**OPERATIONS RESEARCH**
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0030-364X | EISSN 1526-5463 | 00 | 0000 | 0001

**INFORMS**
DOI 10.1287/xxxx.0000.0000


# Near-Optimal Algorithms for the Assortment Planning Problem under Dynamic Substitution and Stochastic Demand

Vineet Goyal
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY. vgoyal@ieor.columbia.edu

Retsef Levi
Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139. retsef@mit.edu

Danny Segev
Department of Statistics, University of Haifa, Haifa 31905. segevd@stat.haifa.ac.il

Assortment planning of substitutable products is a major operational issue that arises in many industries, such as retailing, airlines and consumer electronics. We consider a single-period joint assortment and inventory planning problem under dynamic substitution with stochastic demands, and provide complexity and algorithmic results as well as insightful structural characterizations of near-optimal solutions for important variants of the problem. First, we show that the assortment planning problem is NP-hard even for a very simple consumer choice model, where each consumer prefers only two items. In fact, we show that the problem is hard to approximate within a factor better than $1 - 1/e$. Secondly, we show that for several interesting and practical customer choice models, one can devise a *polynomial-time approximation scheme* (PTAS), i.e., the problem can be solved efficiently to within any level of accuracy. To the best of our knowledge, this is the first efficient algorithm with provably near-optimal performance guarantees for assortment planning problems under dynamic substitution. Quite surprisingly, the algorithm we propose stocks only a *constant* number of different product types, depending only on the desired accuracy level. This provides an important managerial insight that assortments with a relatively small number of product types can obtain almost all of the potential revenue. Furthermore, we show that our algorithm can be easily adapted for more general choice models, and present numerical experiments to show that it performs significantly better than other known approaches.

*Key words*: assortment planning; dynamic substitution; polynomial time approximation schemes

## 1. Introduction

Assortment planning is a major operational issue that arises in many industries, such as retailing, airlines, and consumer electronics. Given a set of products that are differentiated by price, quality and possibly other attributes, one has to decide on the product assortment, and the respective quantities that will be stocked and offered to customers. Such decisions become particularly important when different products are *substitutable* and customers exhibit a *substitution behavior*. For example, customers may a-priori prefer product A to product B, but may still be willing to buy product B if product A is not offered or not available anymore. The substitution behavior can be *assortment-based* (or *static*), i.e., unaffected by the availability of products and depends only on the specific assortment of products, or it can be *stock-out-based* (or *dynamic*), i.e., driven by stock-out events and availability of products. When customers exhibit substitution behavior (static or dynamic), the demands for different product types are correlated, and accounting for product substitutability can lead to significantly higher revenues and profits. However, this requires joint multi-product assortment and inventory decisions, which usually give rise to complex optimization models that are computationally challenging. Assortment planning under substitution forms one of the core problem domains in revenue management, and many variants of these problems have been studied extensively in the literature.

We consider a single-period joint assortment and inventory planning problem with stochastic demand and dynamic substitution. Specifically, we study a single period model with finite number of product types

(or item types), each with a per-unit selling price and potentially other attributes that differentiate between different product types (e.g., quality, size, color). At the beginning of the period, one has to decide jointly on the assortment and the inventory levels, i.e., which product types to offer and how many units to stock from each offered product, subject to a *capacity constraint* on the total number of units that can be stocked. After the assortment and inventory decisions are made, a stochastic number of customers arrive one after the other, each with a *random preference* on the product types. A preference is an ordered list or permutation of product types that reflects the order in which the customer prefers different product types. Note that the no-purchase alternative can appear at any position in the preference. No-purchase alternative at any position denotes that buying nothing is more preferable to buying any product below this position. We assume that the preference of each customer is independent and identically distributed according to a known distribution over all potential preferences. Upon arrival, each customer purchases the first available product type in her preference list. If no product on the customer's list is available, the customer leaves without purchasing any product. The goal is to find the assortment and inventory levels that maximize the expected revenue obtained from the units purchased by customers. Note that both the number of customers and their respective preferences are stochastic in this model and the customers arrive in a sequential manner.

### 1.1. Our Results

**Complexity and approximability.** The model described above is in general computationally intractable. Specifically, we show that the above mentioned problem is *NP-hard* even for the special case when there is only one customer and all preference lists consist of only two product types, i.e., there is no efficient algorithm for solving the problem to optimality, unless $\mathrm{P} = \mathrm{NP}$. Therefore, it is only natural to seek for approximation algorithms that compute near-optimal solutions to the problem. However, even the model with a single customer (but with general preference lists) can be shown to be hard to approximate within some fixed constant. In particular, there is no polynomial-time algorithm that is guaranteed to recover at least $1 - 1/e$ fraction of the optimal expected revenue for all possible instances of the problem unless $\mathrm{P} = \mathrm{NP}$. Therefore, the worst-case approximation factor for any efficient algorithm can not be better than $1 - 1/e$, unless $\mathrm{P} = \mathrm{NP}$.

**Polynomial time approximation scheme.** In view of these hardness of approximation results, we study this problem with two additional assumptions that still capture important practical situations. We first focus attention on *nested* preference lists. Here, the product types are ordered by increasing per-unit selling price. Customers always prefer the cheapest product type available upon arrival. Each preference list corresponds to a different price threshold, and a customer with that preference list is willing to buy all product types with per-unit price lower than the threshold. Nested preference lists arise in situations where the quality of different product types is similar, and customers differentiate only by price. (Similar choice models have been studied by Talluri and van Ryzin (2004).) The second assumption is that the number of customers follows an increasing failure rate (IFR) distribution. (For a definition of an IFR distribution see Shaked and Shanthikumar (1994) and Assumption 2 in Section 4.) This is a well-known class of distributions that includes many of the traditional distributions used in the operations research and operations management literature.

Interestingly, by imposing the above-mentioned assumptions, one can design a *polynomial time approximation scheme* (PTAS) for the problem, i.e., for any accuracy level $0 < \epsilon < 0.5$, one can compute a solution with expected revenue at least $(1 - \epsilon)$ of the optimal expected revenue, in time that is polynomial in the input size for any fixed $\epsilon$. Practically speaking, this result implies that the problem can be solved efficiently to within any degree of accuracy. This stands in contrast to the inapproximability result, stating that there is no approximation algorithm for the general model with worst-case performance guarantee better than $1 - 1/e$ unless $\mathrm{P} = \mathrm{NP}$. To the best of our knowledge, this is the first efficient algorithm with provably-good performance guarantees for assortment planning problems under dynamic substitution. Moreover, our algorithm stocks only a *constant* number of different product types, depending only on the desired accuracy

level $\epsilon$, and not on any other parameter of the problem, including the overall number of product types or the capacity. This provides an important managerial insight that assortments with a relatively small number of product types can obtain almost all of the potential revenue.

There have been relatively few approximation schemes for stochastic optimization models (for example, see the recent results in Halman et al. (2009, 2008)). Most of these results are based on formulating the respective problem as a dynamic program that can be solved in pseudo-polynomial time, and then employ that to devise a *fully polynomial approximation algorithm* (FPTAS). (The running time of an FPTAS depends polynomially on $1/\epsilon$, compared to a PTAS where it can depend on an arbitrary function of $1/\epsilon$.) In contrast, the model studied in this paper does not seem to admit a tractable dynamic program. Instead, we use several structural properties of a near-optimal solution to identify a subset of product types of constant size, which in turn leads to a PTAS. This concept has previously been applied to other combinatorial optimization problems (see, for instance, de la Vega and Lueker (1981), Har-Peled (2011)), but, to our knowledge, not to stochastic optimization problems. We believe that the new ideas introduced in this paper may also be applicable in other substitution and revenue management models, and more generally, stochastic optimization models.

We also show that if the distribution of the number of customers is not IFR, small subsets of product types cannot guarantee near-optimal performance. In particular, we construct an example, in which any solution that stocks a constant number of product types performs arbitrarily bad compared to the optimal policy.

**Extensions and computational experiments.** By employing dynamic programming techniques, we show how to leverage our PTAS to a more general choice model, in which customer choices are affected not only by price but also by quality. Specifically, each customer is willing to buy products within a certain *quality category*, and within that category, product types are differentiated only by price and admit a nested form. The underlying assumption is that the prices of different quality categories are separated to different customer segments. This choice model captures several important practical settings.

In addition, a natural question to explore is whether our approach to compute a near-optimal solution for nested customer preferences, can be extended to handle more general choice models. Even though one can artificially create worst-case examples showing that we may end up with highly sub-optimal solutions, it is of particular interest to examine how our algorithm scales up in practice against adaptations of existing methods. In this context, we describe computational experiments that evaluate the performance of assortments consisting of very few product types, demonstrating that our approach is indeed suitable for fairly general choice models.

**Maximizing revenue vs. maximizing profit.** We would like to emphasize that we consider a model with a capacity constraint on the total number of units being stocked and assume that there are no per-unit purchasing costs. This is in contrast to several dynamic substitution models considered in the literature where each product has a per-unit revenue and a per-unit purchasing cost and the goal is to maximize the expected profit (for instance, see Mahajan and van Ryzin (2001) and Netessine and Rudi (2003)). Capacity constraints arise in many retailing settings, for example, when there is limited shelf space. Ignoring purchasing prices is an appropriate assumption when the cost of buying or producing the products is a sunk cost (e.g., seats in an airplane), or when the costs are identical (e.g., fashion industry), or when unsold products can be fully or almost fully salvaged. Therefore, it is important to study models with capacity constraints instead of modeling purchasing costs. It is worth pointing out that the objective of maximizing revenue (instead of profit) has also been studied is a recent paper by Fisher and Vaidyanathan (2007).

On top of the practical motivation, we show that our method can be extended to compute near-optimal solutions for models with purchasing costs. This result is obtained by essentially reducing such models to a multi-capacitated setting, where carefully-picked subsets of products are given separate capacity constraints. More specifically, under the technical assumptions used to design the above-mentioned approximation scheme, suppose in addition that we are given a budget constraint of $B$ on the total purchasing cost, under which the optimal inventory levels guarantee an expected revenue of $R^*$. Then, by following

the approximation algorithm we sketch in Appendix A.1, one can compute an inventory vector in which the expected revenue is at least $(1-\epsilon)R^*$, without exceeding the total purchasing cost. That being said, the downside of this extension is that the resulting algorithm is not longer a PTAS, but rather a quasi-PTAS (where the running time exponent also involves factors that are polylogarithmic in the input size; see, for instance, Bansal et al. (2006), Remy and Steger (2009), Chan and Elbassioni (2011)).

### 1.2. Literature Review

Joint assortment planning and inventory management problems with substitution have been extensively studied; we refer the reader to directly related papers, surveys, and books (Kok et al. (2006), Lancaster (1990), Ho and Tang (1998), Ramdas (2003)) for a comprehensive review of the recent literature. Pentico (1974) was one of the earliest to consider an assortment planning problem with *downward substitution* and a deterministic sequence of customer arrivals, showing that under several conditions, a certain planning-horizon type policy is optimal. Van Ryzin and Mahajan (1999) consider a static substitution model with *multi-nomial logit* (MNL) demand distributions. They show that in this model the optimal solution consists of the most popular product. Cachon et al. (2005) generalize this static substitution model to incorporate search costs. Hopp and Xu (2005) consider the problem of integrating assortment decisions with pricing decisions, again with MNL demands. Anupindi et al. (2006) consider a *probit* demand model, and include a penalty for customer's disutility in substituting to a less preferred product type.

The above-mentioned papers are primarily focused on static substitution models, where customer preferences and purchasing decisions depend only on the assortment being offered, but not on the specific inventory levels observed at the time of purchase. Specifically, customers do not substitute to other product types just because more preferred types are stocked out. Thus, the demand for each product type is independent of the actual inventory levels of other products, and only depends on the subset of product types being offered. However, in many practical applications such as airlines and retailing, customers do not exhibit static substitution behavior, but rather a dynamic substitution. In particular, customers readily substitute when a more preferred product type is stocked out. Models with dynamic substitution are generally more complex and challenging. The demand for a specific product type is affected not only by the assortment being offered, but also by the respective inventory levels that change dynamically over time as customers arrive and consume products. As a result, assortment and inventory decisions must be made simultaneously.

While the static substitution model has been extensively studied, there is relatively little work on dynamic substitution models. Parlar and Goyal (1984) were the first to study a dynamic substitution model. They consider a probabilistic substitution model and show that the profit function is concave for a wide range of problem parameters. Smith and Agrawal (2000) consider a dynamic substitution model specified by first-choice probabilities and a substitution matrix, and show that static substitution yields bounds on the demand for each product in the dynamic substitution case. Mahajan and van Ryzin (2001) study a joint assortment and inventory planning problem with stochastic demands and general preferences where each product type has per unit revenue and cost and the goal is to maximize the expected profit. Assuming that customer sequences can be sampled, they propose a sample path gradient-based algorithm, and show that under fairly general conditions it converges to a local maximum. However, they do not provide any performance bounds for the expected profit of a local maximum as compared to the optimal expected profit.

Netessine and Rudi (2003) consider a substitution model where each customer preference consists of only two products (a first-choice as well as a second-choice product), in the assortment of an arbitrary number of products, and obtain analytically tractable solutions for the assortment planning problem in both centralized inventory management and competition. It is worth pointing out that we prove the assortment planning problem (in our model) to be NP-hard even when each customer preference consist of only two products. However, our model is different from the model studied by Netessine and Rudi (2003); they consider an uncapacitated problem with a per-unit purchasing cost for each product type and stocking a fractional number of any product type is allowed. Whereas, we consider a model with a capacity constraints on the total number of units that can be stocked across all product types. As mentioned earlier, it is more

useful to consider a model with capacity constraints instead of purchasing costs for some applications (for instance, airline seats and fashion industry). Moreover, we also show that a model with purchasing costs can be approximately reduced to a problem with only capacity constraints (and no purchasing costs).

