Assortment and Inventory Planning under Stockout-Based Substitution: The Many-Products Regime

Jingwei Zhang[1], Will Ma[1], Huseyin Topaloglu[2]

[1]Graduate School of Business, Columbia University, New York, NY 10027
[2]School of Operations Research and Information Engineering, Cornell Tech, New York, NY 10044
jz3608@columbia.edu, wm2428@gsb.columbia.edu, topaloglu@orie.cornell.edu

April 12, 2023

We study the joint assortment and inventory planning problem with stockout-based substitution. In this problem, we pick the number of units to stock for the products at the beginning of the selling horizon. Each arriving customer makes a choice among the set of products with remaining on-hand inventories. Our goal is to pick the stocking quantities to maximize the total expected revenue from the sales net of the stocking cost. Using a fluid approximation for the problem, we give solutions with performance guarantees that significantly improve earlier results. Letting T be the number of time periods in the selling horizon and n be the number of products, when customers choose under a general choice model, we show that we can round the solution to the fluid approximation to obtain stocking quantities with an optimality gap of $O(n + \sqrt{nT})$, improving earlier optimality gaps by a logarithmic factor. More importantly, when customers choose under the multinomial logit model, we develop a rounding scheme that uses the solution to the fluid approximation to generate stocking quantities with an optimality gap of $O(\log T \sqrt{T \log T})$. The optimality gap that we give under the multinomial logit model is the first one that does not depend on the number of products. Such an optimality gap has important implications in the many-products regime. Earlier results cannot guarantee that the stocking quantities generated by the fluid approximation perform well when both the demand volume and number of products are large, which is a regime becoming more relevant for online retail applications with large product variety. In contrast, we can guarantee that the stocking quantities generated by our rounding scheme perform well when both the demand volume and number of products are large.

Key words: assortment optimization, multinomial logit model, fluid approximation.

---

1. Introduction

One of the common challenges faced by the retailers involves deciding which products to offer and how much inventory to stock at the beginning of a selling season when the customers choose and substitute among the products that are available to them. There are multiple tradeoffs to balance in such a problem setting. Due to the substitution possibilities, when the inventory for a product runs out, its demand shifts to other products, so it is difficult to quantify the demand for each product a priori without knowing the stocking quantities. On the other hand, the stocking quantities should, in turn, depend on the demand for the products. Therefore, there is a two-way interaction between the demand for each product and stocking quantities. Moreover, even putting the inventory considerations aside, finding the right variety of products to offer to customers is a non-trivial problem. If the product variety is too low, then a large portion of the customers may not find what they are looking for and leave without a purchase. If the product variety is too high, then there may be too many low-margin products introduced into the the mix, which may end up cannibalizing the sales of high-margin products. These considerations point out that finding the right product variety and stocking quantities requires considering how the demand for each product is shaped by the inventories of all products and the substitution behavior.

Two types of substitution models have attracted attention. The first substitution model is static substitution, which is used when customers cannot observe the product availability, as in online retail. Under static substitution, customers choose within the product assortment without knowing the real-time inventories. If the customer chooses a product without on-hand inventory, then she may leave without a purchase or be served through an emergency procurement. Static substitution takes its name due to the fact that the set of products among which a customer chooses does not depend on real-time inventories. Such a substitution model simplifies the assortment and inventory planning problem significantly. Due to the goodwill cost, however, it may not be desirable to let a customer choose a product without on-hand inventory. Also, customers in brick-and-mortar retail do see the real-time inventories on the shelf when making their choices. Thus, the second substitution model is stockout-based substitution, where customers choose only among the products with on-hand inventories. Under stockout-based substitution, the real-time inventories influence the choices, so the choice probabilities for the products evolve with the on-hand inventories.

Assortment and inventory planning under stockout-based substitution is a notoriously difficult problem that has been of interest to academics and practitioners for a long time. One approach is to calculate stocking quantities using a fluid approximation formulated under the assumption that the choices of the customers take on their expected values. There are three papers critical to us that use this approach. Honhon and Seshadri (2013) consider customers choosing under a general choice model. Letting T be the number of time periods in the selling horizon and n be the number of products, the authors show that implementing the optimal solution of the fluid approximation yields an optimality gap of $O(n\sqrt{T})$. El Housni et al. (2021) also study the problem under a general choice model. Using a fluid approximation, along with a sample average approximation approach, the authors obtain a solution with an optimality gap of $O(n + \sqrt{nT\log(nT)})$. Liang et al. (2022) focus on the case where the customers choose under the multinomial logit model. The authors use a fluid approximation to obtain a solution with an optimality gap of $O(\sqrt{nT\log(nT)})$ when the number of products is smaller than the number of time periods in the selling horizon.

Our Contributions: In many problem settings, it is straightforward to argue that the optimal total expected profit increases linearly with T. In the three papers discussed in the previous paragraph, the optimality gaps of the proposed solutions also increase in T, but sublinearly. In this case, the relative optimality gaps of the proposed solutions in the three papers become arbitrarily close to zero as the number of time periods in the selling horizon gets large. In other words, we have some theoretical assurance for using the solutions in these three papers under large demand volume, all else being equal. On the other hand, it is not possible to argue that the optimal total expected profit increases linearly with n. In fact, it is straightforward to come up with examples where the optimal total expected profit remains constant in n. The optimality gaps of the proposed solutions in the three papers above increase in n. In this case, the relative optimality gaps of the proposed solutions do not become arbitrarily close to zero when both the number of time periods and the number of products get large. In other words, we do not have any theoretical assurance for using the solutions in these three papers under large demand volume, when the number of products is also large. Therefore, the solutions in earlier results are for a regime where the demand volume is large but the number of products remain bounded. However, it is reasonable to think of large systems not only with large demand volume, but also with large number of products.

In this paper, we give improved optimality gaps. When customers choose under a general choice model, we show that we can round the solution from a fluid approximation to get a solution with an optimality gap of $O(n+\sqrt{nT})$. This optimality gap is tight in both T and n. Our optimality gap improves earlier results by a factor of $O(\sqrt{\log(nT)})$. This improvement may be viewed modest, but we build on this result to give much improved optimality gaps under the multinomial logit model, which is one of the most popular choice models. Under the multinomial logit model, we develop a new rounding scheme for the fluid approximation. When the preference weights of the products are not infinitesimal, we show that the solution from our rounding scheme has an optimality gap of $O(\log T\sqrt{T\log T})$, which is independent of the number of products. Our proof involves fully novel components, where we give a rounding scheme to prioritize high-margin products and use a dynamic program to bound the expected no-purchases, yielding a key closed-form bound on the expected no-purchases that exploits the special structure of the multinomial logit model.

To our knowledge, the optimality gap that we give under the multinomial logit model is the first one that is independent of the number of products. Due to the fact that our optimality gap increases sublinearly in T, whereas it does not increase in n, our rounding scheme generates solutions with relative optimality gaps arbitrarily close to zero as both the demand volume and the number of products get large. Furthermore, the optimality gap of $O(\sqrt{nT\log(nT)})$ established in earlier results is not an artifact of loose analysis. In Figure 1, as a function of the number of products in a problem instance, we plot the percent gap between the upper bound on the optimal total expected profit provided by the fluid approximation and the total expected profit obtained by two stocking quantities. The solid line plots the gap for the stocking quantities obtained by our rounding approach, whereas the dotted line plots the gap for the stocking quantities obtained

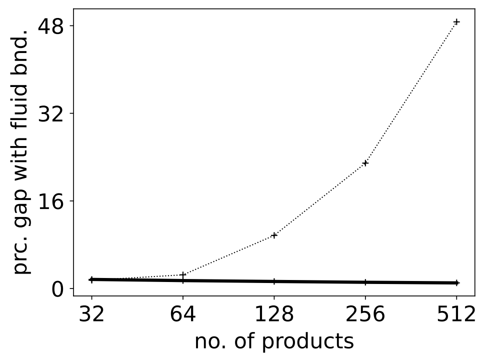

Figure 1 Optimality gaps provided by two socking quantities as a function of the number of products.

by the rounding approach in Liang et al. (2022). Indeed, the gap for our stocking quantities remain stable, whereas the gap for the stocking quantities in Liang et al. (2022) increases in n. Thus, the fact that our rounding scheme yields optimality gaps independent of n is not a technical curiosity. We can compute stocking quantities that perform significantly better.

Other Related Literature: The paper by van Ryzin and Mahajan (1999) studies joint assortment and inventory planning under the multinomial logit model with static substitution and products with identical margins, characterizing the structure of the optimal solution. Cachon et al. (2005) incorporate customer search costs. Topaloglu (2013) relaxes the assumption of identical product margins. Moving on to work under dynamic substitution, Mahajan and van Ryzin (2001) give a stochastic gradient algorithm to compute locally optimal stocking quantities. Gaur and Honhon (2006) use a locational choice model and obtain an upper bound. Hopp and Xu (2008) propose a static approximation for substitution behavior that is based on a fluid network model. Honhon et al. (2010) give a fluid approximation and solve it in $O(8^n)$ operations. Lastly, we distinguish our work from other two streams of literature. The first stream is online assortment optimization, where the assortment of products offered to each customer can be adjusted by the retailer; see, for example, Golrezaei et al. (2014), Rusmevichientong et al. (2020) and Ma et al. (2021). In our problem, the set of products available for the customers is automatically implied by on-hand inventories. The second stream is assortment and inventory planning under stockout-based substitution with an upper bound on the total units stocked; see, for example, Goyal et al. (2016), Aouad et al. (2018), Aouad et al. (2019) and Aouad and Segev (2022). In this problem, stocking costs are ignored but there is an upper bound on the total units stocked. Exploiting the fact that the objective function does not involve a cost component, the focus is on developing multiplicative performance guarantees.

Outline: In Section 2, we formulate the problem and discuss the fluid approximation that drives our stocking quantities. In Section 3, we give our optimality gap under a general choice model. In Section 4, we give our optimality gap under the multinomial logit model. In Section 5, we illustrate our results numerically. In Section 6, we conclude.

2. Problem Formulation

We have n products indexed by $\mathcal{N} = \{1, \ldots, n\}$. Selling a unit of product i yields a revenue of p_i. We procure each unit of product i at a cost of c_i. There are T time periods in the selling horizon indexed by $\mathcal{T} = \{1, \ldots, T\}$. For notational brevity, we follow the convention that there is one customer arrival at each time period. At the beginning of the selling horizon, we decide how many units of each product to stock. Each customer arriving into the system makes a choice within the set of products with remaining inventories, either purchasing one unit of a product or deciding to leave without a purchase. In particular, if the set of products with remaining inventories at a certain time period is S, then the customer purchases product i with probability $\phi_i(S)$. Naturally, we have $\phi_i(S) = 0$ for all $i \notin S$. With probability $\phi_0(S) = 1 - \sum_{i\in\mathcal{N}} \phi_i(S)$, the customer leaves without a purchase. We assume that we have $\phi_i(S \setminus \{j\}) \geq \phi_i(S)$ for all $i, j \in S$, so that dropping a product from the set of available products increases the purchase probability of all other available products. This property is called the substitutability property and it is satisfied by all choice models that are based on random utility maximization. Due to the substitutability property, the purchase probability of a product is non-decreasing over the time periods until it is sold out.

