1$ for all $k = 1, \ldots, K$, the term $E[\sum_{t=1}^{T'} I(X_\epsilon(t) \neq X^* \text{ or } S(t) \neq S^*)]$ is an upper bound of the revenue loss when $X_\epsilon(t) \neq X^*$ or $S(t) \neq S^*$ happens, which we will bound in Part I of this proof. The term $\sum_{k \in X^*} r_k E[\sum_{t=1}^{T'} I(X_\epsilon(t) = X^*, S(t) = S^*) \left(\frac{x_k^*(t)}{\sum_{k=1}^K x_k^*(t)} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)} \right)]$ is the revenue loss in case $X_\epsilon(t) = X^*$ and $S(t) = S^*$, and we consider it in Part II of the proof. The last term $E[\sum_{j=1}^{M} (\sum_{i=1}^{N} \sum_{t=1}^{T'} a_{ij} Z_i(t) - I_j)^+]$ will be bounded in Part III.

Part I: Bound the term $E[\sum_{t=1}^{T'} I(X_\epsilon(t) \neq X^* \text{ or } S(t) \neq S^*)]$.

In this part, we bound the expected number of periods where $X_\epsilon(t) \neq X^*$ or $S(t) \neq S^*$ happens.

Case I(a): First, we consider the case where $X_\epsilon(t) \neq X^*$ and $f(\theta) \geq f^* - \frac{\Delta}{2}$ (or $S(t) \neq S^*$ and $f(\theta) \geq f^* - \frac{\Delta}{2}$). This is the case where $OPT(\theta)$ produces a suboptimal basis; note that the true mean revenue is less than $f^* - \Delta$ if a suboptimal basis is chosen, so intuitively this case should not happen very often.

Since $X_\epsilon(t) \neq X^*$, by Lemma 1, we have $\sum_{k \in X_\epsilon(t)} r_k(\theta) x_k \leq f^* - 3\Delta/4$, so $\sum_{k \notin X_\epsilon(t)} r_k(\theta) x_k \geq \Delta/4$. Because $r_k(\theta) \leq 1$ for all $k = 1, \ldots, K$ and there can be at most $(M+1)$ arms with $x_k > 0$, there exists some arm $k \notin X_\epsilon(t)$ such that $x_k \geq \Delta/(4M+4)$, and thus arm k is selected with probability $x_k / \sum_{k=1}^K x_k \geq x_k \geq \Delta/(4M+4)$. Therefore, $\Delta/(4M+4)$ is the minimum probability that an arm not in $X_\epsilon(t)$ is played in Case 1(a).

By Fact B.1, if arm k has been pulled $N_k(t-1) \geq 2\log(NT)/\epsilon^2$ times (for any $\epsilon > 0$), the probability that $|\theta_{ik}(t) - d_{ik}| \geq \epsilon$ for some $i = 1, \ldots, N$ is bounded by $4/T$. For each period $t = 1, \ldots, T$, we define the event $A_t = \{\text{for all } k \text{ such that } N_k(t-1) \geq 2\log(NT)/\epsilon^2 : |\theta_{ik}(t) - d_{ik}| \leq \epsilon \; \forall i = 1, \ldots, N\}$, and we have $\mathbb{P}(A_t^c) \leq 4K/T$. So

$$\begin{aligned}
& \sum_{t=1}^{T} E\left[\mathbb{P}(X_\epsilon(t) \neq X^*, f(\theta) \geq f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1}) \right] \\
= & \sum_{t=1}^{T} E\left[\mathbb{P}(X_\epsilon(t) \neq X^*, f(\theta) \geq f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1}) \mid A_t \right] \mathbb{P}(A_t) \\
& + \sum_{t=1}^{T} E\left[\mathbb{P}(X_\epsilon(t) \neq X^*, f(\theta) \geq f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1}) \mid A_t^c \right] \mathbb{P}(A_t^c) \\
\leq & \sum_{t=1}^{T} E\left[\mathbb{P}(X_\epsilon(t) \neq X^*, f(\theta) \geq f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1}) \mid A_t \right] \\
& + \sum_{t=1}^{T} \mathbb{P}(A_t^c) \leq \frac{4M+4}{\Delta} \cdot \frac{K \log NT}{\epsilon^2} + \sum_{t=1}^{T} \frac{4K}{T} = O(\frac{MK \log NT}{\Delta^3}).
\end{aligned}$$

The term $\frac{4M+4}{\Delta} \cdot \frac{K \log NT}{\epsilon^2}$ is the expected stopping time that all k arms have been played for $2\log(NT)/\epsilon^2$ times, in which case the event $\{X_\epsilon(t) \neq X^*, f(\theta) \geq f^* - \frac{\Delta}{2}\}$ cannot happen anymore under A_t.

The case $S(t) \neq S^*, f(\theta) \geq f^* - \frac{\Delta}{2}$ can be bounded in the same way so we omit the proof.

Case I(b): We then consider the case $f(\theta) < f^* - \frac{\Delta}{2}$ and therefore either $X_\epsilon(t) \neq X^*$ or $S(t) \neq S^*$. We will bound the total number of times that this case can happen:

$$\sum_{t=1}^{T} E\left[\mathbb{P}(f(\theta) < f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1})\right].$$

Recall that X^* is the support of the optimal solution of the deterministic upper bound OPT_{UB}. If $\theta_{ik}(t) \geq d_{ik} - \epsilon/4$ for all $i = 1, \ldots, N$ and $k \in X^*$, it is easily verifiable that $f(\theta) \geq f^* - \Delta/2$. So

$$\begin{aligned}
&\mathbb{P}(\forall i, k \in X^* : \theta_{ik}(t) \geq d_{ik} - \frac{\epsilon}{4} \mid \mathcal{F}_{t-1}) \leq \mathbb{P}(f(\theta) \geq f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1}) \\
=&\mathbb{P}(X_\epsilon(t) = X^*, S(t) = S^*, f(\theta) \geq f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1}) + \mathbb{P}(X_\epsilon(t) \neq X^* \text{ or } S(t) \neq S^*, f(\theta) \geq f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1}) \\
=&\mathbb{P}(X_\epsilon(t) = X^*, S(t) = S^* \mid \mathcal{F}_{t-1}) + \mathbb{P}(X_\epsilon(t) \neq X^* \text{ or } S(t) \neq S^*, f(\theta) \geq f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1}). \qquad (11)
\end{aligned}$$

In addition, we have

$$\mathbb{P}(f(\theta) < f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1}) \leq \mathbb{P}(\exists i, k \in X^* : \theta_{ik}(t) < d_{ik} - \frac{\epsilon}{4} \mid \mathcal{F}_{t-1}). \qquad (12)$$

Combining (12) with (11), we have

$$\begin{aligned}
&\mathbb{P}(f(\theta) < f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1}) \\
\leq&\mathbb{P}(\exists i, k \in X^* : \theta_{ik}(t) < d_{ik} - \frac{\epsilon}{4} \mid \mathcal{F}_{t-1}) \\
&\cdot \frac{\mathbb{P}(X_\epsilon(t) = X^*, S(t) = S^* \mid \mathcal{F}_{t-1}) + \mathbb{P}(X_\epsilon(t) \neq X^* \text{ or } S(t) \neq S^*, f(\theta) \geq f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1})}{\mathbb{P}(\forall i, k \in X^* : \theta_{ik}(t) \geq d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})} \\
=&\frac{\mathbb{P}(\exists i, k \in X^* : \theta_{ik}(t) < d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})}{\mathbb{P}(\forall i, k \in X^* : \theta_{ik}(t) \geq d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})} \mathbb{P}(X_\epsilon(t) = X^*, S(t) = S^* \mid \mathcal{F}_{t-1}) \\
&+ \frac{\mathbb{P}(\exists i, k \in X^* : \theta_{ik}(t) < d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})}{\mathbb{P}(\forall i, k \in X^* : \theta_{ik}(t) \geq d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})} \mathbb{P}(X_\epsilon(t) \neq X^* \text{ or } S(t) \neq S^*, f(\theta) \geq f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1}) \\
\triangleq&(\Phi_t) + (\Psi_t)
\end{aligned}$$