Kok and Fisher (2007) assume an MNL demand model within a Bayesian framework, propose an algorithm to estimate the model parameters, and also solve the assortment and planning problem with one-level stock-out based substitution. Gaur and Honhon (2006) give a heuristic for the problem under a location choice model based on the solution of the static substitution case. Honhon et al. (2007) consider a general customer choice model with a stochastic demand but the sequence of customer preferences satisfies the following property: if a certain preference list occurs with probability $p$, then for every sequence of $1/p$ customers there will be one customer with this preference list. Therefore, the customer choices are not completely random in their model. The authors provide a novel characterization of the local maxima, and propose a dynamic programming based algorithm to solve the problem. However, the running time of this algorithm is exponential in the number of product types, implying that it is practical only in cases where the number of product types is small. Nagarajan and Rajagopalan (2008) consider a dynamic substitution model where the individual demands for different products are negatively correlated, and show that a *partially-decoupled* policy is optimal under fairly general conditions. Chen and Bassok (2008) study the problem under a general customer choice model where all product types have identical prices and costs. Assuming static allocation (i.e., one see all customers first and decides how to allocate the inventories), they show that if the number of customers is fixed, then with high probability all the demand can be satisfied by stocking a total of $N$ units of the products even when the customer preferences are random. However, the authors assume that the allocation of products to customers is simultaneous instead of sequential. That is, the products can be allocated by the retailer after all customers have arrived and their preferences become known to the retailer, as long as the product appears in the preference list of the customer. Using Normal approximations, the authors also argue that, for a sufficiently large number of customers, the static allocation model is a good approximation for the sequential allocation model, in which customers arrive one after the other and pick the most preferred product type among the available ones.

**Outline**. The rest of this paper is organized as follows. Section 2 provides a mathematical formulation of the model. The hardness results are discussed in Section 3. In Section 4, we describe the PTAS and establish its performance analysis. In Section 5, we discuss the extension to the more general model of quality categories. In Section 6, we show that when the number of customers follows a non-IFR distribution, structural results similar to those in Section 4 need not hold. We conclude by presenting extensive computational experiments in Section 7, where the performance of our approach is tested when applied to more general settings.

## 2. Model Formulation

Consider $n$ product types with per-unit selling prices $p_1 \leq \cdots \leq p_n$, respectively, and a capacity bound $C$ on the total number of units (across all product types) that can be stocked. A random number of customers, say $M$, arrive one after the other, where $M$ is random variable with a known distribution. Each customer $j = 1, \ldots, M$ has a random preference list $L_j$, which specifies a subcollection of products in decreasing order of preference. This list comes from a known distribution and is independent and identically distributed among customers and also independent of the number of customers $M$. Upon arrival, each customer purchases the first available product in her list, assuming that at least one unit of such products exists at that time; otherwise, the customer leaves without purchasing at all. For an inventory vector $(y_1, \ldots, y_n)$, where $y_i$ specifies the number of units stocked from product type $i$, let $R_m(y_1, \ldots, y_n)$ denote the revenue attained if $m$ customers arrive. Observe that this revenue is still random, due to stochasticity in the preference lists of customers. The objective is to determine the inventory level of each product type, subject to the capacity constraint on the total number of units being stocked, so that the expected revenue is maximized, i.e.,

$$\max_{(y_1,\ldots,y_n)\in\mathbb{Z}_+^n} \left\{ \mathrm{E}[R_M(y_1,\ldots,y_n)] : \sum_{i=1}^{n} y_i \leq C \right\} .$$

Note that the expectation is taken with respect to the number of customers $M$ and their stochastic preference lists. We will use $(y_1^*, \ldots, y_n^*)$ to denote the optimal solution and OPT to denote the optimal expected revenue.

## 3. Hardness Results

In this section, we show that the capacitated assortment problem with general preferences is NP-hard even when there is only one customer (i.e., $M = 1$, deterministically) and all possible preferences include only two product types. Our hardness proof makes use of a reduction from the vertex cover problem.

THEOREM 1. *The capacitated assortment planning problem is NP-hard even when there is only one customer and all possible preferences include only two product types.*

*Proof.* In the vertex cover problem, we are given an undirected graph $G = (V, E)$ and an integer parameter $k$. The objective is to decide whether there exists a subset of vertices $V' \subset V$ of cardinality at most $k$, such that each edge $e \in E$ is incident on some vertex in $V'$. This problem is known to be NP-hard (see, for instance, problem [GT1] in Garey and Johnson (1979)). Consider a vertex cover instance $I$, with $V = \{v_1, \ldots, v_n\}$. We construct an instance $I'$ of the assortment problem as follows. Let the number of product types be $n = |V|$, one corresponding to each vertex; all products have identical prices, $p_1 = \cdots = p_n = 1$. In addition, let the set of preference lists be $\{(i, j) : (v_i, v_j) \in E\}$, where a preference list $(i, j)$ implies that the first choice product is $i$ and the second choice is $j$. Finally, there is a single customer, the capacity on the total number of units to be stocked is $k$, and each preference list occurs with probability $1/|E|$.

We show that the optimal expected revenue of the assortment problem instance $I'$ is exactly 1 if and only if there is a vertex cover of size at most $k$ in instance $I$. Suppose there is a vertex cover $V' \subset V$ such that $|V'| \leq k$. Consider the inventory vector $\bar{y} = (y_1, \ldots, y_n)$, where $y_i = 1$ if $v_i \in V'$ and $y_i = 0$ otherwise. In other words, we stock a single unit of product $i$ if and only if $v_i$ is part of the vertex cover $V'$. Clearly, the total number of units stocked is at most $k$. Now,

$$\mathrm{E}\left[R_1(\bar{y})\right] = \frac{1}{|E|} \sum_{(v_i, v_j) \in E} \max\{y_i, y_j\} \; .$$

Since $V'$ is a vertex cover, for each edge $(v_i, v_j) \in E$, at least one of $v_i$ or $v_j$ is in $V'$. Therefore, $\max(y_i, y_j) = 1$ for all $(v_i, v_j) \in E$, and $\mathrm{E}\left[R_1(\bar{y})\right] = 1$.

Conversely, consider an inventory vector $\bar{y} = (y_1, \ldots, y_n)$ that contains at most $k$ units and has expected revenue 1. We proceed by arguing that $V' = \{v_i : y_i \geq 1\}$ is a vertex cover, noting in advance that this set clearly consists of at most $k$ vertices. For this purpose, note that the expected revenue of $\bar{y}$ can be written as

$$\mathrm{E}\left[R_1(\bar{y})\right] = \frac{1}{|E|} \sum_{(v_i, v_j) \in E} \min\{\max(y_i, y_j), 1\}.$$

Since $\mathrm{E}[R_1(\bar{y})] = 1$, it follows that $\max(y_i, y_j) \geq 1$ for all edges $(v_i, v_j) \in E$, implying that at least one of the vertices $v_i$ and $v_j$ belongs to $V'$, making the set $V'$ a vertex cover. □

We further extend the reduction in Theorem 1, and show that it is NP-hard to approximate the assortment planning problem within a factor larger than $1 - 1/e$, even when there is only one customer. In this case, however, customer preferences are not restricted to two products. Our proof is based on an approximation-preserving reduction from the maximum coverage problem. The latter is NP-hard to approximate within a factor better than $1 - 1/e$ (Feige (1998)).

THEOREM 2. *It is NP-hard to approximate the capacitated assortment problem within a factor better than $1 - 1/e$, even when there is only one customer.*

*Proof.* In the maximum coverage problem, we are given a ground set of elements $U$, a set family $\mathcal{F} \subseteq 2^U$, and an integer $k$. The goal is to pick a subset $T \subseteq U$ of at most $k$ elements such that $\text{cov}(T) = |\{S \in \mathcal{F} : S \cap T \neq \emptyset\}|$ is maximized.

Consider an instance $I$ of the maximum coverage problem. We construct an instance $I'$ of the capacitated assortment problem as follows. Let the number of product types be $n = |U|$, one corresponding to each element in $U$; all products have a uniform price of $p_1 = \cdots = p_n = 1$. In addition, construct $|\mathcal{F}|$ preference lists, one corresponding to each set $S_j \in \mathcal{F}$, and let $\mathcal{P}_j$ denote the preference list corresponding to $S_j$. More specifically, suppose $S_j = \{e_{j_1}, \ldots, e_{j_t}\}$, where the element order is arbitrary. Then, the preference $\mathcal{P}_j$ is given by $\mathcal{P}_j = (j_1, \ldots, j_t)$, where $(j_1, \ldots, j_t)$ denote the order of preference for products corresponding to elements in set $S_j$, with $j_1$ being the most preferred product and $j_t$ being the least preferred. Finally, each preference list occurs with probability $1/|\mathcal{F}|$, there is a single customer, and the capacity on the total number of products is $k$.

Consider any subset $T \subseteq U$, $|T| = k$, obtaining the maximum number of covered sets in the instance $I$, and let $\bar{y} = (y_1, \ldots, y_n)$ be the inventory vector defined by $y_i = 1$ if $e_i \in T$ and $y_i = 0$ otherwise. In other words, we stock a single unit of product type $i$ if and only if $e_i$ belongs to $T$. Clearly, the number of units stocked is $\sum_{i=1}^{n} y_i = |T| = k$. We now argue that $\bar{y}$ generates an expected revenue of $\text{cov}(T)/|\mathcal{F}|$. To this end, for any subset $S \subset U$, let $\chi_T(S)$ indicate whether $S$ and $T$ have a non-empty intersection, i.e., $\chi_T(S) = 1$ if $S \cap T \neq \emptyset$ and $\chi_T(S) = 0$ otherwise. Then,

$$\text{E}\left[R_1(\bar{y})\right] = \sum_{S_j \in \mathcal{F}} \Pr[L_1 = \mathcal{P}_j] \cdot \text{E}\left[R_1(\bar{y}) | L_1 = \mathcal{P}_j\right] = \sum_{S_j \in \mathcal{F}} \frac{1}{|\mathcal{F}|} \cdot \chi_T(S_j) = \frac{\text{cov}(T)}{|\mathcal{F}|} \ .$$

Therefore, to complete the proof it is sufficient to show that, given an inventory vector $\bar{y} = (y_1, \ldots, y_n)$ that contains at most $k$ non-zero entries, we can construct a set $T \subseteq U$, $|T| \leq k$, such that $\text{cov}(T) = |\mathcal{F}| \cdot \text{E}\left[R_1(y)\right]$. For this purpose, let $T = \{e_i : y_i \geq 1\}$. Then, clearly $|T| \leq k$, and

$$\text{cov}(T) = \sum_{S_j \in \mathcal{F}} \chi_T(S_j) = |\mathcal{F}| \cdot \text{E}\left[R_1(y)\right] \ .$$

□

## 4. Polynomial Time Approximation Scheme

In Section 3, we show that the general model described in Section 2 is hard to approximate beyond a certain degree of accuracy. In what follows, we consider a special case of the model obtained by imposing additional assumptions and give a polynomial time approximation scheme (PTAS) for computing near-optimal solutions with arbitrary level of accuracy. For any fixed $0 < \epsilon < 0.5$, our algorithm computes an inventory vector that obtains a fraction of at least $(1-\epsilon)$ of the optimal expected revenue, in time polynomial in the input size. We proceed by listing the additional assumptions.

*Assumption 1: Nested preference lists.* We assume that the common distribution from which customers pick their preference lists consists of only lists of the form $L = (1, \ldots, \ell)$ for some $\ell \leq n$. Therefore, there are $n+1$ possible preferences lists, $(), (1), (1,2), \ldots, (1,2,\ldots,n)$, but their respective probabilities can be arbitrary (here, $()$ denotes the empty preference list for customers who prefer the no-purchase alternative to any product type). We refer to such preference lists as *nested lists*. Recall that the products are indexed such that $p_1 \leq \cdots \leq p_n$. (Note that, with the nested lists assumption in place, the latter assumption is without loss of generality: if for some $i_1 < i_2$ we have $p_{i_1} > p_{i_2}$, there is no motivation to stock $i_2$ at all, and this product type can therefore be eliminated). Let $\alpha_i$ denote the probability that the preference list picked from this distribution contains the product $i$, i.e.,

$$\alpha_i = \sum_{\ell=i}^{n} \Pr[L = (1, \ldots, \ell)] \ .$$

It is easy to verify that $\alpha_1 \geq \cdots \geq \alpha_n$. In Section 5, we show that our results extend to a more general setting, in which customer preferences are affected by both quality and price. In particular, the product types are partitioned into quality categories. Customer preferences are characterized by a given quality category, whereas within a given category, the preference lists are nested.

*Assumption 2: IFR.* We assume that the distribution of the number of customers, $M$, has an increasing failure rate (IFR). An integer-valued random variable $X$ is said to be IFR if $\Pr[X = k]/\Pr[X \geq k]$ is non-decreasing over the integer domain. It can be proven (see, for instance, Chapter 1 of Shaked and Shanthikumar (1994)) that this definition is equivalent to requiring that the sequence of random variables $[X - k|X \geq k]_{k\in\mathbb{Z}}$ is stochastically non-increasing in $k$. For definitions of stochastic order and stochastic monotonicity, see Shaked and Shanthikumar (1994) and Definition 2 in Section 4.5.

*Assumption 3: Revenue evaluation.* Given an inventory vector $(y_1, \ldots, y_n)$, there is a polynomial-time procedure for computing $\mathrm{E}[R_M(y_1, \ldots, y_n)]$, possibly up to a multiplicative error of $(1 - \epsilon)$.

In Appendix A.2, we show that when the distribution of the number of customers $M$ has finite support, say on $\{0, \ldots, T\}$, then the expected revenue of any given vector $(y_1, \ldots, y_n)$ can be computed (exactly) by means of dynamic programming in time polynomial in $n$ and $T$. We also explain, in the general case, how to efficiently make the support of $M$ finite, at the cost of $\epsilon$-approximating $\mathrm{E}[R_M(y_1, \ldots, y_n)]$ instead of exactly computing this quantity.