Our goal is to decide how many units of each product to stock at the beginning of the selling horizon to maximize the total expected profit. We use the vector $\boldsymbol{q} = (q_1, \ldots, q_n) \in \mathbb{Z}_+^n$ to capture the inventories on-hand at the beginning of a generic time period, where q_i is the inventory on-hand for product i. If the inventories on-hand at the current time period are $\boldsymbol{q}$, then the set of available products is given by $A(\boldsymbol{q}) = \{i \in \mathcal{N} : q_i > 0\}$, in which case, the sale for product i at the current time period is a Bernoulli random variable with parameter $\phi_i(A(\boldsymbol{q}))$. Given that the inventories on-hand at the current time period are $\boldsymbol{q}$, we use the random vector $\boldsymbol{D}(\boldsymbol{q}) = (D_1(\boldsymbol{q}), \ldots, D_n(\boldsymbol{q}))$ to capture the sales for the products, where $D_i(\boldsymbol{q})$ is the sale for product i. Using $\boldsymbol{e}_i \in \mathbb{Z}_+^n$ to denote the i-th unit vector, note that we have $\boldsymbol{D}(\boldsymbol{q}) = \boldsymbol{e}_i$ with probability $\phi_i(A(\boldsymbol{q}))$. We use the random variable $X_i(\boldsymbol{q}, t)$ to capture the total sales for product i over τ time periods given that the inventories on-hand at the beginning of these τ time periods are $\boldsymbol{q}$. Therefore, the random variable $X_i(\boldsymbol{q}, \tau)$ is recursively defined as $X_i(\boldsymbol{q}, \tau) = D_i(\boldsymbol{q}) + X_i(\boldsymbol{q} - \boldsymbol{D}(\boldsymbol{q}), \tau - 1)$ with the boundary condition that $X_i(\boldsymbol{q}, 0) = 0$. In this case, if the stocking quantities at the beginning of the selling horizon are chosen to be $\boldsymbol{q}$, then the total expected profit that we obtain is given by

$$\Pi(\boldsymbol{q}) \;=\; \sum_{i\in\mathcal{N}} p_i\, \mathbb{E}\{X_i(\boldsymbol{q}, T)\} - \sum_{i\in\mathcal{N}} c_i\, q_i. \tag{1}$$

The optimal total expected profit is $\mathsf{opt} = \max_{\boldsymbol{q}\in\mathbb{Z}_+^n} \Pi(\boldsymbol{q})$. Putting aside solving the last optimization problem, to our knowledge, computing $\Pi(\boldsymbol{q})$ at fixed stocking quantities requires solving a dynamic program with a high-dimensional state variable keeping track of the remaining inventories of each product. We use a fluid approximation to get a tractable benchmark.

Upper Bound from a Fluid Approximation:

Instead of comparing the total expected profit from the approximate stocking quantities that we compute with opt, we will compare it with an upper bound on opt. Consider the problem

$$V^{\mathsf{fluid}} \;=\; T \max_{S\subseteq\mathcal{N}} \left\{ \sum_{i\in\mathcal{N}} \phi_i(S)\,(p_i - c_i) \right\}. \tag{2}$$

The maximization problem above finds a subset of products to offer to maximize the expected profit extracted from the customer arriving at each time period. Using S^{fluid} to denote an optimal solution to the maximization problem above, we define the stocking quantities $\boldsymbol{q}^{\mathsf{fluid}} = (q_1^{\mathsf{fluid}}, \ldots, q_n^{\mathsf{fluid}})$ as $q_i^{\mathsf{fluid}} = T\,\phi_i(S^{\mathsf{fluid}})$ for all $i\in\mathcal{N}$. In this case, if the sale for product i at a time period were to take the deterministic and fractional value of $\phi_i(S^{\mathsf{fluid}})$, then stocking q_i^{fluid} units of product i would imply that the inventories of all products are depleted simultaneously at the end of the selling horizon. Meanwhile, we would be extracting the maximum expected profit from the customer arriving at each time period. Thus, we can show that V^{fluid} is an upper bound on the optimal total expected profit. That is, we have $V^{\mathsf{fluid}} \geq \mathsf{opt}$; see Proposition 2.1 in El Housni et al. (2021). This upper bound on the optimal total expected profit is computed under the assumption that the demand for each product is deterministic and fractional, so we refer to this upper bound on the optimal total expected profit as the fluid upper bound. While it is difficult to compute the optimal total expected profit efficiently, we can often compute the fluid upper bound efficiently. In particular, problem (2) is of combinatorial nature, but we can solve this problem efficiently when the choices of the customers are governed by a variety of choice models, including the multinomial logit, generalized attraction, nested logit, multi-level nested logit and Markov chain choice model; see Talluri and van Ryzin (2004), Davis et al. (2014), Gallego et al. (2015), Li et al. (2015) and Blanchet et al. (2016). For any approximate stocking quantity $\boldsymbol{q}$, the optimality gap of this approximate stocking quantity is given by $\mathsf{opt} - \Pi(\boldsymbol{q})$. Noting that $\mathsf{opt} - \Pi(\boldsymbol{q}) \leq V^{\mathsf{fluid}} - \Pi(\boldsymbol{q})$, to upper bound the optimality gap of the approximate stocking quantity, it will be enough to upper bound $V^{\mathsf{fluid}} - \Pi(\boldsymbol{q})$.

In the next section, we consider the case where the customers choose under a general choice model and give a performance bound for the solutions from the fluid upper bound.

3. Performance Guarantee under a General Choice Model

We consider the case where the choices of the customers are governed by a general choice model and give a performance guarantee for stocking quantities from the fluid upper bound. We obtain the fluid upper bound by using stocking quantities that are depleted simultaneously for all products at the end of the selling horizon. Thus, the set of products with on-hand inventories do not change throughout the selling horizon. Motivated by this observation, we focus on a problem that is formulated under the assumption that the customers always make a choice among the products that are stocked at the beginning of the selling horizon, but if the product they choose does not have on-hand inventory anymore, then they leave without a purchase. If the stocking quantities are $\boldsymbol{q}$, then the set of stocked products is given by $A(\boldsymbol{q})$. A customer chooses product i among this set of products with probability $\phi_i(A(\boldsymbol{q}))$. We use the random vector $\boldsymbol{C}_t(\boldsymbol{q}) = (C_{1t}(\boldsymbol{q}), \ldots, C_{nt}(\boldsymbol{q}))$ to capture the choice of the customer arriving at time period t among the set of products $A(\boldsymbol{q})$. Thus, we have $\boldsymbol{C}_t(\boldsymbol{q}) = \boldsymbol{e}_i$ with probability $\phi_i(A(\boldsymbol{q}))$. Under the assumption that the customers always make a choice among the products that are stocked at the beginning of the selling horizon, if we stock the quantities $\boldsymbol{q}$, then the total expected profit is

$$\Pi^{\mathsf{static}}(\boldsymbol{q}) \;=\; \sum_{i\in\mathcal{N}} p_i\, \mathbb{E}\left\{ \min\left\{ q_i, \sum_{t\in\mathcal{T}} C_{it}(\boldsymbol{q}) \right\}\right\} - \sum_{i\in\mathcal{N}} c_i\, q_i. \tag{3}$$

In the expression in (3), the set of products among which the customers choose does not change during the course of the selling horizon, so we refer (3) as the static approximation. We can argue that the purchase probability of each product at each time period in (3) is no larger than its counterpart in (1). In particular, the customers choose among the set of products stocked at the beginning of the selling horizon in (3), whereas the customers choose among the set of products with on-hand inventories in (1). Thus, by the assumption that $\phi_i(S) \le \phi_i(S \setminus \{j\})$ for all $i, j \in S$, the purchase probability of each product at each time period in (3) is no larger than its counterpart in (1). In this case, the total expected profit in (3) is no larger than the total expected profit in (1), so $\Pi^{\mathsf{static}}(\boldsymbol{q}) \le \Pi(\boldsymbol{q})$ for all $\boldsymbol{q} \in \mathbb{Z}_+^n$. Recalling that S^{fluid} is an optimal solution to problem (2) and $q_i^{\mathsf{fluid}} = T\,\phi_i(S^{\mathsf{fluid}})$, using $\lceil\cdot\rceil$ to denote the round up operator, we define the stocking quantities $\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil = (\lceil q_1^{\mathsf{fluid}} \rceil, \ldots, \lceil q_n^{\mathsf{fluid}} \rceil)$. In the next theorem, we give a performance guarantee for the stocking quantities $\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil$. Throughout the rest of the paper, we set $\overline{p} = \max_{i\in\mathcal{N}} p_i$ and $\overline{c} = \max_{i\in\mathcal{N}} c_i$.

Theorem 3.1 *If $\overline{p}$ and $\overline{c}$ are independent of n and T, then there exist absolute constants C_1 and C_2 such that*

$$V^{\mathsf{fluid}} - \Pi^{\mathsf{static}}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil) \;\le\; C_1\, n + C_2 \sqrt{nT}.$$

We give the proof of the theorem in Appendix A. Using the fact that $V^{\mathsf{fluid}} \ge \mathsf{opt}$ and $\Pi^{\mathsf{static}}(\boldsymbol{q}) \le \Pi(\boldsymbol{q})$ for all $\boldsymbol{q} \in \mathbb{Z}_+^n$, Theorem 3.1 implies that $\mathsf{opt} - \Pi(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil) \le C_1\, n + C_2 \sqrt{nT}$, so the optimality gap of the stocking quantities $\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil$ is upper bounded by $C_1\, n + C_2 \sqrt{nT}$. We briefly outline the proof of the theorem. In the fluid upper bound, we incur the total procurement cost $T \sum_{i \in \mathcal{N}} \phi_i(S^{\mathsf{fluid}})\, c_i = \sum_{i \in \mathcal{N}} c_i\, q_i^{\mathsf{fluid}}$. In the static upper bound, corresponding to the stocking quantities $\boldsymbol{q}^{\mathsf{fluid}}$, we incur the total procurement cost $\sum_{i \in \mathcal{N}} c_i \lceil q_i^{\mathsf{fluid}} \rceil$. The difference in the total procurement cost is upper bounded by $\bar{c}\, n$, so we are left to analyze the difference in the total expected revenue. In the fluid upper bound, the total expected revenue from product i is given by $T\, \phi_i(S^{\mathsf{fluid}})\, p_i = p_i\, q_i^{\mathsf{fluid}}$. In the static upper bound, corresponding to the stocking quantities $\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil$, the total expected revenue from product i is $p_i\, \mathbb{E}\{\min\{\lceil q_i^{\mathsf{fluid}} \rceil, \sum_{t \in \mathcal{T}} C_{it}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil)\}\}$. Thus, the difference between the total expected revenue from product i is given by

$$p_i \left(q_i^{\mathsf{fluid}} - \mathbb{E}\left\{ \min \left\{ \lceil q_i^{\mathsf{fluid}} \rceil, \sum_{t \in \mathcal{T}} C_{it}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil) \right\} \right\} \right).$$

Using the Jensen inequality, we upper bound the difference by $\overline{p}\, \sqrt{\phi_i(S^{\mathsf{fluid}})\, T}$, in which case, using this upper bound for all products with Cauchy-Schwarz inequality yields the result.

If the number of time periods in the selling horizon is larger than the number of products, then Theorem 3.1 gives an optimality gap of $O(\sqrt{nT})$. We shortly give an example where this optimality gap is tight, followed by another example where we improve this optimality gap when the products are highly substitutable. Theorem 3.1 makes use of the static upper bound, which ignores the substitution behavior based on the on-hand product inventories. El Housni et al. (2021) argue that ignoring substitution behavior based on the on-hand product inventories can lead to large optimality gaps and use sample average approximation to get an optimality gap of $O(n + \sqrt{nT\, \log(nT)})$. Theorem 3.1 improves the last optimality gap by a factor of $O(\sqrt{\log(nT)})$ when the number of time periods is larger than the number of products.

