Fluid Approximations for Revenue Management under High-Variance Demand

Yicheng Bai[1], Omar El Housni[1], Billy Jin[1], Paat Rusmevichientong[2], Huseyin Topaloglu[1], David P. Williamson[1]

[1]School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853
[2]Marshall School of Business, University of Southern California, Los Angeles, CA 90089

{yb279, oe46, bzj3}@cornell.edu, rusmevic@marshall.usc.edu, topaloglu@orie.cornell.edu, davidpwilliamson@cornell.edu

December 26, 2022

One of the most prevalent demand models in the revenue management literature is based on dividing the selling horizon into a number of time periods such that there is at most one customer arrival at each time period. This demand model is equivalent to using a discrete-time approximation to a Poisson process, but it has an important shortcoming. If the mean number of customer arrivals is large, then the coefficient of variation of the number of customer arrivals has to be small. In other words, large demand volume and large demand variability cannot co-exist in this demand model. In this paper, we start with a revenue management model that incorporates general mean and variance for the number of customer arrivals. This revenue management model has a random selling horizon length, capturing the distribution of the number of customer arrivals. The question we seek to answer is the form of the fluid approximation that corresponds to this revenue management model. It is tempting to construct the fluid approximation by computing the expected consumption of the resource capacities in the constraints and the total expected revenue in the objective function through the distribution of the number of customer arrivals. We demonstrate that this answer is wrong in the sense that it yields a fluid approximation that is not asymptotically tight as the resource capacities get large. We give an alternative fluid approximation, where, perhaps surprisingly, the distribution of the number of customer arrivals does not play any role in the constraints. We show that this fluid approximation is asymptotically tight as the resource capacities get large. A numerical study also demonstrates that the policies driven by the latter fluid approximation perform substantially better, so there is practical value in getting the fluid approximation right under high-variance demand.

1. Introduction

In the revenue management literature, one of the most prevalent demand models is based on a Bernoulli process, where we divide the selling horizon into a number of time periods such that there is at most one customer arrival at each time period. This demand model is equivalent to using a discrete-time approximation for a Poisson process and it has allowed us to build revenue management models that have had dramatic impact in practice over many decades; see Talluri and van Ryzin (2005). Using such a Bernoulli process as the demand model, however, has an important shortcoming. If the mean number of arrivals is to be large, then the coefficient of variation for the number of customer arrivals must be small. In particular, consider a selling horizon with T time periods. At time period t, we have a customer arrival with probability λ_t. In this case, the total expected number of customer arrivals is $\sum_{t=1}^{T} \lambda_t$, whereas the variance of the number of customer arrivals is $\sum_{t=1}^{T} \lambda_t (1-\lambda_t)$. Noting that $\sum_{t=1}^{T} \lambda_t (1-\lambda_t) \leq \sum_{t=1}^{T} \lambda_t$, the coefficient of variation of

the number of customer arrivals cannot exceed $1/\sqrt{\sum_{t=1}^{T} \lambda_t}$, corresponding to the reciprocal of the square root of the expected number of customer arrivals. Thus, if we would like to model large demand volume so that the expected number of customer arrivals is to be large, then the coefficient of variation of the number of customer arrivals must be small. In other words, large demand volume and large demand variability cannot co-exist in this demand model! Due to this observation, one may even believe that fluid approximations work well simply because a large mean demand under a Bernoulli process gives rise to a small coefficient of variation for the demand.

In this paper, we start with a revenue management model that incorporates a general mean and variance for the number of customer arrivals. Naturally, this revenue management model is based on a dynamic program that has a random selling horizon length. The distribution of the selling horizon length captures the distribution of the number of customer arrivals. The dynamic program allows us to formalize the problem with general mean and variance for the number of customer arrivals, but such a dynamic program involves a high-dimensional state variable when we have a large number of resources in consideration, so it is computationally difficult to solve. Fluid approximations, instead, have been an important workhorse for coming up with implementable policies in practice. The main question that we seek to answer is the form of a sound fluid approximation for our revenue management model with random number of customer arrivals.