### 4.1. An Overview of the Analysis

For ease of exposition, we first provide a high-level overview of the PTAS and its worst-case analysis. We show that, for any accuracy level $0 < \epsilon \leq 0.5$, there exists a solution that stocks only a subset of product types of size $O(\log(1/\epsilon))$, obtaining at least $(1-\epsilon)$ fraction of the optimal revenue. Note that the size of this product set depends on the accuracy level $\epsilon$, but not on any other parameter of the problem. In particular, we establish the following theorem.

THEOREM 3. *For any accuracy level $0 < \epsilon \leq 0.5$, there is an inventory vector $(y_1, \ldots, y_n)$ that satisfies the following properties:*

1. *The total capacity is at most $C$, that is, $\sum_{i=1}^{n} y_i \leq C$.*
2. *The number of non-zero coordinates in $(y_1, \ldots, y_n)$ is only $O(\log(1/\epsilon))$.*
3. $\mathrm{E}[R_M(y_1, \ldots, y_n)] \geq (1 - \epsilon)\mathrm{OPT}$.

Furthermore, our proof gives a constructive method to efficiently identify the corresponding subset of $O(\log(1/\epsilon))$ product types. As a result, one can naively enumerate all possible solutions consisting of these product types, and take the best among those solutions (as mentioned in Assumption 3, the expected revenue of any given solution can be evaluated efficiently). The overall number of such solutions is $O(C^{O(\log(1/\epsilon))})$ since there are a total of $C + 1$ possible choices for the number of units to stock for each of the $O(\log(1/\epsilon))$ product types. However, this is only pseudo polynomial in the size of the input. As explained in Section 4.6, by employing suitable discretization, this can be improved to enumerating only $O((\log C)^{O(\log(1/\epsilon))})$, which is indeed polynomial in the input size.

To obtain a high-revenue subset of product types of size $O(\log(1/\epsilon))$, as described in Theorem 3, we partition the product types into *frequent* and *rare* product types. For any product type $i$, let $X_i$ denote the random number of customers whose preference list contains product type $i$.

DEFINITION 1. A product type $i$ is called *frequent* if the expected number of units sold when $C$ units of product type $i$ are stocked is at least $\epsilon^2 C$, i.e.,

$$\mathrm{E}[\min(X_i, C)] \geq \epsilon^2 C.$$

Otherwise, the product type $i$ is referred to as *rare*.

*Main idea 1: $O(\log(1/\epsilon))$ frequent product types are sufficient.* Since $\alpha_i \geq \alpha_{i'}$ whenever $i < i'$, it follows that the frequent product types must be $1, \ldots, F$, for some $0 \leq F \leq n$, whereas the rare products are $F + 1, \ldots, n$. In Lemma 1 we show that, by applying appropriate truncation and discretization, it is possible to efficiently identify a small subset of frequent product types of size $O(\log(1/\epsilon))$ that obtains a fraction of at least $(1-\epsilon)^3$ of the expected revenue obtained by frequent product types in the optimal solution. (In fact, this would be true for any solution.) The proof of Lemma 1 relies only on the fact that preferences are nested (see Assumption 1).

*Main idea 2: One rare product type is sufficient.* The analysis is completed by showing that a single rare product type can be used to attain at least $(1-\epsilon)$ of the expected revenue obtained by rare product types in the optimal solution. This part of the analysis relies on several central ideas. The first idea is to establish an upper bound on the total achievable expected revenue. This upper bound is derived by considering an *uncapacitated* variant of the problem where there is no capacity constraint and one is allowed to stock any number of units. The uncapacitated variant is clearly a relaxation of the original model, and provides an upper bound on the optimal revenue. In Theorem 4 we show that, in the uncapacitated variant, it is optimal to stock only one product type. Specifically, it is optimal to stock $M$ units (the number of customers is known in advance) of the *maximal product type*, which is the product type $i$ that maximizes the expected marginal revenue $\alpha_i p_i$ from a single customer. The uncapacitated variant is discussed in Section 4.2 below.

Observe that if one stocks $C$ units of a given rare product type, the expected number of units sold is much smaller than the capacity, specifically, less than $\epsilon^2 C$. Intuitively, this observation implies that if one only considers rare product types, the resulting problem is 'almost' uncapacitated, and hence stocking only the maximal product type among rare product types should be near optimal. However, it turns out that this intuition is incorrect in general, unless the distribution of the number of customers $M$ satisfies certain properties. In particular, we prove in Lemmas 2, 3 and 4 that this intuition is indeed valid when the distribution of $M$ is IFR (see Assumption 2 above). This implies that there exists a solution that stocks only $O(\log(1/\epsilon))$ product types and obtains a fraction of at least $(1-\epsilon)^4$ of the optimal revenue. On the other hand, in Section 6, we demonstrate that the latter property does not hold for general distributions of $M$.

The extension of the PTAS to the more general model, with preference lists that capture both price and quality, is discussed in Section 5. In this case, the central idea is to employ the PTAS for nested preference lists as an auxiliary subroutine within a dynamic programming approach.

### 4.2. The Uncapacitated Deterministic-$M$ Problem

In this section, we consider the uncapacitated variant, in which the number of customers is known in advance (i.e., $M$ is deterministic) and there is no constraint on the number of units to be stocked. We show that, for this variant, it is optimal to stock only the product type that maximizes the marginal expected revenue, which is the expected revenue from a single customer if only one unit of the product type is stocked and only one customer arrives. The marginal expected revenue of each product type is exactly the probability that a customer is willing to purchase that product type times the respective per-unit price, i.e., $\alpha_i p_i$. We call this product type the maximal product type and denote it by $i^*$. That is, $i^* = \arg\max_i \alpha_i p_i$. We establish the proof for each fixed $m \in \mathbb{Z}_+$.

THEOREM 4. *For any number of customers, $m \in \mathbb{Z}_+$, it is optimal to stock $m$ units of the maximal product type $i^*$. The resulting optimal expected revenue is $m\alpha_{i^*}p_{i^*}$.*

*Proof.* First observe that since the preference lists are nested, there is no benefit in stocking more than $m$ units. Consider now any solution $y = (y_1, \ldots, y_n)$. For each customer $j = 1, \ldots, m$, let $V_j = V_j(y)$ be the revenue generated by that customer, and let $I_j = I_j(y)$ be the cheapest product type available upon the arrival of customer $j$; note that the distributions $V_j$ and $I_j$ are solution dependent. Since there are $m$ customers and at least $m$ units there will always be some product type available upon the customer arrival. Observe that $\mathrm{E}[V_j] = \mathrm{E}[\mathrm{E}[V_j|I_j]] = \mathrm{E}[\alpha_{I_j} p_{I_j}] \leq \alpha_{i^*} p_{i^*}$. It follows that the total revenue of any solution is upper bounded by $m\alpha_{i^*}p_{i^*}$. However, the solution that stocks $m$ units of product type $i^*$ has expected revenue $m\alpha_{i^*}p_{i^*}$, so it must be optimal. □

Since the number of customers is known in advance, stocking $M$ units of the maximal product type $i^*$ is optimal for the uncapacitated variant and the optimal expected revenue is $\mathrm{E}[M]\alpha_{i^*}p_{i^*}$. Since the uncapacitated variant is a relaxation of the capacitated problem, we conclude that the optimal expected revenue of the original capacitated model is upper bounded by $\mathrm{E}[M]\alpha_{i^*}p_{i^*}$. Moreover, the following is an immediate interesting corollary.

COROLLARY 1. *Consider the capacitated model described in Section 2, with Assumption 1. Suppose that the capacity $C$ is larger or equal to the maximal possible value that the number of customers $M$ can attain. Then, it is optimal to stock $C$ units of the maximal product type $i^*$.*

### 4.3. Frequent Product Types

Recall that, a product type $i$ is called *frequent*, if the expected number of units sold, assuming only $C$ units of this product type are stocked, is at least $\epsilon^2 C$. That is, $\mathrm{E}[\min\{X_i, C\}] \geq \epsilon^2 C$, where $X_i \sim B(M, \alpha_i)$ is the number of customers willing to buy product type $i$. Note that, conditioning on $[M = m]$, the random variable $X_i$ follows a Binomial distribution with parameters $(m, \alpha_i)$. Due to the nested preferences assumption, frequent product types can be numbered as $1, \ldots, F$, and rare product types by $F+1, \ldots, n$, for some $0 \leq F \leq n$. In the remainder of this section, we consider an optimal inventory vector,

$$(\underbrace{y_1^*, \ldots\ldots, y_F^*}_{\text{frequent}}, \underbrace{y_{F+1}^*, \ldots\ldots, y_n^*}_{\text{rare}}).$$

The next lemma shows that there exists a subset of frequent product types of size $O(\log(1/\epsilon))$, which obtains a fraction of at least $(1-3\epsilon)^3$ of the expected revenue obtained by frequent product types in the optimal solution. Furthermore, if the optimal solution stocks more than one unit of rare product types, one can ensure that a capacity of at least $\epsilon C$ is allocated to rare product types (i.e., the total number of units of frequent product types does not exceed $(1-\epsilon)C$). Intuitively, we wish to make sure that the number of units of rare product types that are being sold in expectation (at most $\epsilon^2 C$) is small relative to the capacity allocated to them (at least $\epsilon C$). This property will enable us to use the uncapacitated bound derived in Section 4.2.

LEMMA 1. *Let $(y_1^*, \ldots, y_n^*)$ be an optimal inventory solution. For any $0 < \epsilon < 0.5$, there exists a feasible inventory vector $(y_1, \ldots, y_F)$ of frequent product types with only $O(\log(1/\epsilon))$ non-zero coordinates, such that:*

1. $\mathrm{E}[R_M(y_1, \ldots, y_F, y_{F+1}^*, \ldots, y_n^*)] \geq (1-3\epsilon)^3 \cdot \mathrm{E}[R_M(y_1^*, \ldots, y_n^*)] = (1-3\epsilon)^3 \cdot \mathrm{OPT}$.
2. *If* $\sum_{i=F+1}^{n} y_i^* \geq 2$ *then* $\sum_{i=1}^{F} y_i \leq (1-\epsilon)C$.

*Proof.* We present a constructive proof that shows how to efficiently identify the corresponding subset of frequent product types.

*Phase 1: Eliminating cheap products.* Consider the frequent product types $1, \ldots, F$, and recall that $p_F$ is the price of the most expensive frequent product type. We begin by arguing that it is possible to discard cheap product types from $(y_1^*, \ldots, y_F^*)$, without losing too much revenue in expectation. A product type $i$ is called *cheap* if $p_i \leq \epsilon^3 p_F$.

We argue that if we do not purchase any cheap product type, the expected revenue reduces by at most $\epsilon\mathrm{OPT}$. To understand this claim, note that the total expected revenue from cheap product types is upper bounded by $\epsilon^3 p_F C$, since at most $C$ units of such product types could be sold, at a price of at most $\epsilon^3 p_F$ each. On the other hand, we argue that $\mathrm{OPT} \geq \epsilon^2 C p_F$. Consider the solution where we stock $C$ units of product type $F$. The expected revenue of this solution is given by,

$$\mathrm{E}[\min\{X_F, C\}] \cdot p_F \geq \epsilon^2 C p_F,$$

where the inequality follows since product type $F$ is frequent. Therefore, $\mathrm{OPT} \geq \epsilon^2 C p_F$ and by not purchasing cheap product types, one may lose up to $\epsilon^3 p_F C \leq \epsilon\mathrm{OPT}$.

*Phase 2: Picking $O(\log(1/\epsilon))$ frequent product types.* Phase 1 ensures that all remaining frequent products have per-unit price in $[\epsilon^3 p_F, p_F]$. We proceed by geometrically partitioning this interval by powers of $1+\epsilon$ to obtain $O(\log(1/\epsilon))$ subintervals: $[\epsilon^3 p_F, (1+\epsilon)\epsilon^3 p_F)$, $[(1+\epsilon)\epsilon^3 p_F, (1+\epsilon)^2\epsilon^3 p_F)$, so forth and so on. We modify the optimal solution by considering each of these intervals, and the respective product types with prices within the interval, and reallocating every unit purchased to the cheapest product type that falls within that interval. For instance, if there are 10 product types, say $t, \ldots, t+9$, in the interval $[\epsilon^3 p_F, (1+\epsilon)\epsilon^3 p_F)$, with respecting inventory levels $y_t^*, \ldots, y_{t+9}^*$, then we will stock $y_t^* + \cdots + y_{t+9}^*$ units of type $t$ (which is the cheapest one), and no units whatsoever of types $t+1, \ldots, t+9$. Due to the nested preference lists, it follows that each unit in the modified solution is now consumed with probability at least as high as before. Moreover, if and when it is indeed consumed, the resulting revenue is at least $1/(1+\epsilon) \geq (1-\epsilon)$ times its revenue prior to this transformation, since the endpoints of each price interval differ by a factor of $1+\epsilon$. The solution after this modification has only $O(\log(1/\epsilon))$ frequent product types and obtains at least $(1-\epsilon)^2$ fraction of the optimal revenue.