Independent Demand Model:

We give an example where the optimality gap in Theorem 3.1 is tight. Using $\mathbf{1}(\cdot)$ to denote the indicator function, the choice probabilities are given by $\phi_i(S) = \mathbf{1}(i \in S)\, \frac{1}{1+n}$. Thus, if a product is available, its choice probability is $\frac{1}{1+n}$, irrespective of what other products are available. The unit revenues and procurement cost are respectively $p_i = 2$ and $c_i = 1$ for all $i \in \mathcal{N}$. We assume that $\frac{T}{1+n}$ is an integer. The optimal solution to problem (2) is $S^{\mathsf{fluid}} = \mathcal{N}$, so we have $q_i^{\mathsf{fluid}} = \frac{T}{1+n}$ for all $i \in \mathcal{N}$ and $V^{\mathsf{fluid}} = T\, \frac{n}{1+n}$. We compute the static approximation to the total expected profit at the stocking quantities $\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil$, which is given by $\Pi^{\mathsf{static}}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil)$. Since $q_i^{\mathsf{fluid}} > 0$ for all $i \in \mathcal{N}$, $A(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil) = \mathcal{N}$, so $\phi_i(A(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil)) = \frac{1}{1+n}$. Therefore, $C_{it}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil)$ is a Bernoulli random variable with parameter $\frac{1}{1+n}$, so $\sum_{t \in \mathcal{T}} C_{it}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil)$ is a binomial random variable with parameters $(T, \frac{1}{1+n})$. Using $\mathsf{binomial}(k, p)$ to denote a binomial random variable with parameters (k,p), Lemma A.1 in Appendix A shows that $\mathbb{E}\{[k\,p - \mathsf{binomial}(k,p)]^+\} \geq \frac{1}{\sqrt{2\pi}}\sqrt{k\,p\,(1-p)} - O(1)$. Thus, we obtain

$$\begin{aligned}
V^{\mathsf{fluid}} - \Pi^{\mathsf{static}}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil) &= \frac{Tn}{1+n} - n\left[2\,\mathbb{E}\left\{\min\left\{\frac{T}{1+n}, \mathsf{binomial}\Big(T, \frac{1}{1+n}\Big)\right\}\right\} - \frac{T}{1+n}\right] \\
&= 2\frac{Tn}{1+n} - 2n\,\mathbb{E}\left\{\min\left\{\frac{T}{1+n}, \mathsf{binomial}\Big(T, \frac{1}{1+n}\Big)\right\}\right\} \\
&= 2n\,\mathbb{E}\left\{\left[\frac{T}{1+n} - \mathsf{binomial}\Big(T, \frac{1}{1+n}\Big)\right]^+\right\} \geq 2n\left(\frac{1}{\sqrt{2\pi}}\sqrt{T\frac{n}{(1+n)^2}} - O(1)\right) = \Omega(\sqrt{nT} - n),
\end{aligned}$$

so if the number of time periods in the selling horizon is larger than the number of products, then we have $V^{\mathsf{fluid}} - \Pi^{\mathsf{static}}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil) = \Omega(\sqrt{nT})$.

Fully Substitutable Demand Model:

We give an example where we can improve the optimality gap in Theorem 3.1 under a choice model where the products are fully substitutable. This example suggests that the optimality gap in Theorem 3.1 is, in general, tight, but by focusing on choice models with a special structure, we can improve the optimality gap. The choice probabilities are given by $\phi_i(S) = \frac{n}{1+n}\,\mathbf{1}(i \in S)\,\frac{1}{|S|}$. We interpret these choice probabilities as follows. A customer is interested in making a purchase with probability $\frac{n}{1+n}$. If interested in making a purchase, then the customer chooses any of the available products with equal probability. We have $p_i = 2$ and $c_i = 1$ for all $i \in \mathcal{N}$ for the unit revenues and procurement costs. We continue assuming that $\frac{T}{1+n}$ is an integer. Considering problem (2), any non-empty subset of products is an optimal solution, so we use $S^{\mathsf{fluid}} = \{1\}$. In this case, noting that $\phi_1(\{1\}) = \frac{n}{1+n}$, we get $V^{\mathsf{fluid}} = T\frac{n}{1+n}$ and $q_1^{\mathsf{fluid}} = \frac{Tn}{1+n}$, $q_i^{\mathsf{fluid}} = 0$ for all $i = 2, \ldots, n$. Furthermore, $\sum_{t \in \mathcal{T}} C_{1t}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil)$ is a binomial random variable with parameters $(T, \frac{n}{1+n})$. Lemma 1 Gallego and Moon (1993) shows that $\mathbb{E}\{[k\,p - \mathsf{binomial}(k,p)]^+\} \leq \frac{1}{2}\sqrt{k\,p\,(1-p)}$. Thus, we obtain

$$\begin{aligned}
V^{\mathsf{fluid}} - \Pi^{\mathsf{static}}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil) &= \frac{Tn}{1+n} - \left[2\,\mathbb{E}\left\{\min\left\{\frac{Tn}{1+n}, \mathsf{binomial}\Big(T, \frac{n}{1+n}\Big)\right\}\right\} - \frac{Tn}{1+n}\right] \\
&= 2\frac{Tn}{1+n} - 2\,\mathbb{E}\left\{\min\left\{\frac{Tn}{1+n}, \mathsf{binomial}\Big(T, \frac{n}{1+n}\Big)\right\}\right\} \\
&= 2\,\mathbb{E}\left\{\left[\frac{Tn}{1+n} - \mathsf{binomial}\Big(T, \frac{n}{1+n}\Big)\right]^+\right\} \leq \sqrt{T\frac{n}{(1+n)^2}} = O(\sqrt{T/n}),
\end{aligned}$$

so the optimality gap of the stocking quantities $\boldsymbol{q}^{\mathsf{fluid}}$ is $O(\sqrt{T/n})$. Intuitively speaking, for n large, each customer makes a purchase with probability one, making the problem much easier.

For both examples, because of the choice of the unit revenues and procurement costs, $\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil$ happens to be the optimal stocking quantities. By the first example, the optimality gap of the fluid approximation, in general, cannot be improved beyond $O(\sqrt{nT})$, but we can get better optimality gaps under choice models with a special structure.

4. Performance Guarantee under the Multinomial Logit Model

We focus on the case where the choices of the customers are governed by the multinomial logit model and show that we can obtain stocking quantities from the fluid upper bound with optimality gap independent of the number of products. By the discussion right after Theorem 3.1, there are two contributors to the optimality gap given in the theorem. First, the fluid upper bound works with the stocking quantities $\boldsymbol{q}^{\mathsf{fluid}}$, whereas the static approximation works with the stocking quantities $\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil$. Thus, the first contributor is the rounding error. Second, the fluid upper bound assumes that the demand for product i at each time period takes the deterministic and fractional value $\phi_i(S^{\mathsf{fluid}})$, whereas the static approximation assumes that the demand for product i at each time period is a Bernoulli random variable with parameter $\phi_i(S^{\mathsf{fluid}})$. Under the multinomial logit model, we tighten our analysis to deal with both contributors to the optimality gap more effectively. Instead of rounding up the stocking quantities from the fluid approximation for all products, we develop a rounding scheme that can round up or down the stocking quantity for each product. Furthermore, we use the specific substitution behavior under the multinomial logit model to analyze the stochastic evolution of the demand over the selling horizon.

Multinomial logit model is arguably one of the most popular choice models, both in practical applications and academic research. Under this choice model, a customer associates a preference weight of w_i with product i and a preference weight of w_0 with the no-purchase option. If the set of available products is S, then a customer chooses product i with probability $\phi_i(S) = \mathbf{1}(i \in S) \frac{w_i}{w_0 + \sum_{j \in S} w_j}$. Consider computing the fluid upper bound in (2) when customers choose according to the multinomial logit model. It is known that the optimal solution to problem (2) is margin-ordered in the sense that S^{fluid} offers a certain number of products with the largest margins. In particular, without loss of generality, we index the products such that $p_1 - c_1 \geq p_2 - c_2 \geq \ldots \geq p_n - c_n$. In this case, the optimal solution to problem (2) is of the form $S^{\mathsf{fluid}} = \{1, 2, \ldots, j\}$ for some $j \in \mathcal{N}$. Thus, we can solve problem (2) by checking the objective value provided by each solution of the form $\{1, 2, \ldots, j\}$ and there are at most n solutions of this form. Once we obtain S^{fluid} efficiently in this fashion, we continue computing the stocking quantities from the fluid approximation as $q_i^{\mathsf{fluid}} = T\, \phi_i(S^{\mathsf{fluid}}) = T\, \mathbf{1}(i \in S^{\mathsf{fluid}}) \frac{w_i}{w_0 + \sum_{j \in S^{\mathsf{fluid}}} w_j}$. Next, we describe a procedure to round this solution to obtain integer stocking quantities.

Rounding Procedure:

Using $\lfloor \cdot \rfloor$ to denote the round down operator, the rounding budget is $\delta = \left\lceil \sum_{i \in S^{\mathsf{fluid}}} (q_i^{\mathsf{fluid}} - \lfloor q_i^{\mathsf{fluid}} \rfloor) \right\rceil$. The rounded stocking quantities $\boldsymbol{q}^{\mathsf{round}} = (q_i^{\mathsf{round}}, \ldots, q_n^{\mathsf{round}})$ are given by $q_i^{\mathsf{round}} = \lfloor q_i^{\mathsf{fluid}} \rfloor + 1$ for $i \leq \delta$ and $q_i^{\mathsf{round}} = \lfloor q_i^{\mathsf{fluid}} \rfloor$ for $i > \delta$. In the next lemma, we give properties of the rounded stocking quantities. Intuitively speaking, this lemma states that the rounded stocking quantities shift the stocking quantities in the solution to the fluid upper bound from products with lower margins to those with higher margins, while keeping the total stocking quantities nearly unchanged.

Lemma 4.1 *The rounded stocking quantities $\boldsymbol{q}^{\mathsf{round}}$ and the stocking quantities from the fluid upper bound $\boldsymbol{q}^{\mathsf{fluid}}$ satisfy the inequalities*

$$0 \le \sum_{i \in S^{\mathsf{fluid}}} (q_i^{\mathsf{round}} - q_i^{\mathsf{fluid}}) \le 1, \quad \sum_{i \in S^{\mathsf{fluid}}} (p_i - c_i)\,(q_i^{\mathsf{fluid}} - q_i^{\mathsf{round}}) \le 0, \quad \Big|\{i \in \mathcal{N} : q_i^{\mathsf{round}} > 0\}\Big| \le T + 1.$$

We give the proof of the lemma in Appendix B. To see the benefit from using the stocking quantities $\boldsymbol{q}^{\mathsf{round}}$ instead of $\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil$, setting $X_0(\boldsymbol{q}, \tau) = \tau - \sum_{i \in A(\boldsymbol{q})} X_i(\boldsymbol{q}, \tau)$, we have

$$\begin{aligned}
V^{\mathsf{fluid}} - \Pi(\boldsymbol{q}^{\mathsf{round}}) &= \sum_{i \in S^{\mathsf{fluid}}} (p_i - c_i)\, q_i^{\mathsf{fluid}} - \left[\sum_{i \in S^{\mathsf{fluid}}} p_i\, \mathbb{E}\{X_i(\boldsymbol{q}^{\mathsf{round}}, T)\} - \sum_{i \in S^{\mathsf{fluid}}} c_i\, q_i^{\mathsf{round}} \right] \\
&= \sum_{i \in S^{\mathsf{fluid}}} (p_i - c_i) \left(q_i^{\mathsf{fluid}} - q_i^{\mathsf{round}} \right) + \sum_{i \in S^{\mathsf{fluid}}} p_i \left(q_i^{\mathsf{round}} - \mathbb{E}\{X_i(\boldsymbol{q}^{\mathsf{round}}, T)\} \right) \\
&\overset{(a)}{\le} \sum_{i \in S^{\mathsf{fluid}}} (p_i - c_i) \left(q_i^{\mathsf{fluid}} - q_i^{\mathsf{round}} \right) + \overline{p} \left[\sum_{i \in S^{\mathsf{fluid}}} q_i^{\mathsf{round}} + X_0(\boldsymbol{q}^{\mathsf{round}}, T) - T \right] \\
&\overset{(b)}{=} \sum_{i \in S^{\mathsf{fluid}}} (p_i - c_i) \left(q_i^{\mathsf{fluid}} - q_i^{\mathsf{round}} \right) + \overline{p} \left\{ \mathbb{E}\{X_0(\boldsymbol{q}^{\mathsf{round}}, T)\} - T\,\phi_0(S^{\mathsf{fluid}}) \right\} + \overline{p} \sum_{i \in S^{\mathsf{fluid}}} (q_i^{\mathsf{round}} - q_i^{\mathsf{fluid}}) \\
&\overset{(c)}{\le} \overline{p} \left\{ \mathbb{E}\{X_0(\boldsymbol{q}^{\mathsf{round}}, T)\} - T\,\phi_0(S^{\mathsf{fluid}}) \right\} + \overline{p},
\end{aligned}$$

where (a) uses the definition of $X_0(\boldsymbol{q}, T)$, (b) holds because we have $q_i^{\mathsf{fluid}} = T\,\phi_i(S^{\mathsf{fluid}})$ and $\sum_{i \in S^{\mathsf{fluid}}} \phi_i(S^{\mathsf{fluid}}) + \phi_0(S^{\mathsf{fluid}}) = 1$ and (c) uses the first two inequalities in Lemma 4.1.