Since we find it is more convenient to consider $\sqrt{(\Psi_t)}$ rather than (Ψ_t), we will bound

$$\begin{aligned}
&E\left[\sum_{t=1}^{T} \mathbb{P}(f(\theta) < f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1})\right] \\
\leq&E\left[\sum_{t=1}^{T} \sqrt{\mathbb{P}(f(\theta) < f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1})}\right] \\
\leq&E\left[\sum_{t=1}^{T} \sqrt{(\Phi_t) + (\Psi_t)}\right] \\
\leq&E\left[\sum_{t=1}^{T} \left(\sqrt{(\Phi_t)} + \sqrt{(\Psi_t)}\right)\right].
\end{aligned}$$

To bound (Φ_t), let τ_l denote the stopping time when all arms $k \in X^*$ have been selected for at least $l = 1, \ldots, T$ times.

$$E[\sum_{t=1}^{T} (\Phi_t)] = E\left[\sum_{t=1}^{T} \frac{\mathbb{P}(\exists i, k \in X^* : \theta_{ik}(t) < d_{ik} - \epsilon \mid \mathcal{F}_{t-1})}{\mathbb{P}(\forall i, k \in X^* : \theta_{ik}(t) \geq d_{ik} - \epsilon \mid \mathcal{F}_{t-1})} \mathbb{P}(X_\epsilon(t) = X^*, S(t) = S^* \mid \mathcal{F}_{t-1})\right]$$

$$
\begin{aligned}
=&E\left[\sum_{t=1}^{T}\frac{\mathbb{P}(\exists i,k\in X^*:\theta_{ik}(t)<d_{ik}-\epsilon\mid\mathcal{F}_{t-1})}{\mathbb{P}(\forall i,k\in X^*:\theta_{ik}(t)\geq d_{ik}-\epsilon\mid\mathcal{F}_{t-1})}I(X_\epsilon(t)=X^*,S(t)=S^*)\right]\\
\leq&E\Big[\sum_{t=1}^{\tau_1}\frac{\mathbb{P}(\exists i,k\in X^*:\theta_{ik}(t)<d_{ik}-\epsilon\mid\mathcal{F}_{t-1})}{\mathbb{P}(\forall i,k\in X^*:\theta_{ik}(t)\geq d_{ik}-\epsilon\mid\mathcal{F}_{t-1})}I(X_\epsilon(t)=X^*,S(t)=S^*)\\
&+\sum_{l=1}^{T-1}\sum_{t=\tau_l+1}^{\tau_{l+1}}\frac{\mathbb{P}(\exists i,k\in X^*:\theta_{ik}(t)<d_{ik}-\epsilon\mid\mathcal{F}_{t-1})}{\mathbb{P}(\forall i,k\in X^*:\theta_{ik}(t)\geq d_{ik}-\epsilon\mid\mathcal{F}_{t-1})}I(X_\epsilon(t)=X^*,S(t)=S^*)\Big].
\end{aligned}
$$

The first step is by the Tower rule. The second step is an inequality because $\tau_T \geq T$. By Lemma 2, if $l < 32/\epsilon$, we have

$$
\begin{aligned}
=&E\left[\sum_{t=\tau_l+1}^{\tau_{l+1}}\frac{\mathbb{P}(\exists i,k\in X^*:\theta_{ik}(t)<d_{ik}-\epsilon/4\mid\mathcal{F}_{t-1})}{\mathbb{P}(\forall i,k\in X^*:\theta_{ik}(t)\geq d_{ik}-\epsilon/4\mid\mathcal{F}_{t-1})}I(X_\epsilon(t)=X^*,S(t)=S^*)\right]\\
=&E\left[\sum_{t=\tau_l+1,X_\epsilon(t)=X^*,S(t)=S^*}^{\tau_{l+1}}\frac{\mathbb{P}(\exists i,k\in X^*:\theta_{ik}(t)<d_{ik}-\epsilon/4\mid\mathcal{F}_{t-1})}{\mathbb{P}(\forall i,k\in X^*:\theta_{ik}(t)\geq d_{ik}-\epsilon/4\mid\mathcal{F}_{t-1})}\right]\\
\leq&E\left[\sum_{t=\tau_l+1,X_\epsilon(t)=X^*,S(t)=S^*}^{\tau_{l+1}}O\left(\frac{1}{\Delta^{N(M+1)}}\right)\right]=\frac{|X^*|}{\gamma}\cdot O\left(\frac{1}{\Delta^{N|X^*|}}\right)
\end{aligned}
$$

The last step is due to Lemma 4. Similarly, if $l \geq 32/\epsilon$, we have

$$
\begin{aligned}
&E\left[\sum_{t=\tau_l+1}^{\tau_{l+1}}\frac{\mathbb{P}(\exists i,k\in X^*:\theta_{ik}(t)<d_{ik}-\epsilon/4\mid\mathcal{F}_{t-1})}{\mathbb{P}(\forall i,k\in X^*:\theta_{ik}(t)\geq d_{ik}-\epsilon/4\mid\mathcal{F}_{t-1})}I(X_\epsilon(t)=X^*,S(t)=S^*)\right]\\
=&E\left[\sum_{t=\tau_l+1,X_\epsilon(t)=X^*,S(t)=S^*}^{\tau_{l+1}}\frac{\mathbb{P}(\exists i,k\in X^*:\theta_{ik}(t)<d_{ik}-\epsilon/4\mid\mathcal{F}_{t-1})}{\mathbb{P}(\forall i,k\in X^*:\theta_{ik}(t)\geq d_{ik}-\epsilon/4\mid\mathcal{F}_{t-1})}\right]\\
\leq&E\left[\sum_{t=\tau_l+1,X_\epsilon(t)=X^*,S(t)=S^*}^{\tau_{l+1}}O\left(\frac{1}{l\Delta^{NM+N+1}}\right)\right]\\
=&\frac{|X^*|}{\gamma}O\left(\frac{1}{l\Delta^{N|X^*|+1}}\right)
\end{aligned}
$$

Summing over these terms, we have

$$
\begin{aligned}
&E[\sum_{t=1}^{T}(\Phi_t)]\\
=&E\left[\sum_{t=1}^{T}\frac{\mathbb{P}(\exists i,k\in X^*:\theta_{ik}(t)<d_{ik}-\epsilon/4\mid\mathcal{F}_{t-1})}{\mathbb{P}(\forall i,k\in X^*:\theta_{ik}(t)\geq d_{ik}-\epsilon/4\mid\mathcal{F}_{t-1})}\mathbb{P}(X_\epsilon(t)=X^*,S(t)=S^*\mid\mathcal{F}_{t-1})\right]\\
\leq&E\left[\sum_{l=1}^{32/\epsilon}O\left(\frac{|X^*|}{\gamma\Delta^{N|X^*|}}\right)+\sum_{l=32/\epsilon+1}^{T-1}O\left(\frac{|X^*|}{\gamma l\Delta^{N|X^*|+1}}\right)\right]\\
=&O\left(\frac{|X^*|\log T}{\gamma\Delta^{N|X^*|+1}}\right).
\end{aligned}
$$

Applying the Cauchy-Schwartz inequality twice, it holds that

$$
\begin{aligned}
\sum_{t=1}^{T}E[\sqrt{(\Phi_t)}]\leq&\sum_{t=1}^{T}\sqrt{E\left[(\Phi_t)\right]}\\
=&\sum_{t=1}^{T}\sqrt{E\left[(\Phi_t)\right]}\cdot\sqrt{1}
\end{aligned}
$$

$$\leq \sqrt{\sum_{t=1}^{T} E\left[(\Phi_t)\right]}\sqrt{T}$$
$$= O(\frac{\sqrt{|X^*|T\log T}}{\gamma\Delta^{(N|X^*|+1)/2}})$$

The inequalities use Cauchy-Schwartz inequality (cf. Fact B.3).