*Phase 3: Transferring some capacity to rare product types.* Suppose now that after Phase 2, the capacity allocated to frequent product types is larger than $(1-\epsilon)C$ and $\sum_{i=F+1}^{n} y_i^* \geq 2$. Consider the current solution $\bar{y} = (y_1, \ldots, y_n)$ obtained from the optimal solution after the modifications in Phase 1 and 2. For each unit stocked in $\bar{y}$, compute the expected contribution to the overall revenue, that is, the probability of this unit to be consumed times its per-unit price. We then discard $\sum_{i=1}^{F} y_i - \lfloor (1-\epsilon)C \rfloor$ units, choosing the ones with the smallest contribution. Since the expected revenue from the remaining units can only increase after removing these units, it follows that we lose a fraction of the expected revenue that is upper bounded by

$$\frac{\sum_{i=1}^{F} y_i - \lceil (1-\epsilon)C \rceil + 1}{\sum_{i=1}^{F} y_i} \leq \frac{\epsilon C + 1}{(1-\epsilon)C} \leq \frac{\epsilon C + \epsilon C/2}{(1-\epsilon)C} = \frac{3\epsilon}{2(1-\epsilon)} \leq 3\epsilon.$$

The first inequality follows from $\lceil (1-\epsilon)C \rceil \leq \sum_{i=1}^{F} y_i \leq C$, the second inequality holds since $\epsilon C \geq \sum_{1=F+1}^{N} y_i \geq 2$, and the last inequality since $\epsilon < 0.5$. This concludes the proof of the lemma. □

### 4.4. Rare Product Types

Lemma 1 implies that there exists an inventory vector $(y_1, \ldots, y_F, y_{F+1}^*, \ldots, y_n^*)$ achieving a $(1-3\epsilon)^3$ fraction of the optimal revenue, and that its non-zero components among $y_1, \ldots, y_F$ are contained in an $O(\log(1/\epsilon))$-sized subset of $1, \ldots, F$. The next issue is how to complete each such combination of frequent product types by augmenting it with rare product types. Clearly, if one considers a combination such that $\sum_{i=1}^{F} y_i = C - 1$, it is straightforward to compute the single rare product type that should be stocked to maximize the overall expected revenue simply by enumerating all possibilities. When $\sum_{i=1}^{F} y_i < C - 1$, by Lemma 1 we can assume that $\sum_{i=1}^{F} y_i \leq (1-\epsilon)C$. In other words, there is a capacity of at least $\epsilon C$ to stock units of rare product types. In the next section, we show that in fact one can use the entire residual capacity to stock only one rare product type, and obtain a $(1-\epsilon)$ fraction of the optimal expected revenue. This result is stated in the following lemma.

LEMMA 2. *Consider any inventory vector $(y_1, \ldots, y_F, y_{F+1}^*, \ldots, y_n^*)$ satisfying $\sum_{i=1}^{F} y_i \leq (1-\epsilon)C$. Let $i^*$ be the product type that maximizes $\alpha_i p_i$ among all rare product types, i.e., $i^* = \arg\max_{i \geq F+1} \alpha_i p_i$. Then,*

$$\mathrm{E}\left[ R_M\left( y_1, \ldots, y_F, \underbrace{0, \ldots, 0, C - \sum_{i=1}^{F} y_i, 0, \ldots, 0}_{\text{only } i^*} \right) \right] \geq (1-\epsilon) \cdot \mathrm{E}\left[ R_M(y_1, \ldots, y_F, y_{F+1}^*, \ldots, y_n^*) \right] .$$

In fact, we will show that a fraction of at least $(1-\epsilon)$ of the total expected revenue from the rare product types, in any solution in which frequent product types are allocated a total capacity of at most $(1-\epsilon)C$, can be obtained by stocking only product type $i^*$. We start by introducing some notation:

- Let $R_{\rm I}$ and $R_{\rm II}$ be the random revenues obtained from rare product types in the inventory vectors $(y_1,\ldots,y_F,y^*_{F+1},\ldots,y^*_n)$ and $(y_1,\ldots,y_F,0,\ldots,0,C-\sum_{i=1}^F y_i,0,\ldots,0)$, respectively.
- Let $Z_{i^*}$ be the random number of customers with a preference list containing product type $i^*$, arriving after all units of frequent product types have been consumed; let $Z_{i^*}=0$ if the frequent product types were not fully consumed.
- Let $\mathcal{A}$ denote the event "all units of frequent product are consumed".

Clearly, $\mathrm{E}[R_{\rm I}]=\Pr[\mathcal{A}]\cdot\mathrm{E}[R_{\rm I}|\mathcal{A}]$ and $\mathrm{E}[R_{\rm II}]=\Pr[\mathcal{A}]\cdot\mathrm{E}[R_{\rm II}|\mathcal{A}]$. Therefore, to prove Lemma 2, it is sufficient to prove that $\mathrm{E}[R_{\rm II}|\mathcal{A}]\geq(1-\epsilon)\mathrm{E}[R_{\rm I}|\mathcal{A}]$. The proof relies on two properties of the random variable $Z_{i^*}$. Specifically, we show that two properties of the random variable $M$, the total number of customers, are preserved in $[Z_{i^*}|\mathcal{A}]$. In Lemma 3 it is shown that even conditioning on the event that all units of frequent product types are consumed (i.e., the event $\mathcal{A}$), then product type $i^*$ can still be considered as a rare product type. In Lemma 4, it is proven that, like $M$, the distribution of $[Z_{i^*}|\mathcal{A}]$ is also IFR. We proceed by stating Lemmas 3 and 4; the corresponding proofs are deferred, and we first show that these lemmas can be used to establish Lemma 2.

LEMMA 3. $\mathrm{E}[\min\{Z_{i^*},C-\sum_{i=1}^F y_i\}|\mathcal{A}]\leq\epsilon^2 C$.

LEMMA 4. *The distribution of $[Z_{i^*}|\mathcal{A}]$ is IFR. In particular,* $\mathrm{E}[Z_{i^*}-k|Z_{i^*}\geq k,\mathcal{A}]\leq\mathrm{E}\left[Z_{i^*}|\mathcal{A}\right]$ *for every integer $k\geq 0$.*

*Proof of Lemma 2.* We show that Lemmas 3 and 4 imply that $\mathrm{E}[R_{\rm II}|\mathcal{A}]\geq(1-\epsilon)\mathrm{E}[R_{\rm I}|\mathcal{A}]$, from which the proof of Lemma 2 follows immediately. Note that

$$\begin{aligned}
\mathrm{E}\left[Z_{i^*}|\mathcal{A}\right] &= \mathrm{E}\left[\min\left\{Z_{i^*},C-\sum_{i=1}^F y_i\right\}+\left[Z_{i^*}-\left(C-\sum_{i=1}^F y_i\right)\right]^+\middle|\mathcal{A}\right] \qquad (1)\\
&= \mathrm{E}\left[\min\left\{Z_{i^*},C-\sum_{i=1}^F y_i\right\}\middle|\mathcal{A}\right]\\
&\quad+\Pr\left[Z_{i^*}\geq C-\sum_{i=1}^F y_i\middle|\mathcal{A}\right]\cdot\mathrm{E}\left[Z_{i^*}-\left(C-\sum_{i=1}^F y_i\right)\middle|Z_{i^*}\geq C-\sum_{i=1}^F y_i,\mathcal{A}\right]\\
&\leq \mathrm{E}\left[\min\left\{Z_{i^*},C-\sum_{i=1}^F y_i\right\}\middle|\mathcal{A}\right]+\epsilon\cdot\mathrm{E}\left[Z_{i^*}|\mathcal{A}\right]
\end{aligned}$$

The first inequality holds since by Lemma 4 we have $\mathrm{E}[Z_{i^*}-(C-\sum_{i=1}^F y_i)|Z_{i^*}\geq C-\sum_{i=1}^F y_i,\mathcal{A}]\leq\mathrm{E}\left[Z_{i^*}|\mathcal{A}\right]$ and by Lemma 3, we have $\Pr[Z_{i^*}\geq C-\sum_{i=1}^F y_i|\mathcal{A}]\leq\epsilon$, since

$$\begin{aligned}
\epsilon^2 C &\geq \mathrm{E}\left[\min\left\{Z_{i^*},C-\sum_{i=1}^F y_i\right\}\middle|\mathcal{A}\right]\\
&\geq \Pr\left[Z_{i^*}\geq C-\sum_{i=1}^F y_i\middle|\mathcal{A}\right]\cdot\left(C-\sum_{i=1}^F y_i\right)\\
&\geq \epsilon C\cdot\Pr\left[Z_{i^*}\geq C-\sum_{i=1}^F y_i\middle|\mathcal{A}\right].
\end{aligned}$$

The last inequality follows from the assumption $\sum_{i=1}^F y_i\leq(1-\epsilon)C$. By rearranging Inequality (1), it follows that $\mathrm{E}[\min\{Z_{i^*},C-\sum_{i=1}^F y_i\}|\mathcal{A}]\geq(1-\epsilon)\cdot\mathrm{E}\left[Z_{i^*}|\mathcal{A}\right]$. We conclude the proof by noting that

$$\mathrm{E}[R_{\rm II}|\mathcal{A}]=\mathrm{E}\left[\min\left\{Z_{i^*},C-\sum_{i=1}^F y_i\right\}\middle|\mathcal{A}\right]\cdot p_{i^*}\geq(1-\epsilon)\cdot\mathrm{E}\left[Z_{i^*}|\mathcal{A}\right]\cdot p_{i^*}\geq(1-\epsilon)\cdot\mathrm{E}[R_{\rm I}|\mathcal{A}],$$

where the last inequality follows from the upper bound derived by the uncapacitated variant discussed in Section 4.2. □

### 4.5. Proofs of Lemmas 3 and 4

In what follows, we will use the following definition and theorems.

DEFINITION 2. For two real-valued random variables $X$ and $Y$ with cumulative distribution functions $F_X$ and $F_Y$, respectively, $X$ is *stochastically larger* than $Y$ (written as $X \geq_{st} Y$) if $F_X(x) \leq F_Y(x)$ for every $x \in \mathbb{R}$.

THEOREM 5 **(Theorem 1.2.8 in Müller and Stoyan (2002))**. *For two real-valued random variables $X$ and $Y$, $X \geq_{st} Y$ if and only if $\mathrm{E}[\psi(X)] \geq \mathrm{E}[\psi(Y)]$, for every nondecreasing function $\psi$. In particular, $\mathrm{E}[X] \geq \mathrm{E}[Y]$.*

THEOREM 6 **(Theorem 1.2.1.5 in Müller and Stoyan (2002))**. *Let $X$, $Y$ and $\Theta$ be real-valued random variables. If $[X|\Theta = \theta] \geq_{st} [Y|\Theta = \theta]$ for every value $\theta$ in the support of $\Theta$, then $X \geq_{st} Y$.*

*Proof of Lemma 3.* By Theorem 5, to prove $\mathrm{E}[\min\{Z_{i^*}, C - \sum_{i=1}^{F} y_i\}|\mathcal{A}] \leq \epsilon^2 C$, it is sufficient to show that $[Z_{i^*}|\mathcal{A}] \leq_{st} X_{i^*}$, since the latter claim will imply

$$\mathrm{E}\left[\min\left\{Z_{i^*}, C - \sum_{i=1}^{F} y_i\right\}\middle|\mathcal{A}\right] \leq \mathrm{E}\left[\min\left\{X_{i^*}, C - \sum_{i=1}^{F} y_i\right\}\right] \leq \mathrm{E}\left[\min\left\{X_{i^*}, C\right\}\right] \leq \epsilon^2 C.$$

The last inequality follows since product type $i^*$ is rare.

Now let $L$ be the random index of the customer that consumed the last unit of frequent product types, when the inventory levels of the frequent product types are $(y_1, \ldots, y_F)$; if no such customer exists, we define $L = M + 1$. It is straightforward to check that $[M \geq L] = \mathcal{A}$. Thus, $X_{i^*} \sim B(M, \alpha_{i^*})$, whereas $[Z_{i^*}|\mathcal{A}] \sim B([M - L|M \geq L], \alpha_{i^*})$. Consequently, to prove $[Z_{i^*}|\mathcal{A}] \leq_{st} X_{i^*}$ it remains to show $[M - L|M \geq L] \leq_{st} M$. However, for every $x \geq 0$, we have

$$\begin{aligned}
\Pr\left[M - L \leq x | M \geq L\right] &= \sum_{\ell=\sum_{i=1}^{F} y_i}^{\infty} \Pr\left[L = \ell | M \geq L\right] \cdot \Pr\left[M - L \leq x | M \geq L, L = \ell\right] \\
&= \sum_{\ell=\sum_{i=1}^{F} y_i}^{\infty} \Pr\left[L = \ell | M \geq L\right] \cdot \Pr\left[M - \ell \leq x | M \geq \ell\right] \\
&\geq \sum_{\ell=\sum_{i=1}^{F} y_i}^{\infty} \Pr\left[L = \ell | M \geq L\right] \cdot \Pr\left[M \leq x\right] \\
&= \Pr\left[M \leq x\right] \ .
\end{aligned}$$

The second equation holds since

$$\Pr\left[M - L \leq x | M >, L = \ell\right] = \Pr\left[M - \ell \leq x | M \geq \ell, L = \ell\right] = \Pr\left[M - \ell \leq x | M \geq \ell\right].$$

To better understand the latter equation, note that the event $[L = \ell]$ is measurable with respect to the choices of the first $\ell$ arrivals (i.e., after the first $\ell$ arrivals, we know whether it occurred or not). On the other hand, the number of arrivals after the $\ell$-th arrival, given that there were at least $\ell$ arrivals, is independent on the choices of the first $\ell$ arrivals. The first inequality above holds since $[M - \ell | M \geq \ell] \leq_{st} M$, by Assumption 2 that $M$ is IFR. □

*Proof of Lemma 4.* We will prove that $[Z_{i^*}|\mathcal{A}]$ is IFR, and in particular, $[Z_{i^*} - k|Z_{i^*} \geq k, \mathcal{A}] \leq_{st} [Z_{i^*}|\mathcal{A}]$, for every $k \geq 0$. This implies that, for each integer $k \geq 0$, we have $\mathrm{E}[Z_{i^*} - k|Z_{i^*} \geq k, \mathcal{A}] \leq \mathrm{E}[Z_{i^*}|\mathcal{A}]$ (See Theorem 5). To this end, let $L_k$ be the random index of the $k$-th customer with a preference list containing product type $i^*$, arriving after all the units of frequent product types have been consumed (i.e., a customer that is willing to buy product type $i^*$); if no such customer exists, we define $L_k = M + 1$. It is not difficult to verify that $[Z_{i^*} - k|Z_{i^*} \geq k, \mathcal{A}] \sim B[M - L_k|Z_{i^*} \geq k, \mathcal{A}], \alpha_{i^*}) \sim B([M - L_k|M \geq L_k], \alpha_{i^*})$. In addition, using the notation introduced in proof of Lemma 3, we have already observed that $[Z_{i^*}|\mathcal{A}] \sim B([M - L|\mathcal{A}], \alpha_{i^*})$. Consequently, to prove $[Z_{i^*} - k|Z_{i^*} \geq k, \mathcal{A}] \leq_{st} [Z_{i^*}|\mathcal{A}]$, it suffices to show $[M - L_k|M \geq L_k] \leq_{st} [M - L|\mathcal{A}] = [M - L|M \geq L]$.