We expect a sound fluid approximation to satisfy three properties. First, the optimal objective value of the fluid approximation should be an upper bound on the optimal total expected revenue. Thus, we can use the fluid approximation to assess the optimality gaps of heuristic policies. Second, the relative gap between the optimal objective value of the fluid approximation and the optimal total expected revenue should vanish as the resource capacities get large. In this way, we have confidence in the fluid approximation for systems with large resource availabilities. Third, the fluid approximation should provide a way to implement policies that are asymptotically optimal as the resource capacities get large. We demonstrate that a natural approach to extend the existing fluid approximations to random number of customer arrivals does not satisfy the last two properties. We correct this natural approach and give a fluid approximation satisfying all three properties.

A natural approach for constructing a fluid approximation under random number of customer arrivals uses the distribution of the number of customer arrivals to compute the expected consumption of the resource capacities in the constraints and the total expected revenue in the objective function. While such a fluid approximation provides an upper bound on the optimal total expected revenue, we give a problem instance to demonstrate that the relative gap between the optimal objective value of this fluid approximation and the optimal total expected revenue does not vanish as the resource capacities get large. In particular, we give a problem instance parameterized

by an integer k, where the mean and standard deviation of the number of customer arrivals are, respectively, $k+\sqrt{k}$ and $k\sqrt{k-1}$, yielding a coefficient of variation of $\frac{k\sqrt{k-1}}{k+\sqrt{k}} = \frac{\sqrt{k-1}}{1+\sqrt{1/k}}$. Thus, the mean and coefficient of variation for the number of customer arrivals in this problem instance can both be large when k is large. There is a single resource with a capacity of $k\sqrt{k}+\sqrt{k}$. We set up the customer arrival process so that the optimal total expected revenue turns out to be $2\sqrt{k}$, but the optimal objective value of the natural fluid approximation is $k+\sqrt{k}$. Thus, the ratio between the optimal total expected revenue and the optimal objective value of the fluid approximation is $\frac{2}{\sqrt{k}+1}$, which does not approach one as the resource capacity gets large. Quite the contrary, the ratio converges to zero as the resource capacity gets large, so the natural fluid approximation becomes especially poor, as opposed to being especially good, when the resource capacity gets large.

We give an alternative fluid approximation, where, surprisingly, the distribution of the number of customer arrivals does not play a role in the constraints at all. We show that the optimal objective value of this fluid approximation provides an upper bound on the optimal total expected revenue (Theorem 4.1). Letting $c_{\min}$ be the smallest capacity for a resource, we show that the ratio between the optimal total expected revenue and the optimal objective value of the fluid approximation is $\Omega\left(1-\sqrt{\frac{\log c_{\min}}{c_{\min}}}\right)$, which approaches one as the resource capacities get large (Theorem 5.1). When establishing this result, we also show that we can use our alternative fluid approximation to come up with a policy that obtains at least $\Omega\left(1-\sqrt{\frac{\log c_{\min}}{c_{\min}}}\right)$ fraction of the optimal total expected revenue. For the problem instance in the previous paragraph, the optimal objective value of our fluid approximation is $2\sqrt{k}$, which is exactly the optimal total expected revenue.

Thus, we make three main contributions. First, we give the "right" fluid approximation under random number of customer arrivals to satisfy all three properties mentioned earlier. The form of our fluid approximation is somewhat unexpected, as it does not use the distribution of the number of customer arrivals in the capacity constraints. Second, our work addresses the possible misconception that fluid approximations are asymptotically tight simply because the standard Bernoulli process gives rise to small coefficient of variation for the demand when the mean of the demand is large. We do not need a Bernoulli process to construct asymptotically tight fluid approximations. It is possible to build fluid approximations with sound footing under high-variance demand. Third, it is important to get the fluid approximation right. Policies driven by a naive fluid approximation can have poor performance under random number of customer arrivals.