To bound (Ψ_t), we have

$$\sum_{t=1}^{T} E[\sqrt{(\Psi_t)}]$$
$$= \sum_{t=1}^{T} E\left[\sqrt{\frac{\mathbb{P}(\exists i, k \in X^* : \theta_{ik}(t) < d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})}{\mathbb{P}(\forall i, k \in X^* : \theta_{ik}(t) \geq d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})}\mathbb{P}(X_\epsilon(t) \neq X^* \text{ or } S(t) \neq S^*, f(\theta) \geq f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1})}\right]$$
$$\leq \sum_{t=1}^{T} \sqrt{E\left[\frac{\mathbb{P}(\exists i, k \in X^* : \theta_{ik}(t) < d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})}{\mathbb{P}(\forall i, k \in X^* : \theta_{ik}(t) \geq d_{ik} - \epsilon/4 \mid \mathcal{F}_{t-1})}\right]}\sqrt{E\left[\mathbb{P}(X_\epsilon(t) \neq X^* \text{ or } S(t) \neq S^*, f(\theta) \geq f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1})\right]}$$
$$\leq O(\frac{1}{\Delta^{(N|X^*|)/2}}) \sum_{t=1}^{T} \sqrt{E\left[\mathbb{P}(X_\epsilon(t) \neq X^* \text{ or } S(t) \neq S^*, f(\theta) \geq f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1})\right]}$$
$$\leq O(\frac{1}{\Delta^{(N|X^*|)/2}}) \sqrt{\sum_{t=1}^{T} E\left[\mathbb{P}(X_\epsilon(t) \neq X^* \text{ or } S(t) \neq S^*, f(\theta) \geq f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1})\right]}\sqrt{T}$$
$$\leq O(\frac{1}{\Delta^{(N|X^*|)/2}} \cdot \frac{\sqrt{MK\log NT}}{\Delta^{3/2}}\sqrt{T})$$
$$= O(\frac{1}{\Delta^{(N|X^*|+3)/2}}\sqrt{MKT\log NT})$$

The first inequality uses Fact B.3, the second inequality uses Lemma 2, the third inequality uses Fact B.3, and the last inequality uses the result in Case 1(a).

Finally, we have

$$E\left[\sum_{t=1}^{T} \mathbb{P}(f(\theta) < f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1})\right]$$
$$\leq E\left[\sum_{t=1}^{T}\left(\sqrt{(\Phi_t)} + \sqrt{(\Psi_t)}\right)\right]$$
$$\leq O(\frac{\sqrt{|X^*|T\log T}}{\Delta^{(N|X^*|+1)/2}}) + O(\frac{1}{\Delta^{(N|X^*|+3)/2}}\sqrt{MKT\log NT})$$
$$\leq O(\frac{1}{\Delta^{(N|X^*|+3)/2}}\sqrt{MKT\log NT}).$$

To summarize for the entire Part I, we show that

$$E\left[\sum_{t=1}^{T'} I(X_\epsilon(t) \neq X^* \text{ or } S(t) \neq S^*)\right] \leq E[\sum_{t=1}^{T} I(X_\epsilon(t) \neq X^* \text{ or } S(t) \neq S^*)]$$
$$= E\left[\sum_{t=1}^{T} I(X_\epsilon(t) \neq X^* \text{ or } S(t) \neq S^*, f(\theta) \geq f^* - \frac{\Delta}{2})\right]$$
$$+ E\left[\sum_{t=1}^{T} I(X_\epsilon(t) \neq X^* \text{ or } S(t) \neq S^*, f(\theta) < f^* - \frac{\Delta}{2})\right]$$
$$\leq O(\frac{MK\log NT}{\Delta^3}) + O(\frac{1}{\Delta^{(N|X^*|+3)/2}}\sqrt{MKT\log NT})$$
$$= O(\sqrt{T\log T}).$$

Part II: Bound the term $\sum_{k\in X^*} r_k E[\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*)\left(\frac{x_k^*(t)}{\sum_{k=1}^K x_k^*(t)} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)}\right)]$.

In this part, we bound the revenue loss when $X_\epsilon(t)=X^*$ and $S(t)=S^*$ happens. We first define a sequence of constants as follows:

$$\epsilon_1 = \min\{\epsilon, T'^{-\frac{1}{4}}\}, \epsilon_2 = \min\{\epsilon, T'^{-\frac{3}{8}}\}, \ldots, \epsilon_n = \min\{\epsilon, T'^{-\frac{2^n-1}{2^{n+1}}}\}, \ldots$$

and let τ_ν be the minimum t such that $N_k(t-1) \geq 2\log T'/\epsilon_\nu^2$ for all $k \in X^*$. By Lemma 3 and Fact B.1, there exists a constant L such that if $X_\epsilon(t)=X^*$, $S(t)=S^*$ and $t \geq \tau_\nu$, we have

$$\mathbb{P}\left(\left|\frac{x_k(t)}{\sum_{k=1}^K x_k(t)} - \frac{x_k^*}{\sum_{k=1}^K x_k^*}\right| \geq L\epsilon_\nu\right) \leq \frac{4N}{T'}.$$

Let n be the smallest number such that $2\log T'/\epsilon_n^2 \geq T'$. We have

$$\begin{aligned}
&E[\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*)(\frac{x_k^*}{\sum_{k=1}^K x_k^*} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)})] \\
=&E[\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*, t<\tau_1)(\frac{x_k^*}{\sum_{k=1}^K x_k^*} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)})] \\
&+ E[\sum_{\nu=1}^{n-1}\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*, \tau_\nu \leq t < \tau_{\nu+1})(\frac{x_k^*}{\sum_{k=1}^K x_k^*} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)})] \\
\leq&E[\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*, t<\tau_1)] \\
&+ E[\sum_{\nu=1}^{n-1}\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*, \tau_\nu \leq t < \tau_{\nu+1}) I(\frac{x_k^*}{\sum_{k=1}^K x_k^*} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)} \leq L\epsilon_\nu)\cdot L\epsilon_\nu] \\
&+ E[\sum_{\nu=1}^{n-1}\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*, \tau_\nu \leq t < \tau_{\nu+1}) I(\frac{x_k^*}{\sum_{k=1}^K x_k^*} - x_k(t) > L\epsilon_\nu)] \\
\leq&E[\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*, t<\tau_1)] + E[\sum_{\nu=1}^{n-1}\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*, \tau_\nu \leq t < \tau_{\nu+1})(L\epsilon_\nu + \frac{4N}{T'})]
\end{aligned}$$

By Lemma 4, we have

$$E[\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*, t<\tau_\nu)] \leq \frac{|X^*|}{\gamma} T'^{\frac{2^\nu-1}{2^\nu}} \log T', \text{ for all } \nu.$$