In the remainder of the proof, we will use Theorem 6 to prove that $[M - L_k|M \geq L_k] \leq_{st} [M - L|M \geq L]$ taking $\Theta = L$. Thus, it suffices to show that, for every $\ell$ in the support of $L$, we have

$$[M - L_k|M \geq L_k, L = \ell] \leq_{st} [M - \ell|M \geq \ell, L = \ell] = [M - \ell|M \geq \ell] \ ,$$

where the equality above follows by the same arguments used in the proof of Lemma 3 above. Now for every integer $m$, we have

$$\begin{aligned}
&\Pr[M - L_k \leq m|M \geq L_k, L = \ell] \\
&\quad = \sum_{\ell_k=\ell}^{\infty} \Pr[L_k = \ell_k|M \geq L_k, L = \ell] \cdot \Pr[M - \ell_k \leq m|M \geq L_k, L = \ell, L_k = \ell_k] \\
&\quad = \sum_{\ell_k=\ell}^{\infty} \Pr[L_k = \ell_k|M \geq L_k, L = \ell] \cdot \Pr[M - \ell_k \leq m|M \geq \ell_k] \\
&\quad \geq \sum_{\ell_k=\ell}^{\infty} \Pr[L_k = \ell_k|M \geq L_k, L = \ell] \cdot \Pr[M - \ell \leq m|M \geq \ell] \\
&\quad = \Pr[M - \ell \leq m|M \geq \ell] \ .
\end{aligned}$$

The first equality holds since the event $[L = \ell, L_k = \ell_k]$ is measurable with respect to the choices of the first $\ell_k$ customers, and thus condition on the event $[M \geq \ell_k]$, is independent of $M - \ell_k$. The first inequality follows from $M$ being IFR. The same arguments can be used to show $[M - L_k|M \geq L_k] \geq [M - L'_k|M \geq L'_k]$, for each $k < k'$. This implies that $[Z_i x|\mathcal{A}]$ is indeed IFR. □

### 4.6. Algorithm Running Time

We have provided a constructive proof of Theorem 3. In particular, one can enumerate $O(C^{O(\log(1/\epsilon))})$ solutions to obtain $(1 - 3\epsilon)^4$ fraction of the optimal expected revenue. However, this is only pseudo-polynomial is the size of the input. In fact, one can improve the running time to $O((\log C)^{O(\log(1/\epsilon))})$, which is polynomial in the input size. For this purpose, let $S = \{i_1, \ldots, i_K\}$ be the subset of $O(\log(1/\epsilon))$ product types identified by our algorithm, and let $n^*_k \in \{0, \ldots, C\}$ be the number of units of product type $i_k$ in the best solution that stocks only $i_1, \ldots, i_K$. For every $1 \leq k \leq K$, we consider the following $O(\log C)$ choices for the number of units to stock from product type $i_k$:

$$\{0\} \cup \left\{ \lfloor (1+\epsilon)^\ell \rfloor : 0 \leq \ell \leq \lfloor \log_{(1+\epsilon)} C \rfloor \right\} \ .$$

Clearly, when $n^*_k > 0$, this value is located between two consecutive exponents of $1 + \epsilon$, say $(1 + \epsilon)^{\ell_k} \leq n^*_k < (1 + \epsilon)^{\ell_k+1}$. If we stock $\lfloor (1 + \epsilon)^{\ell_k} \rfloor$ units of product type $i_k$ instead of $n^*_k$, we obtain a fraction of at least $1/(1 + \epsilon) \geq 1 - \epsilon$ from the total expected revenue of each product type. It follows that with running time polynomial is the size of the input, one can obtain at least $(1 - 3\epsilon)^5$ of the optimal expected revenue.

Note that in Theorem 3, we claim that for any $0 \leq \epsilon \leq 0.5$, there is a sparse assortment with only $O(\log(1/\epsilon))$ product types that obtains at least $(1 - \epsilon)$ fraction of the optimal expected revenue. However,

we show instead that the revenue of the approximate sparse assortment computed by our PTAS is at least $(1-3\epsilon)^5$ of the optimal expected revenue. To reconcile this and obtain a solution that guarantees at least $(1-\epsilon)$ of the optimal expected revenue, we can run the algorithm with accuracy level

$$\hat{\epsilon} = \frac{1}{3}\left(1-(1-\epsilon)^{1/5}\right) \approx \frac{\epsilon}{15}.$$

## 5. Extension to Quality Categories

In this section, we extend the PTAS presented above to a model in which customers differentiate between product types not only based on price but also based on quality. This model is richer and captures additional important practical settings. The details of the models are as follows.

*Model description.* As before, we assume that product types are numbered such that the per-unit selling price is monotone non-decreasing, that is, $p_1 \leq \cdots \leq p_n$. In addition, product types are partitioned into *quality categories*. Specifically, let $1 = i_1 < i_2 < \cdots < i_k < i_{k+1} = n+1$, define $k$ quality categories $[i_l, i_{l+1})$, $l = 1, \ldots, k$, where the quality increases in $l$ (i.e., in the price range). For each category $[i_l, i_{l+1})$, the product types $i_l, i_l+1, \ldots, i_{l+1}-1$ are of similar quality, and are differentiated only by price. Each preference list of an arriving customer consists of some quality category, say $[i_l, i_{l+1})$, and a price threshold $p < p_{i_{l+1}}$. Thus, such customer is willing to buy product types within $[i_l, i_{l+1})$, with preference to the cheapest available product type available within this category up to price threshold $p$. The assumption is that the price ranges of different quality categories are well separated. We believe that this choice model captures practical situations in retailing environments.

*Extending the PTAS with known category capacities.* Observe that the total demand for product types in quality category $[i_l, i_{l+1})$ is not affected by the stocking levels of product types in other categories. Now suppose that one has already decided to allocate capacities $C_1, \ldots, C_k$ to the product types in categories $1, \ldots, k$, respectively. Then, the optimal allocation of the capacity $C_l$ to different product types within the $l$-th quality category can be solved separately using the PTAS discussed in Section 4.

Focusing on a single category $[i_l, i_{l+1})$, let $q_j$ be the probability that the preference list of a customer is $i_l, i_l+1, \ldots, j$, for $j = i_l, \ldots, i_{l+1}-1$, and let $Q_l = \sum_{j=1}^{l} q_j$ be the probability that the preference list of the customer belongs to category $l$. Consider the model discussed in Section 2, under Assumptions 1 and 2, with capacity $C_l$ and product types $i_l, \ldots, i_{l+1}-1$. The preference list of an arriving customer is $i_l, i_l+1, \ldots, j$ with probability $q_j/Q_l$. Thus, we can use the PTAS described in the previous section to obtain a near optimal inventory solution for this subproblem.

*Determining near-optimal capacities.* It follows that, when one knows the allocation of capacities to the different quality categories in the optimal solution, it is possible to devise a PTAS. However, these quantities are generally unknown. Instead, we will show that one can find near-optimal capacities by means of dynamic programming. Let $f_k(t)$ be the optimal expected revenue if $t \in \{1, \ldots, C\}$ units are stocked in quality categories $\{k, k+1, \ldots, l\}$ and let $\mu_k(t')$ be the optimal expected revenue from units in quality category $k$ if a total of $t'$ units are allocated to product types within the quality category $[i_k, i_{k+1})$. Note that $(1-\epsilon)$-approximations can be computed to $\mu_k(t)$ for all $k \in \{1, 2, \ldots, l\}, t \in \{1, \ldots, C\}$ using the PTAS described earlier. Now for any $k < l$,

$$f_k(t) = \max_{0 \leq t' \leq t} \left(\mu_k(t') + f_{k+1}(t-t')\right).$$

Using the above dynamic program, we can obtain a PTAS for $f_1(C)$.

THEOREM 7. *There is a PTAS for the model described in Section 2, where the preference lists are based on quality categories.*

Note that the dynamic programming algorithm described above is only pseudo-polynomial due to its linear dependency on $C$. However, we can improve the running time to polynomial in the input size by discretizing the state space of the dynamic program. We sketch the details in Appendix A.3.

## 6. General Stochastic Demand

In this section, we show that if the number of customers $M$ follows a general distribution (not necessarily IFR), then it is not possible in general to construct a near-optimal solution with only a constant number of product types. In particular, we prove the following theorem.

THEOREM 8. *There are instances of the capacitated assortment planning problem with nested preferences, where stocking any single product achieves only $O(1/n)$ of the optimal revenue where $n$ is the number of product types.*

Theorem 8 immediately implies that any solution with only constant number of product types also obtains only $O(1/n)$ of the optimal expected revenue.

Consider the following instance with $n$ product types and capacity $C = n$. Let $M$ denote the random number of customers, and let $n, n^2, \ldots, n^n$ be the possible values that $M$ can take. In what follows, $\beta = \beta(n) \in (0,1)$ is a parameter which value will be specified later. For $1 \le i \le n$, we set

$$\Pr\left[M = n^i\right] = \frac{\beta^{i-1}(1-\beta)}{1-\beta^n}\,, \quad p_i = \frac{1}{\beta^{i-1}(1-\beta)}\,, \quad \alpha_i = \frac{1}{n^i}\,.$$

In the next two lemmas, we prove that the revenue from stocking $n$ units of any product achieves at most $O(1/n)$ times the optimal revenue.

LEMMA 5. *By stocking one unit of each product, the expected revenue obtained is $\Omega(n)$.*

*Proof.* Given that $M = n^i$, the probability that the unit of product $i$ is consumed when the inventory vector contains one unit of each product is greater or equal to the analogous probability when we stock $i-2$ units of product $i-2$, one unit of product $i-1$, and one unit of product $i$. The latter event contains, in particular, the intersection of three (independent) events:

1. Product $i-2$ appears in the preference list of at least $i-2$ customers from within the first $n^i/4$ customers. To lower bound the probability of this event, let $X$ be the random number of customers until product $i-2$ appears in the preference list of exactly $i-2$ customers, in an infinite sequence of customers. Clearly, $X \sim NB(i-2, \alpha_{i-2})$ and $\mathrm{E}[X] = (i-2)/\alpha_{i-2} \le n^{i-1}$. Then, the probability of the desired event is at least $\Pr[X \le n^i/4] \ge \Pr[X \le 2\mathrm{E}[X]] \ge 1/2$, where in the first inequality we assumed $n \ge 4$.
2. Product $i-1$ appears in the preference list of at least one customer from within the next $n^i/4$ customers. This event occurs with probability $1-(1-\alpha_{i-1})^{n^i/4}$.
3. Product $i$ appears in the preference list of at least one customer from within the last $n^i/2$ customers. The probability of this event is $1-(1-\alpha_i)^{n^i/2}$.

As a result of the above discussion, letting $B_i$ denote the event "product $i$ is consumed", we have

$$\begin{aligned}
\Pr\left[B_i | M = n^i\right] &\ge \frac{1}{2} \cdot \left(1-(1-\alpha_{i-1})^{n^i/4}\right) \cdot \left(1-(1-\alpha_i)^{n^i/2}\right) \\
&\ge \frac{1}{2} \cdot \left(1-e^{-\alpha_{i-1}n^i/4}\right) \cdot \left(1-e^{-\alpha_i n^i/2}\right) \\
&= \frac{1}{2} \cdot \left(1-e^{-n/4}\right) \cdot \left(1-e^{-1/2}\right) \ge \frac{1}{40}\,.
\end{aligned}$$

Therefore, the expected revenue from stocking one unit of each product is

$$\begin{aligned}
\mathrm{E}\left[R_M(1,\ldots,1)\right] &= \sum_{i=1}^{n} \Pr\left[M = n^i\right] \cdot \mathrm{E}\left[R_M(1,\ldots,1) | M = n^i\right] \\
&\ge \sum_{i=1}^{n} \Pr\left[M = n^i\right] \cdot \Pr\left[B_i | M = n^i\right] \cdot \mathrm{E}\left[R_M(1,\ldots,1) | B_i, M = n^i\right]
\end{aligned}$$

$$\geq \sum_{i=1}^{n} \Pr\left[M = n^i\right] \cdot \Pr\left[B_i | M = n^i\right] \cdot p_i$$
$$\geq \frac{1}{40} \sum_{i=1}^{n} \frac{\beta^{i-1}(1-\beta)}{1-\beta^n} \cdot \frac{1}{\beta^{i-1}(1-\beta)} \geq \frac{n}{40} .$$

□

LEMMA 6. *By stocking $n$ units of any single product $i$, the expected revenue obtained is $O(1)$.*

*Proof.* Let $\bar{y}$ be the inventory vector, indicating that $n$ units of product $i$ are stocked. Then

$$\mathrm{E}\left[R_M(\bar{y})\right] = \sum_{j=1}^{i} \Pr\left[M = n^j\right] \cdot \mathrm{E}\left[R_M(\bar{y}) | M = n^j\right] + \Pr\left[M \geq n^{i+1}\right] \cdot \mathrm{E}\left[R_M(\bar{y}) | M \geq n^{i+1}\right] .$$

We proceed by separately upper bounding each of the two terms above. Let us now fix $\beta = 2/n$ and assume that $n$ is sufficiently large ($n \geq 4$ is enough).