By the chain of inequalities above, the difference $\sum_{i \in S^{\mathsf{fluid}}} (q_i^{\mathsf{round}} - q_i^{\mathsf{fluid}})$ is $O(1)$, so this difference contributes the term $\overline{p}$ to the optimality gap $V^{\mathsf{fluid}} - \Pi(\boldsymbol{q}^{\mathsf{round}})$, which is independent of the number of products. In contrast, the difference $\sum_{i \in S^{\mathsf{fluid}}} (\lceil q_i^{\mathsf{fluid}} \rceil - q_i^{\mathsf{fluid}})$ is $O(n)$. Thus, using the stocking quantities $\boldsymbol{q}^{\mathsf{round}}$ instead of $\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil$ eliminates the contribution of the rounding error to the optimality gap. Furthermore, by the chain of inequalities above, to bound the optimality gap $V^{\mathsf{fluid}} - \Pi(\boldsymbol{q}^{\mathsf{round}})$, it is enough to upper bound the difference $\mathbb{E}\{X_0(\boldsymbol{q}^{\mathsf{round}}, T)\} - T\,\phi_0(S^{\mathsf{fluid}})$. Note that $\mathbb{E}\{X_0(\boldsymbol{q}^{\mathsf{round}}, T)\}$ is the expected number of customers without a purchase when the stocking quantities are $\boldsymbol{q}^{\mathsf{round}}$, whereas $T\phi_0(S^{\mathsf{fluid}})$ is the deterministic and fractional number of customers without a purchase under the fluid approximation. We will upper bound the difference $\mathbb{E}\{X_0(\boldsymbol{q}^{\mathsf{round}}, T)\} - T\,\phi_0(S^{\mathsf{fluid}})$ by an expression that is independent of the number of products. In this case, the optimality gap of the stocking quantities $\boldsymbol{q}^{\mathsf{round}}$ will be independent of the number of products.

In the next theorem, we give a performance guarantee for the stocking quantities $\boldsymbol{q}^{\mathsf{round}}$ under the multinomial logit model. Throughout the rest of the paper, we set $\underline{w} = \min_{i \in \mathcal{N}} w_i$.

Theorem 4.2 *Considering customers choosing under the multinomial logit model, if* $\overline{p}$ *and* $w_0/\underline{w}$ *are independent of* n *and* T*, then there exist absolute constants* C_1 *and* C_2 *such that*

$$V^{\mathsf{fluid}} - \Pi(\boldsymbol{q}^{\mathsf{round}}) \ \leq \ C_1 \log T \sqrt{T \log T} + C_2.$$

We give the proof of the theorem in Appendix B. Noting that $V^{\mathsf{fluid}} \geq \mathsf{opt}$, Theorem 4.2 implies that $\mathsf{opt} - \Pi(\boldsymbol{q}^{\mathsf{round}}) \leq C_1 \log T \sqrt{T \log T} + C_2$, so optimality gap of the stocking quantities $\boldsymbol{q}^{\mathsf{round}}$ is $O(\log T \sqrt{T \log T})$, which is independent of the number of products. We briefly outline the proof of the theorem. Exploiting the multinomial logit model, we develop a dynamic program to bound the expected no-purchases through a closed-form as $\mathbb{E}\{X_0(\boldsymbol{q}, T)\} \leq 1 + w_0 \max_{i \in \mathcal{N}} \left\{\frac{q_i}{w_i}\right\} \log \sum_{i \in \mathcal{N}} q_i + [T - \sum_{i \in \mathcal{N}} q_i]^+$ for any stocking quantities $\boldsymbol{q} \in \mathbb{Z}_+$. This inequality is the key driver of the proof of the theorem. If the number of customer arrivals is small in the sense that $T \leq K\, |S^{\mathsf{fluid}}| \log T$ for some constant K independent of n and T, then using this inequality along with the fact that $\sum_{i \in \mathcal{N}} q_i^{\mathsf{fluid}} = T \sum_{i \in \mathcal{N}} \phi_i(S^{\mathsf{fluid}}) \leq T$ in our key inequality allows us to bound $\mathbb{E}\{X_0(\boldsymbol{q}, T)\}$.

If, on the other hand, $T \geq K\, |S^{\mathsf{fluid}}| \log T$, then we proceed as follows. Using Chernoff bound, we characterize a time period $\widetilde{T}$ close to T such that no product is sold out by time period $\widetilde{T}$ with high probability. Given that none of the products is sold out by time period $\widetilde{T}$, the total number customers without a purchase by time period $\widetilde{T}$ is a binomial random variable, which allows us to show that the conditional total expected number of customers without a purchase by time period $\widetilde{T}$ is no larger than $T\, \phi_0(S^{\mathsf{fluid}})$. Lastly, given that none of the products is sold out by time period $\widetilde{T}$, we show that the on-hand inventory of each product at time period $\widetilde{T}$ is $O(\sqrt{T})$. In this case, using our key inequality over the remaining $T - \widetilde{T}$ time periods, we show that the expected number of customers without a purchase over these time periods is $O(\log T \sqrt{T \log T})$.

5. Numerical Study

We give a numerical study to test the performance of the solutions $\boldsymbol{q}^{\mathsf{round}}$ and $\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor$ under the multinomial logit model. The solution $\boldsymbol{q}^{\mathsf{round}}$ has an optimality gap of $O(\log T \sqrt{T \log T})$. In Section 3, we work with the solution $\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil$ because this solution is simpler to analyze, but we can show that the solution $\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor$ has an optimality gap of $O(n + \sqrt{nT})$ as well. We use the solution $\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor$ in our numerical study, because Liang et al. (2022) use this solution. In all of our test problems, the unit revenue and procurements costs of all products are $p_i = 2$ and $c_i = 1$ for all $i \in \mathcal{N}$. The preference weights of all products are $w_i = 1$ for all $i \in \mathcal{N}$. The preference weight of the no-purchase option is $w_0 = 1$. We vary the number of products n over $\{2^k : k = 1, \ldots, 9\}$ and the number of time periods T over $\{10 \times 2^k : k = 1, \ldots, 9\}$. For each test problem, we compute the fluid upper bound V^{fluid} and

$T = 1000$

n	2	4	8	16	32	64	128	256	512
V^{fluid}	666.7	800.0	888.9	941.2	969.7	984.6	992.2	996.1	998.1
$\Pi(\boldsymbol{q}^{\mathsf{round}})$	650.3	781.3	870.1	923.0	953.6	970.3	979.4	984.7	987.8
$V^{\mathsf{fluid}} - \Pi(\boldsymbol{q}^{\mathsf{round}})$	16.43	18.69	18.75	18.20	16.12	14.32	12.81	11.40	10.23
$100 \times \frac{1}{V^{\mathsf{fluid}}}(V^{\mathsf{fluid}} - \Pi(\boldsymbol{q}^{\mathsf{round}}))$	2.46	2.34	2.11	1.93	1.66	1.45	1.29	1.14	1.03
$\Pi(\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor)$	650.3	781.6	870.1	923.6	955.1	960.0	896.0	768.0	512.0
$V^{\mathsf{fluid}} - \Pi(\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor)$	16.41	18.42	18.75	17.56	14.60	24.65	96.25	228.1	486.1
$100 \times \frac{1}{V^{\mathsf{fluid}}}(V^{\mathsf{fluid}} - \Pi(\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor))$	2.46	2.30	2.11	1.87	1.51	2.50	9.70	22.9	48.7

Table 1 Optimality gaps for the stocking quantities $\boldsymbol{q}^{\mathsf{round}}$ and $\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor$ for varying n.

$n = 100$

$T/10$	2	4	8	16	32	64	128	256	512
V^{fluid}	19.80	39.60	79.21	158.4	316.8	633.7	1267	2535	5069
$\Pi(\boldsymbol{q}^{\mathsf{round}})$	15.69	34.70	73.61	151.6	308.5	622.7	1252	2515	5042
$V^{\mathsf{fluid}} - \Pi(\boldsymbol{q}^{\mathsf{round}})$	4.11	4.91	5.60	6.85	8.31	10.98	14.83	19.66	27.12
$100 \times \frac{1}{V^{\mathsf{fluid}}}(V^{\mathsf{fluid}} - \Pi(\boldsymbol{q}^{\mathsf{round}}))$	20.76	12.39	7.06	4.33	2.62	1.73	1.17	0.78	0.53
$\Pi(\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor)$	0	0	0	100	300	600	1200	2499	5000
$V^{\mathsf{fluid}} - \Pi(\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor)$	19.80	39.60	79.21	58.42	16.84	33.66	67.33	34.66	69.31
$100 \times \frac{1}{V^{\mathsf{fluid}}}(V^{\mathsf{fluid}} - \Pi(\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor))$	100	100	100	36.87	5.31	5.31	5.31	1.37	1.37

Table 2 Optimality gaps for the stocking quantities $\boldsymbol{q}^{\mathsf{round}}$ and $\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor$ for varying T.

stocking quantities $\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor$ and $\boldsymbol{q}^{\mathsf{round}}$. We estimate the total expected profit obtained by these two sets of stocking quantities using Monte Carlo simulation with 10,000 sample paths.

In Table 1, we fix the number of time periods in the selling horizon at $T = 1000$ and vary the number of products. The first row gives the number of products. The second row gives the value of the fluid upper bound. The block of next three rows focus on the stocking quantities $\boldsymbol{q}^{\mathsf{round}}$. The first row gives the total expected profit obtained by using the stocking quantities $\boldsymbol{q}^{\mathsf{round}}$, the second row gives the absolute gap between the fluid upper bound and the total expected profit from the stocking quantities $\boldsymbol{q}^{\mathsf{round}}$ and the third row gives the same gap in relative percentage terms. The block of last three rows focus on the stocking quantities $\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor$ and provide the same performance measures. Consistent with Theorem 4.2, the absolute gap between the fluid upper bound and the total expected profit from the stocking quantities $\boldsymbol{q}^{\mathsf{round}}$ remains stable as the number of products increases. In contrast, consistent with Theorem 3.1, the absolute gap between the fluid upper bound and the total expected profit from the stocking quantities $\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor$ increases as the number of products increases. For $n = 256$ or 512, the stocking quantities $\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor$ perform rather poorly, with optimality gaps exceeding 20%, whereas the optimality gaps of the stocking quantities $\boldsymbol{q}^{\mathsf{round}}$ stay close to 1%. Other work, such as Liang et al. (2022), works with the stocking quantities $\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor$, but this solution is relevant when the number of time periods is large relative to number of products. When the number of products is large, the stocking quantities $\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor$ can be poor.