By definition of n, we have $2\log T'/\epsilon_{n-1}^2 < T'$, or equivalently $\epsilon_{n-1}^2 = T'^{-\frac{2^{n-1}-1}{2^n}} > 2\log T'/T'$. Simple algebra shows that $n < \log(2\log T'/\log(2\log T')) = O(\log\log T)$. Therefore,

$$\begin{aligned}
&E[\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*)(\frac{x_k^*}{\sum_{k=1}^K x_k^*} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)})] \\
\leq&E[\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*, t<\tau_1)] + \sum_{\nu=1}^{n-1} E[\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*, \tau_\nu \leq t < \tau_{\nu+1})(L\epsilon_\nu + \frac{4N}{T'})] \\
\leq&E[\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*, t<\tau_1)] + \sum_{\nu=1}^{n-1} E[\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*, t<\tau_{\nu+1})]\cdot L\epsilon_\nu \\
&+ \sum_{\nu=1}^{n-1} E[\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*, \tau_\nu \leq t < \tau_{\nu+1})]\cdot\frac{4N}{T'}
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{|X^*|}{\gamma}\sqrt{T'}\log T' + \sum_{\nu=1}^{n-1}\left(\frac{|X^*|}{\gamma}T'^{\frac{2^{\nu+1}-1}{2^{\nu+1}}}\log(T')\cdot L\epsilon_\nu\right) + T'\cdot\frac{4N}{T'}\\
&\leq \frac{|X^*|}{\gamma}\sqrt{T'}\log T' + (n-1)\frac{|X^*|}{\gamma}L\sqrt{T'}\log T' + 4N\\
&= \frac{|X^*|}{\gamma}\sqrt{T'}\log T' + O(\log\log T)\frac{|X^*|}{\gamma}L\sqrt{T'}\log T' + 4N.
\end{aligned}
$$

The second to last step uses the fact that $T'^{\frac{2^{\nu+1}-1}{2^{\nu+1}}}\epsilon_\nu = \sqrt{T'}$ for all ν. To recap, because $r_k \leq 1$ for all k, we have the following bound

$$
\begin{aligned}
&\sum_{k\in X^*} r_k E[\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*)(\frac{x_k^*}{\sum_{k=1}^K x_k^*} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)})]\\
\leq &\frac{|X^*|^2}{\gamma}\sqrt{T'}\log T' + O(\log\log T)\frac{|X^*|^2}{\gamma}L\sqrt{T'}\log T' + 4N\\
= &O(\sqrt{T}\log T\log\log T).
\end{aligned}
$$

Part III: Bound term $E[\sum_{j=1}^M(\sum_{i=1}^N\sum_{t=1}^{T'} a_{ij}Z_i(t) - I_j)^+]$.

For any resource $j = 1,\ldots,M$, we let $c_j' = c_j/(\sum_{k=1}^K x_k^*)$. This is interpreted as the expected rate of inventory consumption of resource j in the first T' time periods. Since $I_j = c_jT \geq c_j'T'$, we have

$$
\begin{aligned}
&E[\left(\sum_{i=1}^N\sum_{t=1}^{T'} a_{ij}Z_i(t) - I_j\right)^+]\\
\leq &E[\sum_{i=1}^N\sum_{t=1}^{T'} a_{ij}I(X_\epsilon(t)\neq X^* \text{ or } S(t)\neq S^*)]\\
&+ E[\left(\sum_{i=1}^N\sum_{t=1}^{T'} a_{ij}Z_i(t)I(X_\epsilon(t)=X^*, S(t)=S^*) - c_j'T'\right)^+].
\end{aligned}
$$

We show in Part I of the proof that

$$
E[\sum_{i=1}^N\sum_{t=1}^{T'} a_{ij}I(X_\epsilon(t)\neq X^* \text{ or } S(t)\neq S^*)] \leq O(\frac{MK\log NT}{\Delta^3}) + O(\frac{1}{\Delta^{(N|X^*|+3)/2}}\sqrt{MKT\log NT}).
$$

Furthermore, we have

$$
\begin{aligned}
&E[\left(\sum_{i=1}^N\sum_{t=1}^{T'} a_{ij}Z_i(t)I(X_\epsilon(t)=X^*, S(t)=S^*) - c_j'T'\right)^+]\\
\leq &E[\Big(\sum_{i=1}^N\sum_{t=1}^{T'}(a_{ij}Z_i(t)I(X_\epsilon(t)=X^*, S(t)=S^*, t<\tau_1) - c_j')\\
&+ \sum_{\nu=1}^{n-1}\sum_{t=1}^{T'}(\sum_{i=1}^N a_{ij}Z_i(t)I(X_\epsilon(t)=X^*, S(t)=S^*, \tau_\nu\leq t<\tau_{\nu+1}) - c_j')\Big)^+]\\
\leq &E[\Big(\sum_{i=1}^N\sum_{t=1}^{T'}(a_{ij}Z_i(t)I(X_\epsilon(t)=X^*, S(t)=S^*, t<\tau_1) - c_j')\Big)^+\\
&+ \sum_{\nu=1}^{n-1}\Big(\sum_{t=1}^{T'}(\sum_{i=1}^N a_{ij}Z_i(t)I(X_\epsilon(t)=X^*, S(t)=S^*, \tau_\nu\leq t<\tau_{\nu+1}) - c_j')\Big)^+]
\end{aligned}
$$

$$\leq E[\sum_{i=1}^{N}\sum_{t=1}^{\tau_1} a_{ij} + \sum_{\nu=1}^{n-1}\Big(\sum_{t=1}^{T'}(\sum_{i=1}^{N} a_{ij} Z_i(t) I(X_\epsilon(t)=X^*, S(t)=S^*, \tau_\nu \leq t < \tau_{\nu+1}) - c'_j)\Big)^+].$$

For each $\nu = 1 \ldots, n$, we define $U_i^\nu(t)$ for each $i = 1, \ldots, N$ and $t = 1, \ldots, T'$ to be a Bernoulli random variable with success probability $\sum_{k=1}^K d_{ik}(x_k^*/(\sum_{k=1}^K x^*) + L\epsilon_\nu)$. Recall that $(\sum_{k=1}^K d_{ik}x_k^*)/(\sum_{k=1}^K x^*)$ is the expected sales of product i by choosing arms according to the optimal solution x^*. Under the case $\{X_\epsilon(t) = X^*, S(t) = S^*, \tau_\nu \leq t < \tau_{\nu+1}\}$, the event $x_k(t)/(\sum_{k=1}^K x_k(t)) \leq x_k^*/(\sum_{k=1}^K x^*) + L\epsilon_\nu$ happens with probability at least $1 - 4N/T$ (by Lemma 3 and B.1), so we can upper bound $Z_i(t)$ by $U_i^\nu(t)$ with high probability.

Conditioning on τ_ν and τ_{nu+1},

$$\begin{aligned}
&E[\Big(\sum_{t=1}^{T'}(\sum_{i=1}^{N} a_{ij} Z_i(t) I(X_\epsilon(t)=X^*, S(t)=S^*, \tau_\nu \leq t < \tau_{\nu+1}) - c'_j)\Big)^+ \mid \tau_\nu, \tau_{\nu+1}] \\
\leq& E[\Big(\sum_{t=\tau_\nu}^{\tau_{\nu+1}-1}(\sum_{i=1}^{N} a_{ij} U_i^\nu(t) - c'_j)\Big)^+ \mid \tau_\nu, \tau_{\nu+1}] + E[\sum_{t=1}^{T'} I(\tau_\nu \leq t < \tau_{\nu+1})\frac{4N}{T'}] \\
\leq& \frac{\sqrt{\sigma_\nu^2(\tau_{\nu+1}-\tau_\nu) + (KL\epsilon_\nu)^2(\tau_{\nu+1}-\tau_\nu)^2} + KL\epsilon_\nu(\tau_{\nu+1}-\tau_\nu)}{2} + E[\sum_{t=1}^{T'} I(\tau_\nu \leq t < \tau_{\nu+1})\frac{4N}{T'}].
\end{aligned}$$

Since $U_i^\nu(t)$ are i.i.d., we use Fact B.4 in the last inequality. The constant σ_ν^2 is the variance of $\sum_{i=1}^N a_{ij}U_i^\nu(t)$; note that $\sigma_\nu^2 \leq (\sum_{i=1}^N a_{ij})/4 \leq 1/4$ because the maximum variance of a Bernoulli random variable is 1/4. We also use the fact that $E[\sum_{i=1}^N a_{ij}U_i^\nu(t) - c'_j] = KL\epsilon_\nu$.