*First term.* Here, we will make use of the inequality $\mathrm{E}[R_M(y)|M = n^j] \leq n^j \alpha_i p_i$ (assuming $n^j$ units can be stocked), to obtain

$$\begin{aligned}
\sum_{j=1}^{i} \Pr\left[M = n^j\right] \cdot \mathrm{E}\left[R_M(y) | M = n^j\right] &\leq \sum_{j=1}^{i} \Pr\left[M = n^j\right] \cdot n^j \alpha_i p_i \\
&= \sum_{j=1}^{i} \frac{\beta^{j-1}(1-\beta)}{1-\beta^n} \cdot n^j \cdot \frac{1}{n^i} \cdot \frac{1}{\beta^{i-1}(1-\beta)} \\
&= \frac{1}{1-\beta^n} \sum_{j=1}^{i} \left(\frac{1}{\beta n}\right)^{i-j} \\
&= \frac{1}{1-(2/n)^n} \sum_{j=1}^{i} \left(\frac{1}{2}\right)^{i-j} \leq 4 .
\end{aligned}$$

*Second term.* In this case, we use the inequality $\mathrm{E}[R_M(y)|M \geq n^{i+1}] \leq np_i$, and get

$$\begin{aligned}
\Pr\left[M \geq n^{i+1}\right] \cdot \mathrm{E}\left[R_M(y) | M \geq n^{i+1}\right] &= \sum_{j=i+1}^{n} \Pr\left[M = n^j\right] \cdot np_i \\
&= n \sum_{j=i+1}^{n} \frac{\beta^{j-1}(1-\beta)}{1-\beta^n} \cdot \frac{1}{\beta^{i-1}(1-\beta)} \\
&\leq \frac{n}{1-\beta^n} \sum_{j=1}^{\infty} \beta^j = \frac{n\beta}{(1-\beta^n)(1-\beta)} \\
&= \frac{2}{(1-(2/n)^n)(1-2/n)} \leq 8 .
\end{aligned}$$

□

## 7. Computational Experiments

In this section, we describe computational experiments that were designed to evaluate the performance of our PTAS on instances where Assumption 1 (nested preference lists) is not satisfied. Even though the algorithm and its analysis are presented in Section 4 under the assumption that the collection of preference lists is nested, we can easily adapt the overall approach to more general settings. Furthermore, the computational results below show that the performance of our algorithm is indeed good for fairly general choice models.

### 7.1. General Choice Model

We consider the following choice model as defined by the random utility model. Let the utility $U_{ij}$ of product type $j$ for customer $i$ be given as,

$$U_{ij} = U_j + \epsilon_{ij},$$

where $\epsilon_{ij}$ is a random utility shock distributed according to a known distribution $f_j$. Note that $\epsilon_{ij}$ is i.i.d for each customer $i$ and $U_j$ is the base utility of product type $j$, where $U_1 \geq U_2 \geq \cdots \geq U_n$. As before, let $p_j$ denote the per-unit price of product type $j$, with $p_1 \leq p_2 \leq \cdots \leq p_n$. We also assume that the no-purchase option (henceforty, product type 0) has a base utility $U_0 = 0$ and each customer $i$ has utility $U_{i0}$ for item 0, given by $U_{i0} = U_0 + \epsilon_{i0} = \epsilon_{i0}$, where $\epsilon_{i0}$ is distributed according to the PDF $f_0$. Recall that for any product type $j$, $\alpha_j$ denotes the probability that a customer is willing to buy product type $j$, i.e., product type $j$ appears in the preference list of the customer. For any product $j$, given the base utility $U_j$ and the PDF $f_j$ for the random shock, we can compute $\alpha_j$ as follows. Product type $j$ appears in the preference list for customer $i$ if and only if $U_j + \epsilon_{ij} \geq \epsilon_{i0}$. Therefore,

$$\alpha_j = \Pr\left[U_j + \epsilon_{ij} \geq \epsilon_{i0}\right],$$

which can be computed when $f_0, \ldots, f_n$ are known. Note that this choice model is fairly general. For instance, if $f_j$ is a log-weibull distribution for all $j = 0, 1, \ldots, n$, this model is equivalent to the widely used multinomial logit (MNL) choice model. For our computational experiments, we assume that $f_j(x) = c_j$ for some constant $c_j$ for all $j = 0, 1, \ldots, n$, i.e., $\epsilon_{ij}$ is distributed according to some uniform distribution (possibly different for different product types). We make this assumption just for simplifying the computation of $\alpha_j$, noting that the algorithm can be adapted for any distributions $f_j$.

### 7.2. Parameter Settings

We next describe the parameters for our computational experiments. We use the following values for the base utilities, $U_j$: $U_0 = 0$ and $U_j = 1/j$ for $j = 1, \ldots, n$. For every customer $i$, the random utility shock $\epsilon_{ij}$ is distributed uniformly between 0 and $R_j$ where

$$R_0 = \frac{3}{2}\,; \quad R_1 = 0.499\,; \quad R_j = \frac{4}{j^2},\ j \geq 2.$$

Knowing the values of $U_j$ and distributions of $\epsilon_{ij}$ for all $j = 0, 1, \ldots, n$, we can compute the values of $\alpha_j, j = 1, \ldots, n$ as follows:

$$\alpha_j = \begin{cases} 1 - \dfrac{(R_0 - U_j)^2}{2R_0R_j} & \text{if } (U_j + R_j) > R_0 \\ \dfrac{U_j}{R_0} + \dfrac{R_j}{2R_0} & \text{otherwise} \end{cases} \tag{2}$$

Each customer $i$ samples the random utility shocks $\epsilon_{ij}$ from the given distribution and computes utilities $U_{ij} = U_j + \epsilon_{ij}$ for all product types $j = 0, 1, \ldots, n$. Let $S_i = \{j : U_{ij} > U_{i0}\}$. Customer $i$ buys the highest utility product from $S_i$ that is available. If none of the products in $S_i$ is available, then he/she buys nothing. We use randomly generated values for the product prices. In particular, we use the following values: $p_0 = 0$, $p_1 = c$, and for any $j > 1$,

$$p_j = \begin{cases} p_{j-1} + (1 + \tau_j)\Delta & \text{w.p. } 0.9 \\ (2 + \tau_j)p_{j-1} & \text{w.p. } 0.1 \end{cases}$$

where $\tau_j$ is uniformly random between 0 and 1, and $\Delta$ is a constant. We use $c = 10$ and $\Delta = 0.8$ in our computational experiments.

### 7.3. Our PTAS Adapted to General Choice Model

Even though our PTAS can be directly adapted for the above choice model, we explicitly describe the resulting procedure. For any product type $j$, let $X_j$ denote the number of customers whose preference list contains $j$. Recall that a product type is frequent if $\mathrm{E}[\min\{X_j, C\}] \geq \epsilon^2 C$ and rare otherwise. Also, recall that a product type $j$ is cheap if $p_j \leq \epsilon^3 p_F$, where $p_F$ is the price of the most expensive frequent product. Note that these definitions are valid for all choice models. Our algorithm proceeds along the lines described in the proof of Lemma 1, as shown in Figure 1.

**Adapted PTAS**

1. Eliminate cheap product types with price smaller than $\epsilon^3 p_F$.
2. Allocate $\epsilon C$ capacity to
$$\mathsf{argmax}_j \, \alpha_j p_j.$$
3. Allocate remaining capacity as follows.
   (a) Geometrically partition the range $[\epsilon^3 p_F, p_F]$
   (b) Consider $O(\log(1/\epsilon))$ product types at the boundary of the geometric partition.
   (c) Select best solution from all possible inventory levels for $O(\log(1/\epsilon))$ types, such that the total capacity is at most $(1-\epsilon)C$.

**Figure 1** Adaptation of the PTAS for more general choice models.

From an implementation point of view, the expected revenue for a given assortment is computed by simulation. In what follows, we denote the expected revenue of the assortment computed by our algorithm as Alg. We set a time limit of 7200 secs and report the best solution found within that time frame.

### 7.4. Other Heuristics

We compare the performance of our PTAS against three different heuristics: $ii)$ local search, $ii)$ projected gradient descent, and $iii)$ hybrid algorithm. We describe the three algorithms below.

**Local Search** We consider the following local search heuristic where in each step, the algorithm either increases the inventory of some product type if total capacity is less than $C$ or finds a pair of product types such that increasing the inventory of one product type and decreasing the inventory of the other product type improves the expected revenue. The algorithm terminates with a local maximum solution when no such local swap is possible. The specifics of this algorithm are given in Figure 2.

**Local Search Algorithm** (LS)

1. Initial feasible solution: $y_j = C$ if product type $j$ maximizes $\alpha_j p_j$ and 0 otherwise.
2. While $(y_1, \ldots, y_n)$ is not local optimum:
   (a) Find product types $i, j$ such that moving one unit from $i$ to $j$ increases the expected revenue.
   (b) The new solution is obtained by $y_i \leftarrow y_i - 1$ and $y_j \leftarrow y_j + 1$.

**Figure 2** The local search algorithm.

We denote the expected revenue computed by the above local search algorithm as LS. We use a time limit of 7200 secs for this algorithm. However, this algorithm terminated in less than 40 seconds, and converged to a local maximum for all instances in our experiments.

**Projected Gradient Descent**. We consider an adaptation of the stochastic gradient descent algorithm of Mahajan and van Ryzin (2001) for our dynamic assortment problem under capacity constraint. Note that Mahajan and van Ryzin (2001) considered a per-unit selling price and per-unit cost for each item type

and the goal was to maximize the expected profit. There were no constraints on the inventory in their model. In our model, there is a capacity constraint on the total number of the units of all product types that can be stocked and the goal is to maximize the expected revenue. Moreover, the revenue function is defined only for integer values of inventory.

To implement a gradient descent algorithm for our problem, we define a continuous extension of the revenue function following Lovász extension of a discrete function. For the revenue function $f : \mathbb{Z}_+^n \to \mathbb{R}_+$, we define the continuous extension $\hat{f} : \mathbb{R}_+^n \to \mathbb{R}_+$ as follows.

$$\hat{f}(y_1, \ldots, y_n) = f(\lfloor y_1 \rfloor, \ldots, \lfloor y_n \rfloor) + \sum_{i=1}^{n} \left[ f\left( \sum_{j=1}^{i} \mathbf{1}_{\pi(j)} \right) - f\left( \sum_{j=1}^{i-1} \mathbf{1}_{\pi(j)} \right) \right] \cdot \left( \lceil y_{\pi(i)} \rceil - \lfloor y_{\pi(i)} \rfloor \right),$$

where $\pi \in \Pi(\{1, \ldots, n\})$ is a permutation such that

$$y_{\pi(1)} \geq y_{\pi(2)} \geq \ldots \geq y_{\pi(n)}.$$

We implement the gradient descent for the continuous extension $\hat{f}$. The gradient of $\hat{f}$ can be expressed as follows.

$$\frac{\partial}{\partial y_{\pi(i)}} \hat{f} = f\left( \sum_{j=1}^{i} \mathbf{1}_{\pi(j)} \right) - f\left( \sum_{j=1}^{i-1} \mathbf{1}_{\pi(j)} \right), \; i = 1, \ldots, n,$$

which can be computed via sampling. Since there is capacity constraint on the total number of items stocked, after each solution update we need to project the solution to the feasible set. Therefore, in any iteration $k$, we update the solution as follows.

$$\mathbf{y}^k = \frac{C \cdot (\mathbf{y}^{k-1} + a_k \nabla \hat{f}(\mathbf{y}^{k-1}))}{\mathbf{e}^T (\mathbf{y}^{k-1} + a_k \nabla \hat{f}(\mathbf{y}^{k-1})},$$

where $\mathbf{e}$ is the vector of all ones and $a_k$ is the step size in iteration $k$. The algorithm terminates when we find a stationary point, i.e., $\mathbf{y}^k = \mathbf{y}^{k-1}$. This essentially translates to the condition

$$\nabla \hat{f}(\mathbf{y}^k) = \tau \mathbf{y}^k,$$

for some $\tau \in \mathbb{R}$. The complete algorithm is described in Figure 3.

**Projected Gradient Descent Algorithm** (PGD)

1. Initial feasible solution: $\mathbf{y}^1 = \mathbf{0}$, $k = 1$.
2. Let step size $a_i = \frac{1}{i}$ in iteration $i$.
3. While $(||\mathbf{y}^k - \mathbf{y}^{k-1}||_\infty \geq \epsilon)$
   (a) Compute $\nabla \hat{f}(\mathbf{y}^k)$.
   (b) Update
   $$\mathbf{y}^{k+1} = \frac{C \cdot (\mathbf{y}^k + a_k \nabla \hat{f}(\mathbf{y}^k))}{\mathbf{e}^T (\mathbf{y}^k + a_k \nabla \hat{f}(\mathbf{y}^k)}.$$
   (c) Update $k \leftarrow k + 1$.
4. Return the continuous solution $\mathbf{y}^k$.

**Figure 3** Projected Gradient Descent algorithm.

Note that the solution $\mathbf{y}$ computed by the above gradient descent is not integral and thus, not precisely feasible. Also, it is not clear how to efficiently compute an integral solution whose expected revenue is close to $\hat{f}(\mathbf{y})$. The solution $\mathbf{y}_u$ where

$$y_j^u = \lceil y_j \rceil, \forall j = 1, \ldots, n,$$

violates the capacity constraint; and solution $\mathbf{y}^l$, where

$$y_j^l = \lfloor y_j \rfloor, \forall j = 1, \ldots, n,$$

can have significantly lower expected revenue as compared to $\hat{f}(\mathbf{y})$. Therefore, it is possible that $\hat{f}(\mathbf{y})$ is higher than the optimal expected revenue for some instances.