In Table 2, we fix the number of the number of products at $n = 100$ and and vary the number of time periods in the selling horizon. The layout of this table is identical to that of Table 1. By

Theorem 4.2, the optimality gap of the stocking quantities $\boldsymbol{q}^{\mathsf{round}}$ is $O(\log T\sqrt{T\log T})$. Accordingly, the absolute gap between the fluid upper bound and the total expected profit obtained by using the stocking quantities $\boldsymbol{q}^{\mathsf{round}}$ increases slightly with T, but noting that V^{fluid} increases linearly in T, the relative gap between the two quantities decrease. When $T = 20, 40, 80$ or 160 so that the number of products is large relative to the number of time periods, the stocking quantities $\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor$ are not practically useful as they have optimality gaps exceeding 30%. When $T = 2560$ or 5120, the optimality gaps of the stocking quantities $\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor$ are just over 1%, but the optimality gaps of the stocking quantities $\boldsymbol{q}^{\mathsf{round}}$ are below 1% for these numbers of time periods in the selling horizon. Overall, the superior theoretical performance guarantee that we can give for $\boldsymbol{q}^{\mathsf{round}}$ over $\lfloor \boldsymbol{q}^{\mathsf{fluid}} \rfloor$ carries over to the practical performance of these stocking quantities.

6. Conclusions

We analyzed fluid relaxations for the assortment and inventory planning problem under stockout-based substitution that has been of interest for a long time. Under a general choice model, we tightened the optimality gap of a solution from the fluid approximation by a factor of $O(\sqrt{\log(nT)})$ when compared with the earlier literature. More interestingly, under the multinomial logit model, we showed that we can obtain a solution from the fluid approximation with an optimality gap of $O(\log T\sqrt{T\log T})$, which is independent of the number of products. The result is significant for both theoretical and practical reasons. In particular, our solutions perform well under large demand volume even when the number of products is large, whereas the solutions in earlier results perform poorly when the number of products gets large. It would be interesting to obtain such tighter performance guarantees under other choice models.

References

Aouad, A., R. Levi, D. Segev. 2018. Greedy-like algorithms for dynamic assortment planning under multinomial logit preferences. *Operations Research* **66**(5) 1321–1345.

Aouad, A., R. Levi, D. Segev. 2019. Approximation algorithms for dynamic assortment optimization models. *Mathematics of Operations Research* **44**(2) 487–511.

Aouad, Ali, Danny Segev. 2022. The stability of MNL-based demand under dynamic customer substitution and its algorithmic implications. *Operations Research* .

Aven, T. 1985. Upper (lower) bounds on the mean of the maximum (minimum) of a number of random variables. *Journal of Applied Probability* **22**(3) 723–728.

Blanchet, J., G. Gallego, V. Goyal. 2016. A Markov chain approximation to choice modeling. *Operations Research* **64**(4) 886–905.

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

Davis, J.M., G. Gallego, H. Topaloglu. 2014. Assortment optimization under variants of the nested logit model. *Operations Research* **62**(2) 250–273.

El Housni, O., O. Mouchtaki, G. Gallego, V. Goyal, S. Humair, S. Kim, A. Sadighian, J. Wu. 2021. Joint assortment and inventory planning for heavy tailed demand. Tech. rep., Columbia Business School, New York, NY.

Gallego, G., I. K. Moon. 1993. The distribution-free newsboy problem: Review and extensions. *Journal of the Operational Research Society* **44**(8) 825–834.

Gallego, G., R. Ratliff, S. Shebalov. 2015. A general attraction model and sales-based linear programming formulation for network revenue management under customer choice. *Operations Research* **63**(1) 212–232.

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

Golrezaei, N., H. Nazerzadeh, P. Rusmevichientong. 2014. Real-time optimization of personalized assortments. *Management Science* **60**(6) 1532–1551.

Goyal, V., R. Levi, D. Segev. 2016. Near-optimal algorithms for the assortment planning problem under dynamic substitution and stochastic demand. *Operations Research* **64**(1) 219–235.

Honhon, D., V. Gaur, S. Seshadri. 2010. Assortment planning and inventory decisions under stockout-based substitution. *Operations Research* **58**(5) 1364–1379.

Honhon, D., S. Seshadri. 2013. Fixed vs. random proportions demand models for the assortment planning problem under stockout-based substitution. *Manufacturing & Service Operations Management* **15**(3) 378–386.

Hopp, W. J., X. Xu. 2008. A static approximation for dynamic demand substitution with applications in a competitive market. *Operations Research* **56**(3) 630–645.

Li, G., P. Rusmevichientong, H. Topaloglu. 2015. The d-level nested logit model: Assortment and price optimization problems. *Operations Research* **62**(2) 325–342.

Liang, A. J., S. Jasin, J. Uichanco. 2022. Assortment and inventory planning under dynamic substitution with MNL model: An LP approach and an asymptotically optimal policy. Tech. rep., University of Michigan, Ann Arbor, Michigan.

Ma, W., D. Simchi-Levi, J. Zhao. 2021. Dynamic pricing (and assortment) under a static calendar. *Management Science* **67**(4) 2292–2313.

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

Nagaev, S. V. 1965. Some limit theorems for large deviations. *Theory of Probability & Its Applications* **10**(2) 214–235.

Rusmevichientong, P., M. Sumida, H. Topaloglu. 2020. Dynamic assortment optimization for reusable products with random usage durations. *Management Science* **66**(7) 2820–2844.

Talluri, K., G. van Ryzin. 2004. Revenue management under a general discrete choice model of consumer behavior. *Management Science* **50**(1) 15–33.

Topaloglu, H. 2013. Joint stocking and product offer decisions under the multinomial logit model. *Production and Operations Management* **22**(5) 1182–1199.

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

Online Supplement
Assortment and Inventory Planning under Stockout-Based Substitution: The Many-Products Regime

Appendix A: Proofs of Results in Section 3

Proof of Theorem 3.1:

Recall $\boldsymbol{C}_t(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil)$ denotes the choice of the customer arriving at time period t among the set of products $A(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil)$ and a customer chooses product i among this set of products with probability $\phi_i(A(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil))$. By (2) and (3), we have

$$
\begin{aligned}
V^{\mathsf{fluid}} - \Pi^{\mathsf{static}}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil) &= \sum_{i\in\mathcal{N}} (p_i - c_i) q_i^{\mathsf{fluid}} - \sum_{i\in\mathcal{N}} p_i \mathbb{E}\left\{ \min\left\{ \lceil q_i^{\mathsf{fluid}} \rceil, \sum_{t\in\mathcal{T}} C_{it}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil) \right\} \right\} + \sum_{i\in\mathcal{N}} c_i \lceil q_i^{\mathsf{fluid}} \rceil \\
&\le \sum_{i\in S^{\mathsf{fluid}}} p_i \mathbb{E}\left\{ q_i^{\mathsf{fluid}} - \min\left\{ q_i^{\mathsf{fluid}}, \sum_{t\in\mathcal{T}} C_{it}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil) \right\} \right\} + \sum_{i\in S^{\mathsf{fluid}}} c_i \\
&= \sum_{i\in S^{\mathsf{fluid}}} p_i \mathbb{E}\left\{ \left[q_i^{\mathsf{fluid}} - \sum_{t\in\mathcal{T}} C_{it}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil) \right]^+ \right\} + \sum_{i\in S^{\mathsf{fluid}}} c_i \le \sum_{i\in S^{\mathsf{fluid}}} p_i \mathbb{E}\left\{ \left| q_i^{\mathsf{fluid}} - \sum_{t\in\mathcal{T}} C_{it}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil) \right| \right\} + \sum_{i\in S^{\mathsf{fluid}}} c_i \\
&\overset{(a)}{\le} \sum_{i\in S^{\mathsf{fluid}}} p_i \sqrt{ \mathbb{E}\left\{ \left(q_i^{\mathsf{fluid}} - \sum_{t\in\mathcal{T}} C_{it}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil) \right)^2 \right\} } + \sum_{i\in S^{\mathsf{fluid}}} c_i \overset{(b)}{=} \sum_{i\in S^{\mathsf{fluid}}} p_i \sqrt{ \mathbb{V}\mathrm{ar}\left\{ \sum_{t\in\mathcal{T}} C_{it}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil) \right\} } + \sum_{i\in S^{\mathsf{fluid}}} c_i \\
&= \sum_{i\in S^{\mathsf{fluid}}} p_i \sqrt{T\phi_i(S^{\mathsf{fluid}})(1-\phi_i(S^{\mathsf{fluid}}))} + \sum_{i\in S^{\mathsf{fluid}}} c_i \le \sum_{i\in S^{\mathsf{fluid}}} p_i \sqrt{T\phi_i(S^{\mathsf{fluid}})} + \sum_{i\in S^{\mathsf{fluid}}} c_i \\
&\overset{(c)}{\le} \bar{p}\sqrt{|S^{\mathsf{fluid}}|\,T} + \bar{c}\left| S^{\mathsf{fluid}} \right|,
\end{aligned}
$$

where (a) holds by Jensen's inequality, (b) holds because $\mathbb{E}\left\{\sum_{t\in\mathcal{T}} C_{it}(\lceil \boldsymbol{q}^{\mathsf{fluid}} \rceil)\right\} = q_i^{\mathsf{fluid}}$ and (c) follows from Cauchy-Schwarz inequality and $\sum_{i\in S^{\mathsf{fluid}}} \phi_i(S^{\mathsf{fluid}}) \le 1$. ■

Lemma A.1 *Suppose $k\,p$ is integral, it holds that*

$$
\mathbb{E}\{[k\,p - \mathsf{binomial}(k,p)]^+\} \ge \frac{1}{\sqrt{2\pi}}\sqrt{k\,p\,(1-p)} - O(1).
$$

Proof: Let $q = 1-p$ and $(X_i)_{i=1}^k$ denote k independent and identical distributed binary random variables such that $X_i = \mathsf{Bernoulli}(p) - p$, then we have

$$
\begin{aligned}
&\mathbb{E}\left\{[k\,p - \mathsf{binomial}(k,p)]^+\right\} = \frac{1}{2}\mathbb{E}\left\{|k\,p - \mathsf{binomial}(k,p)| + k\,p - \mathsf{binomial}(k,p)\right] \\
&= \frac{1}{2}\mathbb{E}\left\{|k\,p - \mathsf{binomial}(k,p)|\right\} = \mathbb{E}\left\{[\mathsf{binomial}(k,p) - k\,p]^+\right\} = \sqrt{k\,p\,q}\,\mathbb{E}\left\{ \left[\frac{\sum_{i=1}^k X_i}{\sqrt{k\,p\,q}} \right]^+ \right\}
\end{aligned}
$$

$$= \sqrt{k\,p\,q}\int_0^\infty \mathbb{P}\left(\frac{\sum_{i=1}^k X_i}{\sqrt{k\,p\,q}} > x\right)dx \geq \sqrt{k\,p\,q}\int_0^\infty \left(\mathbb{P}\left(\mathsf{Norm}(0,1) > x\right) - \frac{C}{(1+x^3)\sqrt{k\,p\,q}}\right)dx$$
$$= \sqrt{\frac{k\,p\,q}{2\pi}} - C\int_0^\infty \frac{1}{1+x^3}dx,$$

where C is a constant and the last inequality follows from the nonuniform Berry-Esseen inequality (Nagaev 1965). Note that

$$\int_0^\infty \frac{1}{1+x^3}dx \leq 1 + \int_1^\infty \frac{1}{1+x^3}dx \leq 1 + \int_1^\infty \frac{1}{x^3}dx \leq \frac{3}{2},$$

therefore,

$$\mathbb{E}\{[k\,p - \mathsf{binomial}(k,p)]^+\} \geq \sqrt{\frac{k\,p\,q}{2\pi}} - \frac{3}{2}C.$$

■

Appendix B: Proofs of Results in Section 4

Proof of Lemma 4.1:

We first note that $\delta = \lceil \sum_{i\in S^{\mathsf{fluid}}}(q_i^{\mathsf{fluid}} - \lfloor q_i^{\mathsf{fluid}} \rfloor)\rceil \leq |S^{\mathsf{fluid}}|$, thus,

$$\sum_{i\in S^{\mathsf{fluid}}}(q_i^{\mathsf{round}} - q_i^{\mathsf{fluid}}) = \sum_{i\in S^{\mathsf{fluid}}}(\lfloor q_i^{\mathsf{fluid}}\rfloor - q_i^{\mathsf{fluid}}) + \delta = \left\lceil \sum_{i\in S^{\mathsf{fluid}}}(q_i^{\mathsf{fluid}} - \lfloor q_i^{\mathsf{fluid}}\rfloor)\right\rceil - \sum_{i\in S^{\mathsf{fluid}}}(q_i^{\mathsf{fluid}} - \lfloor q_i^{\mathsf{fluid}}\rfloor) \in [0,1].$$

Next, we have

$$\sum_{i\in S^{\mathsf{fluid}}}(p_i - c_i)(q_i^{\mathsf{fluid}} - q_i^{\mathsf{round}}) = \sum_{i\in S^{\mathsf{fluid}}}(p_i - c_i)q_i^{\mathsf{fluid}} - \sum_{i=1}^{\delta}(p_i - c_i)(\lfloor q_i^{\mathsf{fluid}}\rfloor + 1) - \sum_{i=\delta+1}^{|S^{\mathsf{fluid}}|}(p_i - c_i)\lfloor q_i^{\mathsf{fluid}}\rfloor$$
$$\leq \sum_{i\in S^{\mathsf{fluid}}}(p_i - c_i)q_i^{\mathsf{fluid}} - \sum_{i=1}^{\delta}(p_i - c_i)q_i^{\mathsf{fluid}} - (p_\delta - c_\delta)\left(\delta + \sum_{i=1}^{\delta}(\lfloor q_i^{\mathsf{fluid}}\rfloor - q_i^{\mathsf{fluid}})\right) - \sum_{i=\delta+1}^{|S^{\mathsf{fluid}}|}(p_i - c_i)\lfloor q_i^{\mathsf{fluid}}\rfloor$$
$$= \sum_{i=\delta+1}^{|S^{\mathsf{fluid}}|}(p_i - c_i)(q_i^{\mathsf{fluid}} - \lfloor q_i^{\mathsf{fluid}}\rfloor) - (p_\delta - c_\delta)\left(\delta - \sum_{i=1}^{\delta}(q_i^{\mathsf{fluid}} - \lfloor q_i^{\mathsf{fluid}}\rfloor)\right)$$
$$\leq (p_\delta - c_\delta)\left(\sum_{i\in S^{\mathsf{fluid}}}(q_i^{\mathsf{fluid}} - \lfloor q_i^{\mathsf{fluid}}\rfloor) - \delta\right) \leq 0.$$

Finally, we have

$$\left|\left\{i : q_i^{\mathsf{round}} > 0\right\}\right| \leq \sum_{i\in S^{\mathsf{fluid}}} q_i^{\mathsf{round}} \leq \sum_{i\in S^{\mathsf{fluid}}} q_i^{\mathsf{fluid}} + 1 \leq T + 1.$$

■

We use a sequence of preliminary lemmas to give a proof for Theorem 4.2.

Lemma B.1 *Given any nonzero inventory $\boldsymbol{q} \in \mathbb{Z}_+^N$ with T customers, it holds that*

$$\mathbb{E}\left\{X_0(\boldsymbol{q},T)\right\} \leq 1 + w_0 \max_{i\in\mathcal{N}}\left\{\frac{q_i}{w_i}\right\}\log\left(\sum_{i\in\mathcal{N}} q_i\right) + \left[T - \sum_{i\in\mathcal{N}} q_i\right]^+.$$

Proof: Note that for each customer t, the no-purchase random variable follows the Bernoulli distribution with success probability $1-\sum_{i\in\mathcal{N}}\phi_i(S_t)$ where S_t is the set of available items observed by the customer and $\phi_i(S_t)$ is the purchase probability of item i. Since each customer can buy at most one item, it is sufficient to consider the problem where the decision maker removes one unit of product per period in order to maximize the total expected no-purchase probability $V_1^0(\boldsymbol{q})$, which can be recursively computed as

$$V_t^0(\boldsymbol{q}')=\frac{w_0}{w_0+\sum\limits_{i:q_i'>0} w_i}+\max_{i:q_i'>0} V_{t+1}^0(\boldsymbol{q}'-\boldsymbol{e}_i),$$

where $V_{T+1}^0(\boldsymbol{q}')=0$ for any $\boldsymbol{q}'$. To bound $V_1^0(\boldsymbol{q})$, we consider an auxiliary problem given by

$$\tilde{V}_t^0(\boldsymbol{q}')=\frac{w_0}{w_0+\sum\limits_{i:q'>0} w_i q_i'/q_i}+\max_{i:q_i'>0} \tilde{V}_{t+1}^0(\boldsymbol{q}'-\boldsymbol{e}_i). \tag{4}$$

We claim that for any time period t and inventory $\boldsymbol{q}'\le\boldsymbol{q}$, $V_t^0(\boldsymbol{q}')\le\tilde{V}_t^0(\boldsymbol{q}')$. We show the claim by induction. It is clear that the claim holds for time period $T+1$. Now assume the claim holds for time period $t+1$, then for time period t, we have

$$\begin{aligned}V_t^0(\boldsymbol{q}')&=\frac{w_0}{w_0+\sum\limits_{i:q_i'>0} w_i}+\max_{i:q_i'>0} V_{t+1}^0(\boldsymbol{q}'-\boldsymbol{e}_i)\le\frac{w_0}{w_0+\sum\limits_{i:q_i'>0} w_i}+\max_{i:q_i'>0} \tilde{V}_{t+1}^0(\boldsymbol{q}'-\boldsymbol{e}_i)\\&\le\frac{w_0}{w_0+\sum\limits_{i:q_i'>0} w_i q_i'/q_i}+\max_{i:q_i'>0} \tilde{V}_{t+1}^0(\boldsymbol{q}'-\boldsymbol{e}_i)=\tilde{V}_t^0(\boldsymbol{q}').\end{aligned}$$

Therefore, the claim holds by induction, which implies that $V_1^0(\boldsymbol{q})\le\tilde{V}_1^0(\boldsymbol{q})$.

The auxiliary problem is equivalent to the problem that splits each item with q_i inventories to q_i items with preference weight w_i/q_i. Without loss of generality, we assume the $\sum_{i\in\mathcal{N}} q_i$ items follow the decreasing order of w_k', where w_k' is the weight of the k-th item in the auxiliary problem. Thus, an optimal policy to the dynamic program in (4) selects an item with largest index. So, we have

$$\begin{aligned}\tilde{V}_1^0(\boldsymbol{q})&\le\sum_{k=1}^{\sum\limits_{i\in\mathcal{N}} q_i}\frac{w_0}{w_0+\sum\limits_{j=k}^{\sum_{i\in\mathcal{N}} q_i} w_k'}+\left[T-\sum_{i\in\mathcal{N}} q_i\right]^+\\&\le\left(\frac{w_0/\min_k\{w_k'\}}{w_0/\min_k\{w_k'\}+1}+\cdots+\frac{w_0/\min_k\{w_k'\}}{w_0/\min_k\{w_k'\}+\sum\limits_{i\in\mathcal{N}} q_i}\right)+\left[T-\sum_{i\in\mathcal{N}} q_i\right]^+\\&\le 1+\frac{w_0}{\min_k\{w_k'\}}\left(\log\left(\frac{w_0}{\min_k\{w_k'\}}+\sum_{i\in\mathcal{N}} q_i\right)-\log\left(\frac{w_0}{\min_k\{w_k'\}}+1\right)\right)+\left[T-\sum_{i\in\mathcal{N}} q_i\right]^+\\&\le 1+w_0\max_{i\in\mathcal{N}}\left\{\frac{q_i}{w_i}\right\}\log\left(\sum_{i\in\mathcal{N}} q_i\right)+\left[T-\sum_{i\in\mathcal{N}} q_i\right]^+.\end{aligned}$$

The result then follows from $\mathbb{E}\{X_0(\boldsymbol{q},T)\} \le \tilde{V}_1^0(\boldsymbol{q})$. ■

In what follows, we let $\tilde{w} = w_0 + \sum_{i\in S^{\mathsf{fluid}}} w_i$ for simplicity.

Lemma B.2 *Suppose* $\tilde{T} = \left\lfloor T - \frac{\tilde{w}}{\underline{w}} - \sqrt{2\frac{\tilde{w}}{\underline{w}} T \log\left(|S^{\mathsf{fluid}}|T\right)}\right\rfloor \ge 0$, *then it holds that*

$$\mathbb{P}\left(X_i(\boldsymbol{q}^{\mathsf{round}}, \tilde{T}) < q_i^{\mathsf{round}}, \forall i \in S^{\mathsf{fluid}}\right) \ge 1 - \frac{1}{T}.$$

Proof: Since $T \ge \tilde{w}/\underline{w}$, we have $q_i^{\mathsf{fluid}} \ge 1$ for any $i \in S^{\mathsf{fluid}}$ and thus $A(\boldsymbol{q}^{\mathsf{round}}) = S^{\mathsf{fluid}}$. Note that when all products are available, the item purchased by a customer within $\tilde{T}$ periods follows a multinomial distribution $(X_0, X_1, \dots, X_{|S^{\mathsf{fluid}}|}) \sim (\tilde{T}, w_0/\tilde{w}, \dots, w_{|S^{\mathsf{fluid}}|}/\tilde{w})$. Therefore, by union bound,

$$\begin{aligned}
&\mathbb{P}\left(X_i(\boldsymbol{q}^{\mathsf{round}}, \tilde{T}) < q_i^{\mathsf{round}}, \forall i \in S^{\mathsf{fluid}}\right) = \mathbb{P}\left(X_i < q_i^{\mathsf{round}}, \forall i \in S^{\mathsf{fluid}}\right) = 1 - \mathbb{P}\left(\exists i \in S^{\mathsf{fluid}}, X_i \ge q_i^{\mathsf{round}}\right) \\
&\quad \ge 1 - \sum_{i\in S^{\mathsf{fluid}}} \mathbb{P}\left(X_i \ge q_i^{\mathsf{round}}\right).
\end{aligned}$$

For any $i \in S^{\mathsf{fluid}}$, since X_i is a binomial random variable, we have

$$\begin{aligned}
&\mathbb{P}\left(X_i \ge q_i^{\mathsf{round}}\right) \le \mathbb{P}\left(X_i \ge \frac{w_i}{\tilde{w}} T - 1\right) = \mathbb{P}\left(X_i \ge \frac{w_i}{\tilde{w}} \tilde{T} + \frac{w_i}{\tilde{w}}(T - \tilde{T}) - 1\right) \\
&\quad = \mathbb{P}\left(X_i \ge \frac{w_i}{\tilde{w}} \tilde{T}\left(1 + \frac{T-\tilde{T}}{\tilde{T}} - \frac{\tilde{w}}{w_i \tilde{T}}\right)\right) \overset{(a)}{\le} \exp\left(-\frac{\frac{w_i}{\tilde{w}}\tilde{T}\left(\frac{T-\tilde{T}}{\tilde{T}} - \frac{\tilde{w}}{w_i\tilde{T}}\right)^2}{1 + \frac{T}{\tilde{T}} - \frac{\tilde{w}}{w_i \tilde{T}}}\right) \\
&\quad = \exp\left(-\frac{\frac{w_i}{\tilde{w}}\left(T - \tilde{T} - \frac{\tilde{w}}{w_i}\right)^2}{\tilde{T} + T - \frac{\tilde{w}}{w_i}}\right) \le \exp\left(-\frac{\left(T - \tilde{T} - \frac{\tilde{w}}{w_i}\right)^2}{2\frac{\tilde{w}}{w_i}T}\right) \le \exp\left(-\frac{\left(T - \tilde{T} - \frac{\tilde{w}}{\underline{w}}\right)^2}{2\frac{\tilde{w}}{\underline{w}}T}\right) \le \frac{1}{|S^{\mathsf{fluid}}|T},
\end{aligned}$$

where (a) follows from Chernoff bound. ■

Lemma B.3 *Given no item is sold out until time period* $\tilde{T}$, *the expected number of no-purchases until time period* $\tilde{T}$ *is bounded by*

$$\mathbb{E}\left\{X_0(\boldsymbol{q}^{\mathsf{round}}, \tilde{T}) \middle| X_i(\boldsymbol{q}^{\mathsf{round}}, \tilde{T}) < q_i^{\mathsf{round}}, \forall i \in S^{\mathsf{fluid}}\right\} \le \frac{w_0}{\tilde{w}} T.$$