Summing over ν, we have

$$\begin{aligned}
&E[\sum_{i=1}^{N}\sum_{t=1}^{\tau_1} a_{ij} + \sum_{\nu=1}^{n-1}\Big(\sum_{t=1}^{T'}(\sum_{i=1}^{N} a_{ij} Z_i(t) I(X_\epsilon(t)=X^*, S(t)=S^*, \tau_\nu \leq t < \tau_{\nu+1}) - c'_j)\Big)^+] \\
\leq& \frac{|X^*|}{\gamma}\sqrt{T'}\log T' + \sum_{\nu=1}^{n-1} E[\frac{\sqrt{\sigma_\nu^2(\tau_{\nu+1}-\tau_\nu) + (KL\epsilon_\nu)^2(\tau_{\nu+1}-\tau_\nu)^2} + KL\epsilon_\nu(\tau_{\nu+1}-\tau_\nu)}{2}] \\
&+ \sum_{\nu=1}^{n-1} E[\sum_{t=1}^{T'} I(\tau_\nu \leq t < \tau_{\nu+1})\frac{4N}{T'}] \\
\leq& \frac{|X^*|}{\gamma}\sqrt{T'}\log T' + \sum_{\nu=1}^{n-1} E[\frac{\sqrt{1/4 * (\tau_{\nu+1}-\tau_\nu)} + KL\epsilon_\nu(\tau_{\nu+1}-\tau_\nu) + KL\epsilon_\nu(\tau_{\nu+1}-\tau_\nu)}{2}] + T' \cdot \frac{4N}{T'} \\
\leq& \frac{|X^*|}{\gamma}\sqrt{T'}\log T' + \sum_{\nu=1}^{n-1} E[\frac{\sqrt{(\tau_{\nu+1}-\tau_\nu)}}{4} + LK\epsilon_\nu(\tau_{\nu+1}-\tau_\nu)] + 4N \\
\leq& \frac{|X^*|}{\gamma}\sqrt{T'}\log T' + \frac{1}{4}E[\sum_{\nu=1}^{n-1}\sqrt{(\tau_{\nu+1}-\tau_\nu)}] + O(K\frac{|X^*|}{\gamma}\sqrt{T}\log T\log\log T) + 4N \\
\leq& \frac{|X^*|}{\gamma}\sqrt{T'}\log T' + \frac{1}{4}E[\sqrt{\sum_{\nu=1}^{n-1}(\tau_{\nu+1}-\tau_\nu)}\sqrt{\sum_{\nu=1}^{n-1}1}] + O(K\frac{|X^*|}{\gamma}\sqrt{T}\log T\log\log T) + 4N \\
\leq& \frac{|X^*|}{\gamma}\sqrt{T'}\log T' + \frac{1}{4}\sqrt{E[\sum_{\nu=1}^{n-1}(\tau_{\nu+1}-\tau_\nu)]}\sqrt{n-1} + O(K\frac{|X^*|}{\gamma}\sqrt{T}\log T\log\log T) + 4N \\
\leq& \frac{|X^*|}{\gamma}\sqrt{T'}\log T' + O(\sqrt{T})O(\sqrt{\log\log T}) + O(K\frac{|X^*|}{\gamma}\sqrt{T}\log T\log\log T) + 4N.
\end{aligned}$$

The first inequality uses the fact that $E[\tau_1] \leq \frac{|X^*|}{\gamma}\sqrt{T'}\log T'$. The fourth inequality uses the result from

Part II:

$$\sum_{\nu=1}^{n-1} E[\epsilon_\nu(\tau_{\nu+1}-\tau_\nu)] \le O(\frac{|X^*|}{\gamma}\sqrt{T}\log T\log\log T).$$

The fifth and sixth inequalities are due to the Cauchy-Schwartz inequality (Fact B.3). The last inequality is due to $E[\tau_n] = O(T)$ and $n = O(\log\log T)$, which are results from Part II.

To summarize, we have

$$\begin{aligned}E[\sum_{j=1}^{M}(\sum_{i=1}^{N}\sum_{t=1}^{T'} a_{ij}Z_i(t) - I_j)^+] \le& M\Big(O(\frac{MK\log NT}{\Delta^3}) + O(\frac{1}{\Delta^{(N|X^*|+3)/2}}\sqrt{MKT\log NT})\\ &+ O(\sqrt{T}\log T) + O(\sqrt{T})O(\sqrt{\log\log T}) + O(\sqrt{T}\log T\log\log T)\Big)\\ =&O(\sqrt{T}\log T\log\log T).\end{aligned}$$

Combining Parts I, II and III completes the proof.

The regret is bounded by

$$\begin{aligned}Regret(T) \le& f^*T - E[R(T')]\\ \le& f^*T - f^*T + 1 + (\max_{k\in X^*} r_k)E[\sum_{t=1}^{T'} I(X_\epsilon(t)\neq X^* \text{ or } S(t)\neq S^*)]\\ &+ \sum_{k\in X^*} r_k E[\sum_{t=1}^{T'} I(X_\epsilon(t)=X^*, S(t)=S^*)\left(\frac{x_k^*(t)}{\sum_{k=1}^K x_k^*(t)} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)}\right)]\\ &+ E[\sum_{j=1}^{M}(\sum_{i=1}^{N}\sum_{t=1}^{T'} a_{ij}Z_i(t) - I_j)^+]\\ \le& O(\sqrt{T\log T}) + O(\sqrt{T}\log T\log\log T) + O(\sqrt{T}\log T\log\log T)\\ =& O(\sqrt{T}\log T\log\log T).\end{aligned}$$

Extending the proof for multiple optimal solutions.

We consider the case where the optimal solution of OPT_{UB} is any convex combination of two extreme points (x^*, s^*) and $(\bar{x}, \bar{s})$. The case of more than two optimal extreme points can be resolved using the same method. Define $X^* = \{k \mid x_k^* > 0\}, S^* = \{j \mid s_j^* > 0\}, \bar{X} = \{k \mid \bar{x}_k > 0\}, \bar{S} = \{j \mid \bar{s_j} > 0\}$.

In order to modify Lemma 1, we define constant Δ as follows:

$$\begin{aligned}\Delta = \min\big\{X, S :& (X^* \subsetneq X \text{ or } S^* \subsetneq S) \text{ and } (\bar{X} \subsetneq X \text{ or } \bar{S} \subsetneq S) :\\ &\max_y f^* - \sum_{j=1}^{M} c_j y_j - y_0\\ &\text{subject to } \sum_{j=1}^{M} b_{jk}y_j + y_0 \ge r_k \text{ for all } k \in X, y_j \ge 0 \text{ for all } j \in S\big\}.\end{aligned}$$

Then we define constant ϵ the same as in Section A.1. The equivalence of Lemma 1 is the following, and we omit the proof since it is almost identical to the proof of Lemma 1.