**Hybrid Algorithm**. We also consider a hybrid algorithm that is a combination of projected gradient descent algorithm and the local search algorithm. We first compute a fractional stationary point using the projected gradient search algorithm and then a local search algorithm to compute an integer local maximum close to the fractional stationary point. The algorithm can be described as follows.

**Hybrid Algorithm** (Hybrid)
1. Compute a stationary point $\mathbf{y}$ using Proj-Gradient-Descent.
2. Let $y_j^l = \lfloor y_j \rfloor$ for all $j = 1, \ldots, n$.
3. Update $y_1^l = C - (y_2^l + \ldots + y_n^l)$.
4. While $(y_1^l, \ldots, y_n^l)$ is not local optimum:
   (a) Find product types $i, j$ such that moving one unit from $i$ to $j$ increases the expected revenue.
   (b) The new solution is obtained by $y_i^l \leftarrow y_i^l - 1$ and $y_j^l \leftarrow y_j + 1$.
5. Return $\mathbf{y}^l$.

**Figure 4** Hybrid algorithm.

Note that the hybrid algorithm starts with a feasible integer solution obtained by rounding down the fractional solution obtained by the projected gradient descent algorithm. The expected revenue of the rounded down solution can be significantly lower than the value of the fractional stationary solution. However, after the local swap operations, it is possible that the hybrid algorithm finds a feasible integer solution with value higher than that of the fractional stationary point.

## 7.5. Results

In addition to testing the algorithms mentioned above, we also compute the optimal expected revenue, denoted by OPT, using enumeration for additional comparisons. However, for several instances the exhaustive does not terminate even after 96 hours. Therefore, we report the best solution obtained within a time limit of 7200 secs as well. Table 1 summarizes the experimental results.

The results of the computational experiments show that our algorithm performs significantly better than all the three heuristics. The projected gradient descent algorithm performs better than the local search heuristic on almost all instances. However, we should note that the solution obtained using the projected gradient descent algorithm is fractional and thus, infeasible for the original problem. The hybrid algorithm performs the best among the three heuristics. However, our adapted PTAS (Alg) obtains a solution with a significantly higher expected revenue than even the hybrid algorithm for all instances. The value of the solution obtained by our algorithm is better by around $8.5\%$ on average as compared to the hybrid algorithm.

For instances where OPT found an optimal solution within 96 hours, the expected revenue of the solution computed by Alg is within $4\%$ of the optimal expected revenue. We use $\epsilon = 0.1$ for Alg in our experiments which provide a theoretical guarantee of being within $10\%$ of the optimal. Therefore, the numerical results for instances from a more general choice model are significantly better than even the worst-case theoretical

| | $n$ | $m$ | $C$ | Alg | LS | Grad-descent | Hybrid | OPT* | OPT |
|---|---|---|---|---|---|---|---|---|---|
| 1. | 10 | 40 | 15 | 189.76 | 175.14(8.35%) | 170.48(11.31%) | 188.75(0.54%) | 194.46 | 194.46 |
| 2. | 10 | 40 | 20 | 240.19 | 215.31(11.56%) | 215.84(11.28%) | 234.58(2.39%) | 243.639* | 244.70 |
| 3. | 10 | 40 | 25 | 291.01 | 247.15(17.75%) | 258.13(12.74%) | 272.86(6.65%) | 247.191* | 293.87 |
| 4. | 10 | 50 | 30 | 352.03 | 300.10(17.30%) | 308.19(14.22%) | 319.35(10.23%) | 240.263* | – |
| 5. | 10 | 100 | 40 | 507.3 | 412.80(22.89%) | 475.26(6.74%) | 498.5(1.77%) | 381.025* | – |
| 6. | 10 | 100 | 50 | 609.16 | 483.80(25.91%) | 540.67(12.67%) | 554.07(9.94%) | 379.663* | – |
| 7. | 10 | 150 | 80 | 967.51 | 678.93(42.50%) | 793.97(21.86%) | 829.74(16.60%) | 487.144* | – |
| 8. | 10 | 200 | 100 | 1227.87 | 905.85(35.54%) | 1007.55(21.87%) | 1026.55(19.61%) | 642.956* | – |
| 9. | 12 | 100 | 50 | 563.083 | 496.70(13.36%) | 515.56(9.22%) | 534.05(5.44%) | 294.715* | – |
| 10. | 15 | 40 | 15 | 227.18 | 191.55(18.60%) | 168.52(34.81%) | 197.96(14.76%) | 153.2* | 235.42 |
| 11. | 15 | 40 | 20 | 264.2 | 206.39(28.01%) | 212.27(24.46%) | 245.34(7.69%) | 150.914* | – |
| 12. | 20 | 50 | 25 | 327.36 | 314.76(4%) | 250.41(30.73%) | 306.07(6.96%) | 218.952* | – |

**Table 1** Here, $(^*)$ = the algorithm reached the time limit, and $(-)$ = the optimal algorithm did not terminate within 96 hours. For each heuristic, LS, Grad-descent, and Hybrid, we note the % improvement in expected revenue Alg over each of the heuristics in brackets.

bounds of our algorithm for the special case of nested preferences. We would like to re-emphasize that our algorithm is fairly general and can be adapted for very general choice models. While we present theoretical guarantees only for the case of nested preference lists, our algorithm performs reasonably well as compared to OPT for significantly more general choice models.

Furthermore, Alg performs significantly better than other known heuristics including local search and projected gradient descent (suitably adapted from Mahajan and van Ryzin (2001)). It is worth noting that, during our experiments, we discovered that there are many local maxima in this problem, and that the convergence to a particular solution is very sensitive to the initial solution. The running time of LS is small due to the presence of many local maxima, which make the algorithm converge to one near the initial solution very quickly. We experimented with several initial solutions including: (1) $C$ units of the most expensive product type, (2) $C$ units of the least expensive product type, and (3) $C$ units of product type $j$ that maximizes $\alpha_j p_j$. In Table 1, we report results for the best initial solution, which was the one in item (3) for all instances.

## 8. Concluding Remarks

*Approximability of general model.* Our main results indicate that the model under consideration is indeed NP-hard for general choice models, and can be efficiently approximated within any degree of accuracy for certain classes of choice models. However, even though we were able to establish a lower bound of $1-1/e$ on the approximability of an extremely simple model (with a single customer), it is quite possible that stronger complexity results could be obtained for the model in its utmost generality (i.e., arbitrary number of customers with arbitrary preferences). For this reason, it would be interesting to investigate such complexity issues as part of future research, and possibly complement them by corresponding upper bounds through algorithmic methods.

*Hardness of nested preferences.* Under the technical restrictions listed in Section 4, where customer preferences are assumed to be nested and where the number of arriving customers is assumed to be drawn from an IFR distribution, we were able to devise a polynomial-time approximation scheme (PTAS). However, we

do not know if this variant is NP-hard or can be solved in polynomial time. We pose the task of fully characterizing the hardness of nested preferences model as an interesting direction for future research. Although we would like to note that even if this model can be optimally solved in polynomial time, the managerial insights from the structure of a near-optimal solution (that a constant number of product types are sufficient) obtained by our algorithm are useful. Moreover, our algorithm provides

1. An immediate way to recognize the products involved.
2. A very efficient enumeration method, to spread the capacity $C$ between these products. Here, only $O((\log C)^{O(\log 1/\epsilon)}) = O((1/\epsilon)^{O(\log \log C)})$ inventory vectors need to be evaluated.

Furthermore, the computational experiments show that our PTAS performs well for significantly general choice models derived from random utility models and thus, can be used as a practical method for assortment optimization problems.

## Acknowledgments

The first author's research is partially supported by NSF Grant CMMI-1201116. The second author's research is partially supported by NSF grants DMS-0732175 and CMMI-0846554 (CAREER Award), an AFOSR award FA9550-08-1-0369, an SMA grant and the Buschbaum Research Fund of MIT.

## References

Anupindi, R., S. Gupta, M. A. Venkataramanan. 2006. Managing variety on the retail space: Using household scanner panel data to rationalize assortments. Tech. rep., University of Michigan.

Bansal, Nikhil, Amit Chakrabarti, Amir Epstein, Baruch Schieber. 2006. A quasi-PTAS for unsplittable flow on line graphs. *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*. 721–729.

Cachon, G., C. Terwiesch, Y. Xu. 2005. Retail assortment planning in the presence of consumer search. *Manufacturing & Service Operations Management* **7**(4) 330–346.

Chan, T.-H. Hubert, Khaled M. Elbassioni. 2011. A QPTAS for TSP with fat weakly disjoint neighborhoods in doubling metrics. *Discrete & Computational Geometry* **46**(4) 704–723.

Chen, F., Y. Bassok. 2008. Substitution and variety. Working paper, submitted.

de la Vega, Wenceslas Fernandez, George S. Lueker. 1981. Bin packing can be solved within 1+epsilon in linear time. *Combinatorica* **1**(4) 349–355.

Feige, U. 1998. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM* **45**(4) 634–652.

Fisher, M. L., R. Vaidyanathan. 2007. An algorithm and demand estimation procedure for retail assortment optimization. Manuscript.

Garey, M. R., D. S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman.

Gaur, V., D. Honhon. 2006. Assortment planning and inventory decisions under a locational choice model. *Management Science* **52**(10) 1528–1543.

Halman, N., D. Klabjan, C.L. Li, J. Orlin, D. Simchi-Levi. 2008. Fully polynomial time approximation schemes for stochastic dynamic programs. *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 700–709.

Halman, Nir, Diego Klabjan, Mohamed Mostagir, James B. Orlin, David Simchi-Levi. 2009. A fully polynomial-time approximation scheme for single-item stochastic inventory control with discrete demand. *Mathematics of Operations Research* **34**(3) 674–685.

Har-Peled, Sariel. 2011. *Geometric Approximation Algorithms*, *Mathematical Surveys and Monograph*, vol. 173. American Mathematical Society. Chapter 23.

Ho, T.H., C.S. Tang. 1998. *Product Variety Management: Research Advances*. Springer.

Honhon, D., V. Gaur, S. Seshadri. 2007. Assortment planning and inventory decisions under stock-out based substitution. Working paper.

Hopp, W.J., X. Xu. 2005. Product line selection and pricing with modularity in design. *Manufacturing & Service Operations Management* **7**(3) 172–187.

Kok, A.G., M.L. Fisher. 2007. Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research* **55**(6) 1001–1021.

Kok, AG, ML Fisher, R. Vaidyanathan. 2006. Assortment planning: Review of literature and industry practice. N. Agrawal, S. A. Smith, eds., *Retail Supply Chain Management*. Kluwer.

Lancaster, K. 1990. The economics of product variety: A survey. *Marketing Science* **9**(3) 189–206.

Mahajan, S., G. van Ryzin. 2001. Stocking retail assortments under dynamic consumer substitution. *Operations Research* **49**(3) 334–351.

Müller, Alfred, Dietrich Stoyan. 2002. *Comparison Methods for Stochastic Models and Risks*. Wiley.

Nagarajan, M., S. Rajagopalan. 2008. Inventory models for substitutable products: Optimal policies and heuristics. *Management Science* **54**(8) 1453–1466.

Netessine, S., N. Rudi. 2003. Centralized and competitive inventory models with demand substitution. *Operations Research* **51**(2) 329–335.

Parlar, M., SK Goyal. 1984. Optimal ordering decisions for two substitutable products with stochastic demands. *Operations Research* **21**(1) 1–15.

Pentico, D.W. 1974. The assortment problem with probabilistic demands. *Management Science* **21**(3) 286–290.

Ramdas, K. 2003. Managing product variety: An integrative review and research directions. *Production and Operations Management* **12**(1) 79–101.

Remy, Jan, Angelika Steger. 2009. A quasi-polynomial time approximation scheme for minimum weight triangulation. *Journal of the ACM* **56**(3).

Shaked, Moshe, J. George Shanthikumar. 1994. *Stochastic Orders and Their Applications*. Academic Press.

Smith, S.A., N. Agrawal. 2000. Management of multi-item retail inventory systems with demand substitution. *Operations Research* **48**(1) 50–64.

Talluri, K., G. van Ryzin. 2004. Revenue management under a general discrete choice model of consumer behavior. *Management Science* **50** 15–33.

Van Ryzin, G., S. Mahajan. 1999. On the relationship between inventory costs and variety benefits in retailassortments. *Management Science* **45**(11) 1496–1509.

## Appendix A: Additional Proofs

### A.1. Quasi-PTAS for Purchasing Costs

Let us consider a model where, instead of capacities, we are given per-unit purchasing costs $(c_1, \ldots, c_n)$, and per-unit selling prices $(p_1, \ldots, p_n)$. The objective is to compute an inventory vector such that its total purchasing costs falls within a budget of $B$ and such that the expected total revenue is maximized. Our basic assumption, in addition to the nested preference structure (see Section 4) is that the purchasing cost of any product increases with its selling price, i.e., $c_1 \leq \cdots \leq c_n$ and $p_1 \leq \cdots \leq p_n$; on the other hand, we will not be needing Assumption 2, meaning that the number of customers $M$ can follow any distribution over the non-negative integers, instead of being restricted to IFR ones.

Suppose that, in the optimal solution, we gain an expected revenue of $R^*$. Then, based on the technical ideas given in Section 4.3, the following algorithm can be easily shown to obtain a fraction of at least $1-\epsilon$ of the optimal revenue $R^*$ while meeting the budget constraint:

- We partition the collection of products $1, \ldots, n$ into cost classes $\mathcal{C}_1, \ldots, \mathcal{C}_K$, as well as into price classes $\mathcal{P}_1, \ldots, \mathcal{P}_L$, where

$$\mathcal{C}_k = \left\{ i : (1+\epsilon)^{k-1} c_1 \leq c_i \leq (1+\epsilon)^k c_1 \right\} \quad \text{and} \quad \mathcal{P}_\ell = \left\{ i : (1+\epsilon)^{\ell-1} p_1 \leq p_i \leq (1+\epsilon)^\ell p_1 \right\} .$$

Here, $K = O(\log(c_n/c_1))$ whereas $L = O(\log(p_n/p_1))$. With these definitions at hand, let $\mathcal{I}_{k\ell}$ be the set of products that belong to the $k$-th cost class and to the $\ell$-th price class, that is, $\mathcal{I}_{k\ell} = \mathcal{C}_k \cap \mathcal{P}_\ell$. It is not difficult to verify that each $\mathcal{I}_{k\ell}$ forms a sub-interval of $1, \ldots, n$ and that only $O(K+L)$ sets out of $\{\mathcal{I}_{k\ell}\}_{k,\ell}$ are non-empty.