Proof: Using Lemma B.2, we have

$$\begin{aligned}
&\mathbb{E}\left\{X_0(\boldsymbol{q}^{\mathsf{round}}, \tilde{T}) \middle| X_i(\boldsymbol{q}^{\mathsf{round}}, \tilde{T}) < q_i^{\mathsf{round}}, \forall i \in S^{\mathsf{fluid}}\right\} \\
&\quad = \mathbb{E}\left\{\tilde{T} - \sum_{i\in\mathcal{N}} X_i(\boldsymbol{q}^{\mathsf{round}}, \tilde{T}) \middle| X_i(\boldsymbol{q}^{\mathsf{round}}, \tilde{T}) < q_i^{\mathsf{round}}, \forall i \in S^{\mathsf{fluid}}\right\} \\
&\quad = \mathbb{E}\left\{\tilde{T} - \sum_{i\in\mathcal{N}} X_i \middle| X_i < q_i^{\mathsf{round}}, \forall i \in S^{\mathsf{fluid}}\right\} = \mathbb{E}\left\{X_0 \middle| X_i < q_i^{\mathsf{round}}, \forall i \in S^{\mathsf{fluid}}\right\} \\
&\quad \le \frac{\mathbb{E}\{X_0\}}{\mathbb{P}(X_i < q_i^{\mathsf{round}}, \forall i \in S^{\mathsf{fluid}})} \le \frac{T}{T-1}\frac{w_0}{\tilde{w}}\tilde{T} \le \frac{w_0}{\tilde{w}} T.
\end{aligned}$$

■

Lemma B.4 *Given no item is sold out until time period $\tilde{T}$, there exist constants C_1 and C_2 such that the expected number of no-purchase after time period $\tilde{T}$ is bounded by*

$$\mathbb{E}\left\{X_0(\boldsymbol{q}^{\mathsf{round}},T)-X_0(\boldsymbol{q}^{\mathsf{round}},\tilde{T})\middle|X_i(\boldsymbol{q}^{\mathsf{round}},\tilde{T})<q_i^{\mathsf{round}},\forall i\in S^{\mathsf{fluid}}\right\}\le C_1\log T\sqrt{T\log T}+C_2.$$

Proof: Conditional on that no item is sold out until time period $\tilde{T}$, then the inventory left after time period $\tilde{T}$ is $\tilde{\boldsymbol{q}}=\boldsymbol{q}^{\mathsf{round}}-\boldsymbol{X}$, where $\boldsymbol{X}$ is the multinomial random variable defined in the proof of Lemma B.2. By Lemma B.1, we have

$$\begin{aligned}
&\mathbb{E}\left\{X_0(\boldsymbol{q}^{\mathsf{round}},T)-X_0(\boldsymbol{q}^{\mathsf{round}},\tilde{T})\middle|X_i(\boldsymbol{q}^{\mathsf{round}},\tilde{T})<q_i^{\mathsf{round}},\forall i\in S^{\mathsf{fluid}}\right]\\
&\le\mathbb{E}\left\{\underbrace{1+w_0\max_{i\in S^{\mathsf{fluid}}}\left\{\frac{\tilde{q}_i}{w_i}\right\}\log\left(\sum_{i\in S^{\mathsf{fluid}}}\tilde{q}_i\right)}_{(a)}+\underbrace{\left[T-\tilde{T}-\sum_{i\in S^{\mathsf{fluid}}}\tilde{q}_i\right]^+}_{(b)}\middle|X_i(\boldsymbol{q}^{\mathsf{round}},\tilde{T})<q_i^{\mathsf{round}},\forall i\in S^{\mathsf{fluid}}\right\}.
\end{aligned}$$

We focus on the second term (b) first. Let $\kappa=w_0/\underline{w}$,

$$\begin{aligned}
&\mathbb{E}\left\{\left[T-\tilde{T}-\sum_{i\in S^{\mathsf{fluid}}}\tilde{q}_i\right]^+\middle|X_i(\boldsymbol{q}^{\mathsf{round}},\tilde{T})<q_i^{\mathsf{round}},\forall i\in S^{\mathsf{fluid}}\right\}\\
&=\mathbb{E}\left\{\left[T-\sum_{i\in S^{\mathsf{fluid}}}q_i^{\mathsf{round}}-\tilde{T}+\sum_{i\in\in S^{\mathsf{fluid}}}X_i\right]^+\middle|X_i<q_i^{\mathsf{round}},\forall i\in S^{\mathsf{fluid}}\right\}\\
&\overset{(c)}{\le}\mathbb{E}\left\{\left[T-\sum_{i\in S^{\mathsf{fluid}}}q_i^{\mathsf{fluid}}-\tilde{T}+\sum_{i\in S^{\mathsf{fluid}}}X_i\right]^+\middle|X_i<q_i^{\mathsf{round}},\forall i\in S^{\mathsf{fluid}}\right\}\\
&=\mathbb{E}\left\{\left[\frac{w_0}{\tilde{w}}T-\tilde{T}+\sum_{i\in S^{\mathsf{fluid}}}X_i\right]^+\middle|X_i<q_i^{\mathsf{round}},\forall i\in S^{\mathsf{fluid}}\right\}\le\mathbb{E}\left\{\left[\frac{w_0}{\tilde{w}}T-\tilde{T}+\sum_{i\in S^{\mathsf{fluid}}}X_i\right]^+\right\}\\
&\le\sqrt{\mathbb{E}\left\{\left(\sum_{i\in S^{\mathsf{fluid}}}X_i-\tilde{T}+\frac{w_0}{\tilde{w}}T\right)^2\right\}}\overset{(d)}{=}\sqrt{\mathbb{E}\left\{\left(\frac{w_0}{\tilde{w}}T-X_0\right)^2\right\}}\\
&=\sqrt{\frac{w_0(\tilde{w}-w_0)}{\tilde{w}^2}\tilde{T}+\left(\frac{w_0}{\tilde{w}}(T-\tilde{T})\right)^2}\le\sqrt{\left(\frac{w_0}{\tilde{w}}\right)^2\left(\frac{\tilde{w}}{\underline{w}}+1+\sqrt{2\frac{\tilde{w}}{\underline{w}}T\log(|S^{\mathsf{fluid}}|T)}\right)^2+\frac{w_0}{\tilde{w}}\tilde{T}}\\
&=\sqrt{\left(\frac{w_0}{\underline{w}}\right)^2+2\frac{w_0^2}{\tilde{w}\underline{w}}\left(1+T\log(T|S^{\mathsf{fluid}}|)\right)+\left(\frac{w_0}{\tilde{w}}\right)^2\left(1+2\left(1+\frac{\tilde{w}}{\underline{w}}\right)\sqrt{2\frac{\tilde{w}}{\underline{w}}T\log(|S^{\mathsf{fluid}}|T)}\right)+\frac{w_0}{\tilde{w}}\tilde{T}}\\
&\le\sqrt{\kappa^2+\frac{2\kappa^2}{|S^{\mathsf{fluid}}|}\left(1+T\log(|S^{\mathsf{fluid}}|T)\right)+\frac{\kappa^2}{|S^{\mathsf{fluid}}|^2}+2\left(\sqrt{\frac{\kappa^4}{|S^{\mathsf{fluid}}|^3}}+\sqrt{\frac{\kappa^4}{|S^{\mathsf{fluid}}|}}\right)\sqrt{2T\log(|S^{\mathsf{fluid}}|T)}+\frac{\kappa}{|S^{\mathsf{fluid}}|}\tilde{T}}\\
&\le\sqrt{4\kappa^2+2\kappa^2\frac{T}{|S^{\mathsf{fluid}}|}\log(|S^{\mathsf{fluid}}|T)+4\kappa^2\sqrt{2\frac{T}{|S^{\mathsf{fluid}}|}\log(|S^{\mathsf{fluid}}|T)}+\kappa\frac{T}{|S^{\mathsf{fluid}}|}},
\end{aligned}$$

where (c) follows from Lemma 4.1 and (d) holds because $\boldsymbol{X}$ is a multinomial random variable with $\tilde{T}$ trials. Therefore, there exists constants D_1 and D_2 such that

$$\mathbb{E}\left\{\left[T-\tilde{T}-\sum_{i\in S^{\mathsf{fluid}}}\tilde{q}_i\right]\middle| X_i(\boldsymbol{q}^{\mathsf{round}},\tilde{T})<q_i^{\mathsf{round}},\forall i\in S^{\mathsf{fluid}}\right\}\le D_1\sqrt{T\log(T|S^{\mathsf{fluid}}|)}+D_2.$$

Now we analyze the first item (a). We have

$$\begin{aligned}
&\mathbb{E}\left\{w_0\max_{i\in S^{\mathsf{fluid}}}\left\{\frac{\tilde{q}_i}{w_i}\right\}\log\left(\sum_{i\in S^{\mathsf{fluid}}}\tilde{q}_i\right)\middle| X_i(\boldsymbol{q}^{\mathsf{round}},\tilde{T})<q_i^{\mathsf{round}},\forall i\in S^{\mathsf{fluid}}\right\}\\
&=\mathbb{E}\left\{w_0\max_{i\in S^{\mathsf{fluid}}}\left\{\frac{q_i^{\mathsf{round}}-X_i}{w_i}\right\}\log\left(\sum_{i\in S^{\mathsf{fluid}}}\left(q_i^{\mathsf{round}}-X_i\right)\right)\middle| X_i<q_i^{\mathsf{round}},\forall i\in S^{\mathsf{fluid}}\right\}\\
&\overset{(e)}{\le}\mathbb{E}\left\{w_0\max_{i\in S^{\mathsf{fluid}}}\left\{\frac{q_i^{\mathsf{fluid}}-X_i+1}{w_i}\right\}\log\left(\sum_{i\in S^{\mathsf{fluid}}}q_i^{\mathsf{fluid}}+1\right)\middle| X_i<q_i^{\mathsf{round}},\forall i\in S^{\mathsf{fluid}}\right\}\\
&\le\mathbb{E}\left\{w_0\max_{i\in S^{\mathsf{fluid}}}\left\{\frac{q_i^{\mathsf{fluid}}-X_i+1}{w_i}\right\}\log(T+1)\middle| X_i<q_i^{\mathsf{round}},\forall i\in S^{\mathsf{fluid}}\right\}\\
&\le\left(\mathbb{E}\left\{\max_{i\in S^{\mathsf{fluid}}}\left\{\frac{q_i^{\mathsf{fluid}}-X_i}{w_i}\right\}\middle| X_i<q_i^{\mathsf{round}},\forall i\in S^{\mathsf{fluid}}\right\}+\frac{1}{\underline{w}}\right)w_0\log(T+1),
\end{aligned}$$

where inequality (e) follows from Lemma 4.1. We are left to bound the first term in the inequality above. Note that