Lemma 5. *Consider problem $OPT(\theta)$ where the decision variables are restricted to a subset X and S that do not satisfy either of the conditions: 1) $X^* \subset X$ and $S^* \subset S$; 2) $\bar{X} \subset X$ and $\bar{S} \subset S$. If $|\theta_{ik}(t) - d_{ik}| \le \epsilon$ for all $i = 1, \ldots, N$ and $k \in X$, the optimal value of $OPT(\theta)$ satisfies $f(\theta) \le f^* - \frac{3\Delta}{4}$.*

For any period $t = 1, \ldots, T$, the definitions of the sets $X_\epsilon(t)$ and $S(t)$ remain the same. We define

indicator functions $I_t^* = I(X_\epsilon(t) = X^*, S(t) = S^*)$ and $\bar{I}_t = I(X_\epsilon(t) = \bar{X}, S(t) = \bar{S})$. Let

$$T' = \min\{\tau : \sum_{t=1}^{\tau} \left(\frac{I_t^*}{\sum_{k=1}^K x_k^*} + \frac{\bar{I}_t}{\sum_{k=1}^K \bar{x}_k} + (1 - I_t^* - \bar{I}_t) \right) \geq T\}.$$

The new definition of T' represents a random variable interpreted as the time when inventory runs out. We again consider the retailer's revenue before period T' (note that $T' \leq T$ almost surely):

$$\begin{aligned}
E[R(T)] \geq & E[R(T')] \\
\geq & \sum_{k=1}^K r_k E[N_k(T')] - E[\sum_{j=1}^M (\sum_{i=1}^N \sum_{t=1}^{T'} a_{ij} Z_i(t) - I_j)^+] \\
= & \sum_{k=1}^K r_k E\left[\sum_{t=1}^{T'} \frac{x_k(t)}{\sum_{k=1}^K x_k(t)} \right] - E[\sum_{j=1}^M (\sum_{i=1}^N \sum_{t=1}^{T'} a_{ij} Z_i(t) - I_j)^+] \\
= & \sum_{k=1}^K r_k E\left[\sum_{t=1}^{T'} \frac{x_k(t)}{\sum_{k=1}^K x_k(t)} (I_t^* + \bar{I}_t) \right] + \sum_{k=1}^K r_k E\left[\sum_{t=1}^{T'} \frac{x_k(t)}{\sum_{k=1}^K x_k(t)} (1 - \bar{I}_t - I_t^*) \right] \\
& - E[\sum_{j=1}^M (\sum_{i=1}^N \sum_{t=1}^{T'} a_{ij} Z_i(t) - I_j)^+] \\
= & \sum_{k=1}^K r_k E\left[\sum_{t=1}^{T'} (\frac{x_k^*}{\sum_{k=1}^K x_k^*} I_t^* + \frac{\bar{x}_k}{\sum_{k=1}^K \bar{x}_k} \bar{I}_t + (1 - \bar{I}_t - I_t^*)) \right] - \sum_{k=1}^K r_k E\left[\sum_{t=1}^{T'} \left(\frac{x_k^*(t)}{\sum_{k=1}^K x_k^*(t)} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)} \right) I_t^* \right] \\
& - \sum_{k=1}^K r_k E\left[\sum_{t=1}^{T'} \left(\frac{\bar{x}_k}{\sum_{k=1}^K \bar{x}_k} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)} \right) \bar{I}_t) \right] + \sum_{k=1}^K r_k E\left[\sum_{t=1}^{T'} (\frac{x_k(t)}{\sum_{k=1}^K x_k(t)} - 1)(1 - \bar{I}_t - I_t^*) \right] \\
& - E[\sum_{j=1}^M (\sum_{i=1}^N \sum_{t=1}^{T'} a_{ij} Z_i(t) - I_j)^+] \\
\geq & f^* T - \sum_{k=1}^K r_k E\left[\sum_{t=1}^{T'} \left(\frac{x_k^*(t)}{\sum_{k=1}^K x_k^*(t)} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)} \right) I_t^* \right] - \sum_{k=1}^K r_k E\left[\sum_{t=1}^{T'} \left(\frac{\bar{x}_k}{\sum_{k=1}^K \bar{x}_k} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)} \right) \bar{I}_t) \right] \\
& - E[\sum_{t=1}^{T'} (1 - \bar{I}_t - I_t^*)] - E[\sum_{j=1}^M (\sum_{i=1}^N \sum_{t=1}^{T'} a_{ij} Z_i(t) - I_j)^+].
\end{aligned}$$

The last inequality uses the definition of T'. Similar to the case of having a unique optimal solution, we bound $E[\sum_{t=1}^{T'} (1 - \bar{I}_t - I_t^*)]$ in Part I of this proof, the term $\sum_{k=1}^K r_k E\left[\sum_{t=1}^{T'} \left(\frac{x_k^*(t)}{\sum_{k=1}^K x_k^*(t)} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)} \right) I_t^*\right] + \sum_{k=1}^K r_k E\left[\sum_{t=1}^{T'} \left(\frac{\bar{x}_k}{\sum_{k=1}^K \bar{x}_k} - \frac{x_k(t)}{\sum_{k=1}^K x_k(t)} \right) \bar{I}_t)\right]$ in Part II of the proof, and the last term $E[\sum_{j=1}^M (\sum_{i=1}^N \sum_{t=1}^{T'} a_{ij} Z_i(t) - I_j)^+]$ in Part III.

To prove Part I, we consider two cases: (a) $I_t^* = \bar{I}_t = 0$, and $f(\theta) \geq f^* - \Delta/2$; (b) $f(\theta) \leq f^* - \Delta/2$. The proof of Case (a) follows almost step by step, except that we now use Lemma 5, which replaces Lemma 1 used in the unique optimal solution case. In Case (b), we have

$$\begin{aligned}
& \mathbb{P}(f(\theta) < f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1}) \\
\leq & \mathbb{P}(\left\{ \exists i, k \in X^* : \theta_{ik}(t) < d_{ik} - \frac{\epsilon}{4} \right\} \cap \left\{ \exists i, k \in \bar{X} : \theta_{ik}(t) < d_{ik} - \frac{\epsilon}{4} \right\} \mid \mathcal{F}_{t-1}) \\
& \cdot \frac{\mathbb{P}(I_t^* = 1 \mid \mathcal{F}_{t-1}) + \mathbb{P}(\bar{I}_t = 1 \mid \mathcal{F}_{t-1}) + \mathbb{P}(I_t^* = \bar{I}_t = 0, f(\theta) \geq f^* - \frac{\Delta}{2} \mid \mathcal{F}_{t-1})}{\mathbb{P}(\{\forall i, k \in X^* : \theta_{ik}(t) \geq d_{ik} - \epsilon/4\} \cup \left\{ \forall i, k \in \bar{X} : \theta_{ik}(t) \geq d_{ik} - \epsilon/4 \right\} \mid \mathcal{F}_{t-1})} \\
= & \frac{\mathbb{P}(\left\{ \exists i, k \in X^* : \theta_{ik}(t) < d_{ik} - \frac{\epsilon}{4} \right\} \cap \left\{ \exists i, k \in \bar{X} : \theta_{ik}(t) < d_{ik} - \frac{\epsilon}{4} \right\} \mid \mathcal{F}_{t-1})}{\mathbb{P}(\{\forall i, k \in X^* : \theta_{ik}(t) \geq d_{ik} - \epsilon/4\} \cup \left\{ \forall i, k \in \bar{X} : \theta_{ik}(t) \geq d_{ik} - \epsilon/4 \right\} \mid \mathcal{F}_{t-1})} \mathbb{P}(I_t^* = 1 \mid \mathcal{F}_{t-1})
\end{aligned}$$