Studying fluid approximations under arrival processes other than a Bernoulli process is not an intellectual curiosity. We give a numerical study to check the practical benefits from getting the fluid approximation right under high-variance demand. Policies driven by our fluid approximation perform up to 13% better than those driven by the naive fluid approximation. Furthermore, demand

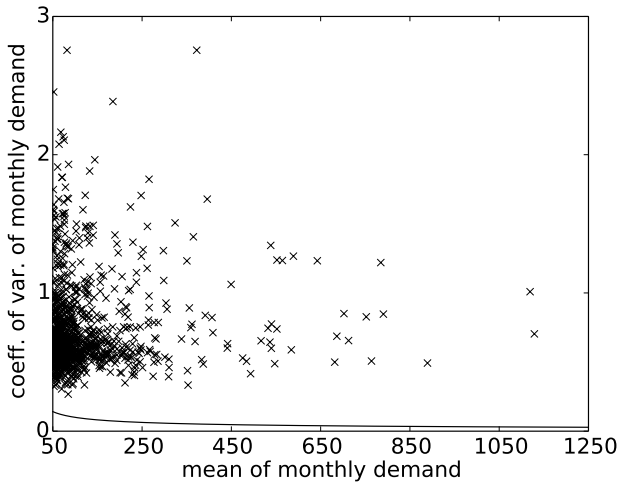

Figure 1 Mean and coefficient of variation of the monthly demand for different products.

data can display variance significantly larger than what is implied by a discrete-time approximation to a Poisson process. We studied a publicly available dataset from an online electronics retailer; see Kaggle (2021). The dataset includes order transactions over nine months. In Figure 1, each cross corresponds to a different product, plotting the mean and coefficient variation of the monthly demand for the product. There are many products with mean monthly demand of about 50 and coefficient of variation exceeding one. If the demand were driven by a Poisson process, then a mean of 50 for the demand would imply a coefficient of variation of $1/\sqrt{50}$, which is only about 0.14. In the same figure, the solid line plots $1/\sqrt{x}$ as a function of x, which is the coefficient of variation under Poisson demand arrivals corresponding to a mean monthly demand of x. The crosses lie substantially above the solid line, indicating that the coefficient of variation of the monthly demand is larger than what we would observe if the demand were driven by a Poisson process.

We give our results for revenue management problems under the independent demand model with multiple resources, where we control the availability of the products, the sale of a product consumes a combination of resources and each customer arrives into the system to purchase a fixed product in mind, purchasing only this product if it is available for sale. This model is known as the independent demand model with a network of resources. Our results easily extend when the demand for each product depends on the prices or the assortment of available products.

Related Literature. Fluid approximations have been studied under demand models based on a Bernoulli process. Gallego and van Ryzin (1994) focus on the case with a single resource and a single product, where the demand for the product depends on its price and the sale for the product consumes the capacity of the resource. The authors consider an asymptotic regime where the expected number of customer arrivals and the resource capacity scale linearly with k. They use a fluid approximation to give a policy with a relative performance guarantee of $\Omega\left(1-\frac{1}{\sqrt{k}}\right)$. The

policy is based on solving the fluid approximation once at the beginning of the selling horizon. Gallego and van Ryzin (1997) generalize these results to the case with multiple resources and products, where the sale of a product consumes the capacities of a combination of resources. Talluri and van Ryzin (1998) show the asymptotic optimality of a policy driven by a dual solution to the fluid approximation in the same asymptotic regime. Liu and van Ryzin (2008) and Gallego et al. (2004) generalize these results to models with customer choice, where the customers choose among the products offered to them. Cooper (2002) focuses on demand processes where the total demand divided by k converges to a fixed number in distribution and characterizes the same type of relative gaps. Jasin and Kumar (2012) consider the case where the fluid approximation is solved periodically and show that the ratio between the total expected revenue of the policy derived from the fluid approximation and the optimal total expected revenue can be much larger than $\Omega\left(1-\frac{1}{\sqrt{k}}\right)$. Balseiro et al. (2021) generalize the ideas in the last paper to give a unified analysis for different demand models while periodically solving the fluid approximation. Rusmevichientong et al. (2020) show that the ratio between the optimal total expected revenue and the optimal objective value of a fluid approximation is $\Omega\left(1-\frac{1}{\sqrt[3]{c_{\min}}}\right)$, where $c_{\min}$ is the smallest capacity for a resource. The expected demand in this paper does not necessarily have to be scaled, so their asymptotic regime is more general. Similarly, Feng et al. (2022) establish a ratio of $\Omega\left(1-\frac{1}{\sqrt{c_{\min}}}\right)$.