• For every non-empty set $\mathcal{I}_{k\ell}$, we approximately guess (from below, up to $1-\epsilon$) the number of units $\mathcal{U}_{k\ell}$ that are purchased from $\mathcal{I}_{k\ell}$ in the optimal solution. As a result, we get an estimate of $(1-\epsilon)\mathcal{U}_{k\ell} \leq \hat{\mathcal{U}}_{k\ell} \leq \mathcal{U}_{k\ell}$. The total number of guesses to consider is

$$\left(O\left(\log\frac{B}{c_1}\right)\right)^{O(K+L)} = \left(O\left(\log\frac{B}{c_1}\right)\right)^{O(\log(c_n/c_1)+\log(p_n/p_1))} = O\left(\frac{c_n}{c_1}\cdot\frac{p_n}{p_1}\right)^{O(\log\log(B/c_1))} ,$$

since $\lfloor B/c_1 \rfloor$ is an obvious upper bound on the number of units purchased from any $\mathcal{I}_{k\ell}$.

• Based on these guesses, in each $\mathcal{I}_{k\ell}$ we order $\hat{\mathcal{U}}_{k\ell}$ units of the minimal index product, which also has to be the least expensive and the least profitable product in $\mathcal{I}_{k\ell}$ by the assumptions $c_1 \leq \cdots \leq c_n$ and $p_1 \leq \cdots \leq p_n$.

Due to the nested preference lists, it follows that within each $\mathcal{I}_{k\ell}$, the $t$-th unit we order (of the minimal index product) is consumed with probability at least as high as the $t$-th unit ordered in the optimal solution (counting from the least profitable to the most profitable product in $\mathcal{I}_{k\ell}$). Furthermore, if and when this unit is indeed consumed, the resulting revenue is at least $1/(1+\epsilon) \geq (1-\epsilon)$ times its revenue in the optimal solution, since the selling prices within $\mathcal{I}_{k\ell}$ differ by a factor of at most $1+\epsilon$. It follows that, out of the optimal expected revenue $R^*$, we obtain at least $(1-\epsilon)R^*$, as our guess for the number of units purchased from $\mathcal{I}_{k\ell}$ satisfies $\hat{\mathcal{U}}_{k\ell} \geq (1-\epsilon)\mathcal{U}_{k\ell}$. This explanation handles the revenue question, but still, it remains to argue that the budget $B$ is not exceeded. For this purpose, if we use $(y_1^*, \ldots, y_n^*)$ to denote the optimal inventory vector, the total purchasing cost can be bounded by

$$\sum_{k,\ell} \hat{\mathcal{U}}_{k\ell} \cdot \min_{i\in\mathcal{I}_{k\ell}} c_i \leq \sum_{k,\ell} \mathcal{U}_{k\ell} \cdot \min_{i\in\mathcal{I}_{k\ell}} c_i \leq \sum_{k,\ell}\sum_{i\in\mathcal{I}_{k\ell}} c_i y_i^* = \sum_{i=1}^{n} c_i y_i^* \leq B ,$$

where the first inequality holds since $\hat{\mathcal{U}}_{k\ell} \leq \mathcal{U}_{k\ell}$, and the second inequality follows from observing that the optimal solution pays at least $\min_{i\in\mathcal{I}_{k\ell}} c_i$ for each of the $\mathcal{U}_{k\ell}$ units it purchases from $\mathcal{I}_{k\ell}$. As far as running time is concerned, it is sufficient to evaluate the expected revenue for each of the guesses mentioned above and pick the best one, which indeed amounts to quasi-polynomial time when the purchasing costs and selling prices are not extremely far apart; $c_n/c_1 = O(n^{\mathrm{polylog}(n)})$ and $p_n/p_1 = O(n^{\mathrm{polylog}(n)})$ is enough.

## A.2. Computing the Expected Revenue

Suppose that the number of arriving customers, $M$, has a finite support, say on $\{0, \ldots, T\}$. We argue that the expected revenue of every inventory vector $(y_1, \ldots, y_n)$ can be computed in time polynomial in $n$ and $T$. We begin by pointing out that since

$$\mathrm{E}\left[R_M(y_1, \ldots, y_n)\right] = \sum_{m=0}^{T} \Pr\left[M = m\right] \cdot \mathrm{E}\left[R_m(y_1, \ldots, y_n)\right] ,$$

it is sufficient to evaluate the expected revenue for each fixed number of customers. For this purpose, number the units stocked in increasing order of product indices as $1, \ldots, \sum_{i=1}^{n} y_i$, and denote by $\mathrm{type}(u)$ the product type of unit $u$. In particular, when $u \in \{1, \ldots, y_1\}$ we have $\mathrm{type}(u) = 1$, when $u \in \{y_1+1, \ldots, y_1+y_2\}$ we have $\mathrm{type}(u) = 2$, and so on. For a unit $1 \leq u \leq \sum_{i=1}^{n} y_i$ and a customer $1 \leq k \leq m$, let $P[u,k]$ be the probability that unit $u$ is consumed by customer $k$. With this definition in mind, note that

$$\mathrm{E}\left[R_m(y_1, \ldots, y_n)\right] = \sum_{u=1}^{\sum_{i=1}^{n} y_i} \sum_{k=1}^{m} P[u,k] \cdot p_{\mathrm{type}(u)} ,$$

implying that it remains to compute the probabilities $P[u,k]$, a task that can be accomplished by means of dynamic programming, as

$$P[u,k] = \sum_{\ell=1}^{k-1} P[u-1,\ell] \cdot \left(1-\alpha_{\mathrm{type}(u)}\right)^{k-\ell-1} \cdot \alpha_{\mathrm{type}(u)} .$$

*Making the support of $M$ finite.* In the remainder of this section, we show that even though the number of customers $M$ may not be upper bounded, we can still define a corresponding random variable $\tilde{M}$ that satisfies the following properties:

1. $\tilde{M} \in \{0, \ldots, \lceil \frac{p_n C \mathrm{E}[M]}{\epsilon \mathrm{OPT}} \rceil\}$.
2. $\tilde{M}$ is stochastically smaller than $M$.
3. For any inventory vector $(y_1, \ldots, y_n)$ with capacity at most $C$, we have

$$\mathrm{E}\left[R_{\tilde{M}}(y_1, \ldots, y_n)\right] - \mathrm{E}\left[R_M(y_1, \ldots, y_n)\right] \leq \epsilon \cdot \mathrm{OPT} \; .$$

To this end, we define $\tilde{M}$ by truncating the original number of customers at $\lceil \frac{p_n C \mathrm{E}[M]}{\epsilon \mathrm{OPT}} \rceil$, that is, $\tilde{M} = \min\{M, \lceil \frac{p_n C \mathrm{E}[M]}{\epsilon \mathrm{OPT}} \rceil\}$. In this case,

$$\begin{aligned}
\mathrm{E}\left[R_{\tilde{M}}(y_1, \ldots, y_n)\right] &= \Pr\left[\tilde{M} \leq \left\lceil \frac{p_n C \mathrm{E}\left[M\right]}{\epsilon \mathrm{OPT}} \right\rceil\right] \cdot \mathrm{E}\left[R_{\tilde{M}}(y_1, \ldots, y_n) \,\middle|\, \tilde{M} \leq \left\lceil \frac{p_n C \mathrm{E}\left[M\right]}{\epsilon \mathrm{OPT}} \right\rceil\right] \\
&\quad + \Pr\left[\tilde{M} > \left\lceil \frac{p_n C \mathrm{E}\left[M\right]}{\epsilon \mathrm{OPT}} \right\rceil\right] \cdot \mathrm{E}\left[R_{\tilde{M}}(y_1, \ldots, y_n) \,\middle|\, \tilde{M} > \left\lceil \frac{p_n C \mathrm{E}\left[M\right]}{\epsilon \mathrm{OPT}} \right\rceil\right] \\
&\leq \mathrm{E}\left[R_M(y_1, \ldots, y_n)\right] + \frac{\epsilon \mathrm{OPT}}{p_n C} \cdot C p_n \\
&= \mathrm{E}\left[R_M(y_1, \ldots, y_n)\right] + \epsilon \mathrm{OPT} \; ,
\end{aligned}$$

where the inequality above follows from applying Markov's inequality, and since the expected revenue from any inventory vector with capacity at most $C$ is clearly upper bounded by $Cp_n$. It is worth mentioning that, for our particular purposes, the inventory vectors whose revenue needs to be evaluated do not contain any cheap items (see proof of Lemma 1), implying that $\mathrm{OPT} \geq \epsilon^2 C p_F$. Consequently, the newly-defined random variable $\tilde{M}$ is actually upper bounded by $\lceil \frac{p_n C \mathrm{E}[M]}{\epsilon \mathrm{OPT}} \rceil \leq \lceil \frac{\mathrm{E}[M]}{\epsilon^3} \cdot \frac{p_n}{p_F} \rceil$.

### A.3. Sketch of the Polynomial Time Algorithm for Quality Categories

We can assume without loss of generality that a polynomial-factor estimate $\widetilde{\mathrm{OPT}} \in [\mathrm{OPT}, k \cdot \mathrm{OPT}]$ of the optimal objective value is known in advance. To enforce this assumption, one can separately apply our basic algorithm (described in Section 4) to each of the quality classes $1, \ldots, k$, and consider the one whose expected revenue is maximized. Clearly, this revenue resides within the interval $[\mathrm{OPT}/k, \mathrm{OPT}]$, and we can therefore obtain $\widetilde{\mathrm{OPT}}$ with appropriate scaling.

For every $1 \leq \ell \leq k$ and $\phi \in [0, \widetilde{\mathrm{OPT}}]$, let $h(\ell, \phi)$ be the minimum number of units to stock from classes $1, \ldots, \ell$ such that an overall expected revenue of at least $\phi$ is achieved. If we were able to compute $h(\ell, \phi)$ for all possible values of $\ell$ and $\phi$, then the optimal objective value would obviously be $\max\{\phi : h(k, \phi) \leq C\}$. However, since $\phi$ is continuous, we will limit this parameter to take integer multiples of $\Delta = \epsilon \widetilde{\mathrm{OPT}}/k$, noting that if $t$ is the maximal integer for which $t\Delta \leq \phi$, then $h(k, t\Delta) \leq h(k, \phi)$ whereas the revenue loss is at most $\phi - t\Delta \leq \Delta = \epsilon \widetilde{\mathrm{OPT}}/k \leq \epsilon \mathrm{OPT}$. With this discretization in mind, we will focus on computing an approximate version $\widetilde{h}(\cdot, \cdot)$ of the function $h(\cdot, \cdot)$, satisfying the following properties: (1) $\widetilde{h}(\ell, t\Delta) \leq h(\ell, t\Delta)$ for every $1 \leq \ell \leq k$ and $0 \leq t \leq \lceil k/\epsilon \rceil$; and (2) it is possible to stock at most $\widetilde{h}(\ell, t\Delta)$ units from classes $1, \ldots, \ell$ such that an overall expected revenue of at least $(1-\epsilon)t\Delta - \ell\Delta$ is achieved.

*Computing $\widetilde{h}(1, t\Delta)$.* These values can be determined by applying our basic algorithm to the quality class 1. To ensure that the running time is polynomial in $\log C$, instead of testing $0, \ldots, C$ as the potential capacity of class 1, we conduct a binary search over this interval to find the minimal number of units that obtain an expected revenue of at least $(1-\epsilon)t\Delta$ . Clearly, $\widetilde{h}(1, t\Delta)$ will satisfy properties (1) and (2) above, since $h(1, t\Delta)$ units are sufficient to obtain an expected revenue of at least $t\Delta$ (by definition), and since the performance guarantee of our algorithm is $(1-\epsilon)$.

*Computing* $\widetilde{h}(\ell, t\Delta)$ *for* $\ell \geq 2$. In this case, for every $0 \leq t' \leq t$, we try to obtain an expected revenue of at least $t'\Delta$ from classes $1, \ldots, \ell - 1$ (using $\widetilde{h}(\ell - 1, t'\Delta)$), as well as an expected revenue of at least $(t - t')\Delta$ from class $\ell$, as explained in the previous item. It remains to show that this method is good enough to have $\widetilde{h}(\ell, t\Delta)$ satisfy properties (1) and (2). For this purpose, suppose that to achieve an expected revenue of $t\Delta$ from classes $1, \ldots, \ell$, one has to stock $u$ units of classes $1, \ldots, \ell - 1$ (with revenue $\phi$) and $h(\ell, t\Delta) - u$ units of class $\ell$ (with revenue at least $t\Delta - \phi$). Let $t'$ be the maximal integer for which $t'\Delta \leq \phi$. Then, in terms of capacities, we will stock at most $\widetilde{h}(\ell - 1, t'\Delta) \leq h(\ell - 1, t'\Delta) \leq h(\ell - 1, \phi) \leq u$ units of classes $1, \ldots, \ell - 1$, and at most $h(\ell, t\Delta) - u$ units of class $\ell$. In addition, we are guaranteed to obtain an expected revenue of at least $(1 - \epsilon)t'\Delta - (\ell - 1)\Delta + (1 - \epsilon)(t - t' - 1)\Delta \geq (1 - \epsilon)t\Delta - \ell\Delta$.