$$
\begin{aligned}
&+\frac{\mathbb{P}(\{\exists i,k\in X^*:\theta_{ik}(t)<d_{ik}-\frac{\epsilon}{4}\}\cap\{\exists i,k\in\bar{X}:\theta_{ik}(t)<d_{ik}-\frac{\epsilon}{4}\}\mid\mathcal{F}_{t-1})}{\mathbb{P}(\{\forall i,k\in X^*:\theta_{ik}(t)\geq d_{ik}-\epsilon/4\}\cup\{\forall i,k\in\bar{X}:\theta_{ik}(t)\geq d_{ik}-\epsilon/4\}\mid\mathcal{F}_{t-1})}\mathbb{P}(\bar{I}_t=1\mid\mathcal{F}_{t-1})\\
&+\frac{\mathbb{P}(\{\exists i,k\in X^*:\theta_{ik}(t)<d_{ik}-\frac{\epsilon}{4}\}\cap\{\exists i,k\in\bar{X}:\theta_{ik}(t)<d_{ik}-\frac{\epsilon}{4}\}\mid\mathcal{F}_{t-1})}{\mathbb{P}(\{\forall i,k\in X^*:\theta_{ik}(t)\geq d_{ik}-\epsilon/4\}\cup\{\forall i,k\in\bar{X}:\theta_{ik}(t)\geq d_{ik}-\epsilon/4\}\mid\mathcal{F}_{t-1})}\mathbb{P}(I_t^*=\bar{I}_t=0,f(\theta)\geq f^*-\frac{\Delta}{2}\mid\mathcal{F}_{t-1})\\
\leq&\frac{\mathbb{P}(\exists i,k\in X^*:\theta_{ik}(t)<d_{ik}-\frac{\epsilon}{4}\mid\mathcal{F}_{t-1})}{\mathbb{P}(i,k\in X^*:\theta_{ik}(t)\geq d_{ik}-\epsilon/4\mid\mathcal{F}_{t-1})}\mathbb{P}(I_t^*=1\mid\mathcal{F}_{t-1})\\
&+\frac{\mathbb{P}(\exists i,k\in\bar{X}:\theta_{ik}(t)<d_{ik}-\frac{\epsilon}{4}\mid\mathcal{F}_{t-1})}{\mathbb{P}(\forall i,k\in\bar{X}:\theta_{ik}(t)\geq d_{ik}-\epsilon/4\mid\mathcal{F}_{t-1})}\mathbb{P}(\bar{I}_t=1\mid\mathcal{F}_{t-1})\\
&+\frac{\mathbb{P}(\exists i,k\in X^*:\theta_{ik}(t)<d_{ik}-\frac{\epsilon}{4}\mid\mathcal{F}_{t-1})}{\mathbb{P}(\forall i,k\in X^*:\theta_{ik}(t)\geq d_{ik}-\epsilon/4\mid\mathcal{F}_{t-1})}\mathbb{P}(I_t^*=\bar{I}_t=0,f(\theta)\geq f^*-\frac{\Delta}{2}\mid\mathcal{F}_{t-1})\\
\triangleq&(\Phi_t)+(\Gamma_t)+(\Psi_t)
\end{aligned}
$$

Therefore, we get

$$
\begin{aligned}
&E\left[\sum_{t=1}^{T}\mathbb{P}(f(\theta)<f^*-\frac{\Delta}{2}\mid\mathcal{F}_{t-1})\right]\\
\leq&E\left[\sum_{t=1}^{T}\sqrt{\mathbb{P}(f(\theta)<f^*-\frac{\Delta}{2}\mid\mathcal{F}_{t-1})}\right]\\
\leq&E\left[\sum_{t=1}^{T}\sqrt{(\Phi_t)+(\Gamma_t)+(\Psi_t)}\right]\\
\leq&E\left[\sum_{t=1}^{T}\left(\sqrt{(\Phi_t)}+\sqrt{(\Gamma_t)}+\sqrt{(\Psi_t)}\right)\right].
\end{aligned}
$$

We can then bound each of these terms as before: $E[\sum_{t=1}^{T}\sqrt{(\Phi_t)}]$, $E[\sum_{t=1}^{T}\sqrt{(\Gamma_t)}]$, $E[\sqrt{(\Psi_t)}]$. This finishes the proof of Part I.

Part II also requires little change: we can establish the bound for $E\left[\sum_{t=1}^{T'}\left(\frac{x_k^*(t)}{\sum_{k=1}^{K}x_k^*(t)}-\frac{x_k(t)}{\sum_{k=1}^{K}x_k(t)}\right)I_t^*\right]$ and $E\left[\sum_{t=1}^{T'}\left(\frac{\bar{x}_k}{\sum_{k=1}^{K}\bar{x}_k}-\frac{x_k(t)}{\sum_{k=1}^{K}x_k(t)}\right)\bar{I}_t)\right]$ respectively, using the same proof as before.

As for Part III, for any resource $j=1,\ldots,M$, we have the following decomposition

$$
\begin{aligned}
&E[\left(\sum_{i=1}^{N}\sum_{t=1}^{T'}a_{ij}Z_i(t)-c_jT\right)^+]\\
\leq&E[\sum_{i=1}^{N}\sum_{t=1}^{T'}a_{ij}(1-\bar{I}_t-I_t^*)]+E[\left(\sum_{i=1}^{N}\sum_{t=1}^{T'}(a_{ij}Z_i(t)-\frac{c_j}{\sum_{k=1}^{K}x_k^*(t)})I_t^*\right)^+]\\
&+E[\left(\sum_{i=1}^{N}\sum_{t=1}^{T'}(a_{ij}Z_i(t)-\frac{c_j}{\sum_{k=1}^{K}x_k(t))})\bar{I}_t\right)^+].
\end{aligned}
$$

Then, the original proof of Part III can be applied separately to each of the three terms.

Extending the proof for TS-general algorithm.

We extend the regret bound for TS-general by reducing it to the setting of TS-fixed. To this end, given any multi-armed bandit problem with global constraints described in Section 3.2, we suppose there are $N=M+1$ products, indexed by $i=0,1,\ldots,M$. If $i\neq 0$, the demand of product i under arm k is a binary random variable, which takes the value of $\bar{b}_{ik}$ or 0 with mean b_{ik}. Product i consumes one unit of resource j if $i=j$, and zero units of resource j if $i\neq j$. These products do not generate any revenue. In the special case of $i=0$, the demand of product 0 under arm k is a binary random variable, which takes value of $\bar{r}_k$or 0 with mean r_k. Product 0 does not consume any resources, but each unit sold generates

one unit of revenue. With this reduction, applying TS-general to the multi-armed bandit problem with global constraints is equivalent to solving a network revenue management problem using TS-fixed.