The papers discussed so far use a Bernoulli process. Walczak (2006) incorporates high variance into demand by using dynamic programs with batch arrivals but does not give efficiently computable policies with performance guarantees. Besbes and Saure (2014) consider dynamic pricing problems when the price-demand function changes at a random time period. Using a price-demand function of zero after the change, the authors can capture random number of customer arrivals. They study the structural properties of the solution obtained through a fluid approximation similar to ours and give performance guarantees in an asymptotic regime where the expected number of customer arrivals and the resource capacity scale linearly. In a concurrent and independent work, Aouad and Ma (2022) consider matching problems and develop a fluid approximation similar to ours. The authors give a policy with a performance guarantee of $\Omega\left(1-\frac{1}{\sqrt{c_{\min}}}\right)$, but they do not consider products consuming combinations of resources. When we submitted our work, their paper was not publicly available, but they apparently had already derived their fluid approximation, so our paper and theirs are independent discoveries.

Organization. In Section 2, we give our revenue management model with random number of customer arrivals. In Section 3, we show the pitfalls of a naive fluid approximation. In Section 4, we formulate our fluid approximation and show that it yields an upper bound on the optimal total expected revenue. In Section 5, we establish the gap of $\Omega\left(1-\sqrt{\frac{\log c_{\min}}{c_{\min}}}\right)$ for the optimal objective value of our fluid approximation. In Section 6, we give a numerical study.

2. Revenue Management with Random Number of Arrivals

We give a revenue management model under a general distribution for the number of customer arrivals. The set of resources is $\mathcal{L}$. The capacity of resource i is c_i. The set of products is $\mathcal{J}$. The revenue of product j is f_j. If we make a sale for product j, then we consume the capacities of the resources in the set $A_j \subseteq \mathcal{L}$. The number of customer arrivals is a random variable taking values in $\mathcal{T} = \{1, \ldots, T\}$. Using the random variable D to capture the number of customer arrivals, we characterize this random variable with its survival rate $\rho_t = \mathbb{P}\{D \geq t+1 \,|\, D \geq t\}$. For simplicity, there is a customer arrival at each time period with probability one, so the number of customer arrivals corresponds to the number of time periods. With probability λ_{jt}, the customer arriving at time period t is interested in purchasing product j. If this product is available for purchase, then the customer purchases it. Otherwise, she leaves without a purchase. The number of customer arrivals, as well as the product of interest to each arriving customer, are all independent. We want to find a policy to offer a set of products at each time period to maximize the total expected revenue. We use the vector $\boldsymbol{x} = (x_i : i \in \mathcal{L})$ to capture the state of the system, where x_i is the remaining capacity for resource i. Using $\boldsymbol{e}_i \in \{0,1\}^{|\mathcal{L}|}$ to denote the i-th unit vector, we can find the optimal policy by computing the value functions $\{J_t : t \in \mathcal{T}\}$ through the dynamic program

$$J_t(\boldsymbol{x}) = \sum_{j \in \mathcal{J}} \lambda_{jt} \max \left\{ f_j + \rho_t \, J_{t+1}\Big(\boldsymbol{x} - \sum_{i \in A_j} \boldsymbol{e}_i\Big) \,,\, \rho_t \, J_{t+1}(\boldsymbol{x}) \right\},$$

with the boundary condition that $J_{T+1} = 0$. In the dynamic program above, we follow the convention that $J_t(\boldsymbol{x}) = -\infty$ whenever $x_i < 0$ for some $i \in \mathcal{L}$.

Given that the customer arriving at time period t is interested in purchasing product j, the two terms in the maximum operator in the dynamic program above correspond to making product j available and not available. In either case, we have another customer arrival only with probability ρ_t. In our model, each customer arrives with the intention of purchasing a fixed product. Our decision is whether to make this product available. This approach keeps our fluid approximation as simple as possible, while allowing us to discuss the intricacies under random number of customer arrivals, but we can give analogous fluid approximations when the demand for each product depends on the prices or the assortment of available products. Furthermore, the sale of a product consumes at most one unit of the capacity for a resource, but we can work with the case where the sale of a product consumes multiple units of the capacity for a resource. These extensions bring notational overhead without making our results more insightful. Lastly, having a finite upper bound of T on the number of customer arrivals is reasonable from practical perspective, because we can choose the upper bound as large as we would like. Nevertheless, later in the paper, we discuss dealing with the case without a finite upper bound on the number of customer arrivals. We proceed to discussing fluid approximations corresponding to the dynamic program above.