$$\begin{aligned}
&w_0\mathbb{E}\left\{\max_{i\in S^{\mathsf{fluid}}}\left\{\frac{q_i^{\mathsf{fluid}}-X_i}{w_i}\right\}\middle| X_i<q_i^{\mathsf{round}},\forall i\in S^{\mathsf{fluid}}\right\}\\
&\le\frac{w_0}{\mathbb{P}\left(X_i<q_i^{\mathsf{round}},\forall i\in S^{\mathsf{fluid}}\right)}\mathbb{E}\left\{\max_{i\in S^{\mathsf{fluid}}}\left\{\left[\frac{q_i^{\mathsf{fluid}}-X_i}{w_i}\right]^+\right\}\right\}\\
&\overset{(f)}{\le}\frac{T}{T-1}\left(\max_{i\in S^{\mathsf{fluid}}}\left\{\mathbb{E}\left\{\frac{w_0}{w_i}\left[q_i^{\mathsf{fluid}}-X_i\right]^+\right\}\right\}+\sqrt{\sum_{i\in S^{\mathsf{fluid}}}\mathbb{V}\mathrm{ar}\left\{\frac{w_0}{w_i}\left[q_i^{\mathsf{fluid}}-X_i\right]^+\right\}}\right)\\
&\le\frac{T}{T-1}\left(\max_{i\in S^{\mathsf{fluid}}}\left\{\sqrt{\mathbb{E}\left\{\left(\frac{w_0}{w_i}\left(q_i^{\mathsf{fluid}}-X_i\right)\right)^2\right\}}\right\}+\sqrt{\sum_{i\in S^{\mathsf{fluid}}}\mathbb{E}\left\{\left(\frac{w_0}{w_i}\left(q_i^{\mathsf{fluid}}-X_i\right)\right)^2\right\}}\right),
\end{aligned}$$

where (f) follows from Theorem 2.1 in Aven (1985). Finally, we have

$$\begin{aligned}
&\mathbb{E}\left\{\left(\frac{w_0}{w_i}\left(q_i^{\mathsf{fluid}}-X_i\right)\right)^2\right\}=\mathbb{V}\mathrm{ar}\left\{\frac{w_0}{w_i}X_i\right\}+\left(\frac{w_0}{w_i}q_i^{\mathsf{fluid}}-\mathbb{E}\left\{\frac{w_0}{w_i}X_i\right\}\right)^2\\
&=\frac{w_0^2\left(\tilde{w}-w_i\right)}{w_i\tilde{w}^2}\tilde{T}+\left(\frac{w_0}{\tilde{w}}T-\frac{w_0}{\tilde{w}}\tilde{T}\right)^2\le\frac{w_0^2\tilde{T}}{w_i\tilde{w}}+\frac{w_0^2}{\tilde{w}^2}\left(\frac{\tilde{w}}{\underline{w}}+1+\sqrt{2\frac{\tilde{w}}{\underline{w}}T\log(|S^{\mathsf{fluid}}|T)}\right)^2\\
&\le\frac{w_0^2}{\underline{w}\tilde{w}}\tilde{T}+\left(\frac{w_0}{\underline{w}}+\frac{w_0}{\tilde{w}}\right)^2+2\frac{w_0^2}{\tilde{w}^2}\left(\frac{\tilde{w}}{\underline{w}}+1\right)\sqrt{2\frac{\tilde{w}}{\underline{w}}T\log\left(|S^{\mathsf{fluid}}|T\right)}+2\frac{w_0^2}{\tilde{w}\underline{w}}T\log(|S^{\mathsf{fluid}}|T)
\end{aligned}$$

$$\leq \frac{\kappa^2}{|S^{\text{fluid}}|}\tilde{T} + 4\kappa^2 + 2\kappa^2\sqrt{2\frac{T}{|S^{\text{fluid}}|}\log(|S^{\text{fluid}}|T)} + 2\kappa^2\sqrt{\frac{T}{|S^{\text{fluid}}|^3}\log(|S^{\text{fluid}}|T)} + 2\kappa^2\frac{T}{|S^{\text{fluid}}|}\log(|S^{\text{fluid}}|T)$$
$$\leq 4\kappa^2 + 3\kappa^2\frac{T}{|S^{\text{fluid}}|}\log(|S^{\text{fluid}}|T) + 6\kappa^2\sqrt{\frac{T}{|S^{\text{fluid}}|}\log(|S^{\text{fluid}}|T)}.$$

Combining the preceding inequalities, term (a) can be bounded by

$$\mathbb{E}\left\{w_0 \max_{i\in S^{\text{fluid}}}\left\{\frac{\tilde{q}_i}{w_i}\right\}\log\left(\sum_{i\in S^{\text{fluid}}}\tilde{q}_i\right)\middle| X_i(\boldsymbol{q}^{\text{round}},\tilde{T}) < q_i^{\text{round}}, \forall i\in S^{\text{fluid}}\right\}$$
$$\leq \kappa\left(1 + 2\left(1+\sqrt{|S^{\text{fluid}}|}\right)\left(\sqrt{4 + 3\frac{T}{|S^{\text{fluid}}|}\log(|S^{\text{fluid}}|T) + 6\sqrt{\frac{T}{|S^{\text{fluid}}|}\log(|S^{\text{fluid}}|T)}}\right)\right)\log(T+1)$$
$$\leq \kappa\left(1 + 4\sqrt{4|S^{\text{fluid}}| + 3T\log(|S^{\text{fluid}}|T) + \sqrt{|S^{\text{fluid}}|T\log(|S^{\text{fluid}}|T)}}\right)\log(T+1).$$

Since $\tilde{T}\geq 0$, i.e.,

$$T > \frac{\tilde{w}}{\underline{w}} + \sqrt{2\frac{\tilde{w}}{\underline{w}}T\log(|S^{\text{fluid}}|T)},$$

we have $q_i^{\text{fluid}} \geq 1$ for any product $i\in S^{\text{fluid}}$ and then $|S^{\text{fluid}}| \leq \sum_{i\in S^{\text{fluid}}} q_i^{\text{fluid}} \leq T$, thus there exist constants D_3 and D_4 such that

$$\mathbb{E}\left\{w_0 \max_{i\in S^{\text{fluid}}}\left\{\frac{\tilde{q}_i}{w_i}\right\}\log\left(\sum_{i\in S^{\text{fluid}}}\tilde{q}_i\right)\middle| X_i(\boldsymbol{q}^{\text{round}},\tilde{T}) < q_i^{\text{round}}, \forall i\in S^{\text{fluid}}\right\} \leq \left(D_3\sqrt{T\log(T)} + D_4\right)\log(T).$$

In conclusion, combining the analysis for terms (a) and (b), there exist constants C_1 and C_2 such that

$$\mathbb{E}\left\{X_0(\boldsymbol{q}^{\text{round}},T) - X_0(\boldsymbol{q}^{\text{round}},\tilde{T})\middle| X_i(\boldsymbol{q}^{\text{round}},\tilde{T}) < q_i^{\text{round}}, \forall i\in S^{\text{fluid}}\right\} \leq C_1\log(T)\sqrt{T\log(T)} + C_2.$$

■

Proof of Theorem 4.2:

We focus on problems with $T\geq 3$, otherwise, we have $V^{\text{fluid}} - \Pi(\boldsymbol{q}^{\text{round}}) \leq 4\bar{p}$. We consider two possible cases:

(i) The demand is relatively small, specifically,

$$T \leq \frac{\tilde{w}}{\underline{w}} + \sqrt{2\frac{\tilde{w}}{\underline{w}}T\log(|S^{\text{fluid}}|T)}.$$

We claim that $T \leq 8(\tilde{w}/\underline{w})\log(|S^{\text{fluid}}|T)$. Suppose not, then

$$\frac{\tilde{w}}{\underline{w}} + \sqrt{2\frac{\tilde{w}}{\underline{w}}T\log(|S^{\text{fluid}}|T)} < \frac{T}{2\log(|S^{\text{fluid}}|T)} + \frac{T}{2} < T,$$

which leads to a contradiction and thus the claim holds. By Lemma B.1, we have

$$\begin{aligned}
V^{\mathsf{fluid}} - \Pi(\boldsymbol{q}^{\mathsf{round}}) &\le \bar{p}\left(\mathbb{E}\left\{X_0(\boldsymbol{q}^{\mathsf{round}}, T)\right\} - \frac{Tw_0}{\tilde{w}} + 1\right) \\
&\le \bar{p}\left(w_0 \max_{i\in\mathcal{N}}\left\{\frac{q_i^{\mathsf{round}}}{w_i}\right\}\log\left(\sum_{i\in\mathcal{N}} q_i^{\mathsf{round}}\right) + \left[T - \sum_{i\in\mathcal{N}} q_i^{\mathsf{round}}\right]^+ - \frac{Tw_0}{\tilde{w}}\right) + 2\bar{p} \\
&\overset{(a)}{\le} \bar{p}\left(w_0 \max_{i\in\mathcal{N}}\left\{\frac{\lceil q_i^{\mathsf{fluid}}\rceil}{w_i}\right\}\log\left(\sum_{i\in S^{\mathsf{fluid}}} q_i^{\mathsf{fluid}} + 1\right) + T - \sum_{i\in S^{\mathsf{fluid}}} q_i^{\mathsf{fluid}} - \frac{w_0}{\tilde{w}}T\right) + 2\bar{p} \\
&= \bar{p} w_0 \max_{i\in\mathcal{N}}\left\{\frac{\lceil q_i^{\mathsf{fluid}}\rceil}{w_i}\right\}\log\left(\sum_{i\in S^{\mathsf{fluid}}} q_i^{\mathsf{fluid}} + 1\right) + 2\bar{p} \le \bar{p} w_0\left(\frac{T}{\tilde{w}} + \frac{1}{\underline{w}}\right)\log\left(T+1\right) + 2\bar{p} \\
&\overset{(b)}{\le} \bar{p}\kappa\left(8\log\left(|S^{\mathsf{round}}|T\right) + 1\right)\log\left(1+T\right) + 2\bar{p} \\
&\overset{(c)}{\le} \bar{p}\kappa\left(8\log\left(T\left(T+1\right)\right) + 1\right)\log\left(1+T\right) + 2\bar{p},
\end{aligned}$$

where (a) and (c) follows from Lemma 4.1 and (b) holds because $T \le 8(\tilde{w}/\underline{w})\log(|S^{\mathsf{fluid}}|T)$ and if $T > \tilde{w}/\underline{w}$, then $S^{\mathsf{round}} = S^{\mathsf{fluid}}$.

(ii) The demand is relatively large, specifically,

$$T > \frac{\tilde{w}}{\underline{w}} + \sqrt{2\frac{\tilde{w}}{\underline{w}} T \log(|S^{\mathsf{fluid}}|T)}.$$

Combining Lemma B.1, B.2, B.3 and B.4, we have

$$\begin{aligned}
V^{\mathsf{fluid}} - \Pi(\boldsymbol{q}^{\mathsf{round}}) &\le \bar{p}\left(\mathbb{E}\left\{X_0(\boldsymbol{q}^{\mathsf{round}}, T)\right\} - \frac{Tw_0}{\tilde{w}}\right) + \bar{p} \\
&\le \bar{p}\left(\mathbb{E}\left\{X_0(\boldsymbol{q}^{\mathsf{round}}, T)\middle| X_i(\boldsymbol{q}^{\mathsf{round}}, \tilde{T}) < q_i^{\mathsf{round}}, \forall i \in S^{\mathsf{fluid}}\right\}\mathbb{P}\left(X_0(\boldsymbol{q}^{\mathsf{round}}, \tilde{T}) < q_i^{\mathsf{round}}, \forall i\in S^{\mathsf{fluid}}\right) - \frac{w_0}{\tilde{w}}T + 2\right) \\
&\le \bar{p}\left(\mathbb{E}\left\{X_0(\boldsymbol{q}^{\mathsf{round}}, T)\middle| X_i(\boldsymbol{q}^{\mathsf{round}}, \tilde{T}) < q_i^{\mathsf{round}}, \forall i \in S^{\mathsf{fluid}}\right\} + 2 - \frac{w_0}{\tilde{w}}T\right) \\
&\le \bar{p}\left(C_1\log(T)\sqrt{T\log(T)} + C_2 + 2\right).
\end{aligned}$$

Combining case (i) and (ii) leads to the result. ■