B Useful Facts

Fact B.1 Let $\theta_{ik}(t)$ be the sampled demand of product i under price p_k at time t, and d_{ik} be the mean demand of product i under price p_k. Let $N_k(t-1)$ be the number of times that price p_k is offered before time t. For any $\epsilon > 0$, Agrawal and Goyal (2013) (Lemma 2 and Lemma 3) show that

$$\mathbb{P}(|\theta_{ik}(t) - d_{ik}| \geq \epsilon \mid \mathcal{F}_{t-1}) \leq 4e^{-\epsilon^2 N_k(t-1)/2}.$$

In particular, if $N_k(t-1) \geq 2\log T/\epsilon^2$, we have

$$\mathbb{P}(|\theta_{ik}(t) - d_{ik}| \geq \epsilon \mid \mathcal{F}_{t-1}) \leq \frac{4}{T}.$$

Proof. This proof is slightly modified from Agrawal and Goyal (2013). We can derive the above inequalities using the fact that the $(i+1)$th order statistic of $n+1$ uniformly distributed variables is distributed as $Beta(i+1, n-i+1)$. Therefore, we have

$$\begin{aligned}
&\mathbb{P}(|\theta_{ik}(t) - d_{ik}| \geq \epsilon \mid \mathcal{F}_{t-1}) \\
=&\mathbb{P}(\theta_{ik}(t) \geq d_{ik} + \epsilon \mid \mathcal{F}_{t-1}) + \mathbb{P}(\theta_{ik}(t) \leq d_{ik} - \epsilon \mid \mathcal{F}_{t-1}) \\
=&\mathbb{P}(\sum_{i=1}^{N_k(t-1)+1} X_i \leq \hat{d}_{ik}(t) N_k(t-1) \mid \mathcal{F}_{t-1}) + \mathbb{P}(\sum_{i=1}^{N_k(t-1)+1} Y_i \geq \hat{d}_{ik}(t) N_k(t-1) + 1 \mid \mathcal{F}_{t-1}),
\end{aligned}$$

where X_i's are i.i.d. Bernoulli random variables with mean $d_{ik} + \epsilon$, Y_i's are i.i.d. Bernoulli random variables with mean $d_{ik} - \epsilon$, and $\hat{d}_{ik}(t)$ is the empirical mean demand of product i under price p_k for the first $t-1$ periods. By Hoeffding's inequality, we have

$$\mathbb{P}(|\hat{d}_{ik}(t) - d_{ik}| \geq \frac{\epsilon}{2} \mid \mathcal{F}_{t-1}) \leq 2e^{-\epsilon^2 N_k(t-1)/2}.$$

So

$$\begin{aligned}
&\mathbb{P}(\sum_{i=1}^{N_k(t-1)+1} X_i \leq \hat{d}_{ik}(t) N_k(t-1) \mid \mathcal{F}_{t-1}) + \mathbb{P}(\sum_{i=1}^{N_k(t-1)+1} Y_i \geq \hat{d}_{ik}(t) N_k(t-1) + 1 \mid \mathcal{F}_{t-1}) \\
\leq&\mathbb{P}(\sum_{i=1}^{N_k(t-1)+1} X_i \leq (d_{ik} + \frac{\epsilon}{2}) N_k(t-1) \mid \mathcal{F}_{t-1}) + \mathbb{P}(\sum_{i=1}^{N_k(t-1)+1} Y_i \geq (d_{ik} - \frac{\epsilon}{2}) N_k(t-1) + 1 \mid \mathcal{F}_{t-1}) + 2e^{-\epsilon^2 N_k(t-1)/2} \\
\leq&\mathbb{P}(\sum_{i=1}^{N_k(t-1)} X_i \leq (d_{ik} + \frac{\epsilon}{2}) N_k(t-1) \mid \mathcal{F}_{t-1}) + \mathbb{P}(\sum_{i=1}^{N_k(t-1)} Y_i \geq (d_{ik} - \frac{\epsilon}{2}) N_k(t-1) \mid \mathcal{F}_{t-1}) + 2e^{-\epsilon^2 N_k(t-1)/2} \\
\leq&e^{-\epsilon^2 N_k(t-1)/2} + e^{-\epsilon^2 N_k(t-1)/2} + 2e^{-\epsilon^2 N_k(t-1)/2} = 4e^{-\epsilon^2 N_k(t-1)/2}.
\end{aligned}$$

The last inequality again uses the Hoeffding bound. □

Fact B.2 Let $\theta_{ik}(t)$ be the sampled demand of product i under price p_k at time t, and let d_{ik} be the mean demand of product i under price p_k. Let $N_k(t-1)$ be the number of times that price p_k is offered before time t. For any $\epsilon > 0$, Agrawal and Goyal (2013) (Lemma 4) show that

$$E[\frac{1}{\mathbb{P}(\theta_{ik}(t) \geq d_{ik} - \epsilon \mid \mathcal{F}_{t-1})} \mid N_k(t-1)] \leq 1 + \frac{3}{\epsilon}.$$

Furthermore, if $N_k(t-1) \geq 8/\epsilon$, it holds that

$$E[\frac{1}{\mathbb{P}(\theta_{ik}(t) \geq d_{ik} - \epsilon \mid \mathcal{F}_{t-1})} \mid N_k(t-1)] \leq 1 + O(\frac{1}{\epsilon^2 N_k(t-1)}).$$

(In fact, the bound proved by Agrawal and Goyal (2013) is tighter than $O(\frac{1}{\epsilon^2 N_k(t-1)})$, but we only need the looser bound above in the proof.)

Fact B.3 We use two forms of the Cauchy-Schwartz inequality in the proof. Suppose that $a_i \geq 0$ and $b_i \geq 0$ for all $i = 1, \ldots, n$, we have

$$\sqrt{\sum_{i=1}^{n} a_i} \sqrt{\sum_{i=1}^{n} b_i} \geq \sum_{i=1}^{n} \sqrt{a_i b_i}.$$

Suppose that $A \geq 0$ and $B \geq 0$ almost surely, we have

$$\sqrt{E[A]}\sqrt{E[B]} \geq E[\sqrt{AB}].$$

Fact B.4 Given constants $c > 0$, $\epsilon \geq 0$ and $\sigma \geq 0$, suppose C_i are i.i.d random variables with mean $c + \epsilon$ and variance σ^2 for all $i = 1, \ldots, t$. Gallego and Van Ryzin (1994) uses the following inequality (in the proof of Theorem 3):

$$E[(\sum_{i=1}^{t} U_i - ct)^+] \leq \frac{\sqrt{\sigma^2 t + (\epsilon t)^2} + \epsilon t}{2}.$$

In fact, the following looser bound is enough for our proof:

$$E[(\sum_{i=1}^{t} U_i - ct)^+] \leq E[|\sum_{i=1}^{t} U_i - ct|] \leq \sqrt{E[(\sum_{i=1}^{t} U_i - ct)^2]} = \sqrt{\sigma^2 t + (\epsilon t)^2}.$$

C Thompson Sampling with Unlimited Inventory

The original Thompson sampling algorithm proposed by Thompson (1933) is described below. Note that it can be viewed as a special case of the TS-fixed algorithm proposed in Section 3 if inventory is unlimited.

Suppose there are K arms, indexed by $k = 1, \ldots, K$. At each time period $t = 1, \ldots, T$, the decision maker chooses to pull one of the arms. If arm k is pulled, the decision maker receives a reward of either 0 or 1 with fixed mean r_k. We assume that the rewards in different time periods are independent, and r_k is unknown to the decision maker. Let $N_k(t-1)$ be the number of times that arm k is pulled before period t, and let $R_k(t-1)$ be the total rewards from arm k before period t. Thompson (1933) propose a strategy shown in Algorithm 6.

Algorithm 6 Thompson Sampling with Unlimited Inventory

Let $N_k(0) = 0, R_k(0) = 1$ for all arms $k = 1, \ldots, K$. Repeat the following steps for all $t = 1, ..., T$:

1. Sample Demand: For each arm k, sample $\theta_k(t)$ from a $Beta(R_k(t-1)+1, N_k(t-1) - R_k(t-1) + 1)$ distribution.
2. Select Arm: The decision maker pulls an arm $\hat{k} \in \arg\max_k \theta_k(t)$ and receives reward.
3. Update: Let $N_{\hat{k}}(t) \leftarrow N_{\hat{k}}(t-1)+1$. Let $R_{\hat{k}}(t) \leftarrow R_{\hat{k}}(t-1)+1$ if the reward is 1, and $R_{\hat{k}}(t) \leftarrow R_{\hat{k}}(t-1)$ if the reward is 0.