3. A Natural Fluid Approximation and its Pitfalls

We give a first fluid approximation that corresponds to the dynamic program in Section 2. This fluid approximation could be viewed as a natural one, because we use the distribution of the number of customer arrivals to compute the expected capacity consumption of each resource in the constraints and the total expected revenue in the objective function. Furthermore, if the number of customer arrivals is fixed, so that $D = T$ with probability one, then this fluid approximation reduces to the traditional fluid approximation under a Bernoulli process that already appears in the literature. However, we will see that if the number of customer arrivals is random, then the relative gap between the optimal objective value of this fluid approximation and the optimal total expected revenue does not vanish as the resource capacities get large. We use the decision variables $\boldsymbol{y} = (y_{jt} : j \in \mathcal{J},\ t \in \mathcal{T}) \in \mathbb{R}_+^{|\mathcal{J}||\mathcal{T}|}$, where y_{jt} is the expected number of purchases for product j at time period t given that the length of the selling horizon reaches beyond time period t. Using $\mathbf{1}_{(\cdot)}$ to denote the indicator function, consider the linear program

$$\max_{\boldsymbol{y} \in \mathbb{R}_+^{|\mathcal{J}||\mathcal{T}|}} \left\{ \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} f_j \, \mathbb{P}\{D \geq t\} \, y_{jt} \ : \ \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} \mathbf{1}_{(i \in A_j)} \, \mathbb{P}\{D \geq t\} \, y_{jt} \leq c_i \quad \forall \, i \in \mathcal{L} \right.$$
$$\left. y_{jt} \leq \lambda_{jt} \quad \forall \, j \in \mathcal{J},\ t \in \mathcal{T} \right\}, \qquad \text{(Traditional Fluid)}$$

where the objective function accounts for the total expected revenue and the first constraint ensures that the total expected capacity consumption of resource i does not exceed its capacity.

If $D = T$ with probability one, so that the number of customer arrivals is fixed, then we have $\mathbb{P}\{D \geq t\} = 1$ for all $t = 1, \ldots, T$ and the problem above reduces to the fluid approximation under a Bernoulli process given in Section 3.3.1 in Talluri and van Ryzin (2005). Thus, we refer to the linear program above as the Traditional Fluid. If the number of customer arrivals is fixed, then the relative gap between the optimal objective value of the Traditional Fluid and the optimal total expected revenue vanishes as the resource capacities get large, so the Traditional Fluid is appropriate under a Bernoulli process. We give a problem instance to show that if the number of customer arrivals is random, then the relative gap between the optimal objective value of the Traditional Fluid and the optimal total expected revenue does not vanish as the resource capacities get large. Thus, one may believe that we cannot give a fluid approximation satisfying the vanishing relative gap property under random number of customer arrivals, but this belief is not correct. In the next section, we will give an alternative fluid approximation satisfying the vanishing relative gap property.

Consider a problem instance with one resource and one product. The product has a revenue of one and it consumes the capacity of the single resource. The number of time periods in the

selling horizon has two possible values with $\mathbb{P}\{D=\sqrt{k}\}=1-\frac{1}{k}$ and $\mathbb{P}\{D=k^2+\sqrt{k}\}=\frac{1}{k}$. Thus, $\mathbb{E}\{D\}=k+\sqrt{k}$ and $\mathrm{Var}(D)=k^2(k-1)$. The capacity of the resource is $k\sqrt{k}+\sqrt{k}$. At all time periods, an arriving customer requests the product with probability one. We compute the optimal total expected revenue. Because there is a single product, it is optimal to accept all customer requests as much as the capacity allows. There are only two possible values for D. If $D=\sqrt{k}$, then the capacity of $k\sqrt{k}+\sqrt{k}$ allows us to accept all customer requests, so we get a total expected revenue of $\sqrt{k}$, where we use the fact that the product has a revenue of one. If $D=k^2+\sqrt{k}$, then we can sell all of the capacity of $k\sqrt{k}+\sqrt{k}$, so we get a total expected revenue of $k\sqrt{k}+\sqrt{k}$. Thus, the optimal total expected revenue is given by $\mathbb{P}\{D=\sqrt{k}\}\sqrt{k}+\mathbb{P}\{D=k^2+\sqrt{k}\}(k\sqrt{k}+k)=\left(1-\frac{1}{k}\right)\sqrt{k}+\frac{1}{k}(k\sqrt{k}+\sqrt{k})=2\sqrt{k}$. On the other hand, dropping the indices for the single resource and product, the Traditional Fluid approximation for this problem instance is

$$\max_{\boldsymbol{y}\in\mathbb{R}_+^{k^2}}\left\{\sum_{t=1}^{\sqrt{k}}y_t+\frac{1}{k}\sum_{t=\sqrt{k}+1}^{k^2+\sqrt{k}}y_t\ :\ \sum_{t=1}^{\sqrt{k}}y_t+\frac{1}{k}\sum_{t=\sqrt{k}+1}^{k^2+\sqrt{k}}y_t\le k\sqrt{k}+\sqrt{k},\ \ y_t\le 1\ \ \forall\, t=1,\ldots,k^2+\sqrt{k}\right\}.$$

Setting all of the decision variables to their upper bounds of one provides a feasible, as well as an optimal, solution with the objective value $\sqrt{k}+\frac{1}{k}k^2=k+\sqrt{k}$, as $k+\sqrt{k}\le k\sqrt{k}+\sqrt{k}$.

Thus, the ratio between the optimal total expected revenue and the optimal objective value of the Traditional Fluid approximation is $\frac{2\sqrt{k}}{k+\sqrt{k}}=\frac{2}{\sqrt{k}+1}$, which does not approach one as k gets large. Quite the contrary, this ratio converges to zero, which indicates that the Traditional Fluid approximation gets arbitrarily poor as k gets large. In the next section, we give another fluid approximation that makes up for the shortcomings of the Traditional Fluid approximation. Our formulation for the Traditional Fluid approximation uses one decision variable for each product and time period. Closing this section, we give an equivalent reformulation of the Traditional Fluid that aggregates the decision variables for each product over the time periods, using one decision variable for each product. In our equivalent reformulation, we use the decision variables $\boldsymbol{w}=(w_j:j\in\mathcal{J})\in\mathbb{R}_+^{|\mathcal{J}|}$, where w_j is the total expected number of purchases for product j over the whole selling horizon. In this case, we can write the Traditional Fluid approximation equivalently as

$$\max_{\boldsymbol{w}\in\mathbb{R}_+^{|\mathcal{J}|}}\left\{\sum_{j\in\mathcal{J}}f_j\,w_j\ :\ \sum_{j\in\mathcal{J}}\mathbf{1}_{(i\in A_j)}\,w_j\le c_i\ \ \forall\, i\in\mathcal{L},\quad w_j\le\sum_{t\in\mathcal{T}}\mathbb{P}\{D\ge t\}\,\lambda_{jt}\ \ \forall\, j\in\mathcal{J}\right\}.\quad (\textsf{Compact})$$

In particular, if $\boldsymbol{w}^*$ is optimal to the Compact problem, then setting $\hat{y}_{jt}=\lambda_{jt}\frac{w_j^*}{\sum_{k\in\mathcal{T}}\mathbb{P}\{D\ge k\}\,\lambda_{jk}}$ yields an optimal solution to the Traditional Fluid approximation.

Total expected demand for product j is $\sum_{t\in\mathcal{T}}\mathbb{P}\{D\ge t\}\,\lambda_{jt}$, so by the second constraint above, the total expected purchases for product j does not exceed its total expected demand.

4. A Fluid Approximation that Checks the Boxes

To give our alternative approximation, using the decision variables $\boldsymbol{y} = (y_{jt} : j \in \mathcal{J},\ t \in \mathcal{T}) \in \mathbb{R}_+^{|\mathcal{J}||\mathcal{T}|}$ with the same interpretation as in the previous section, we consider the linear program

$$Z_{\mathsf{LP}}^* = \max_{\boldsymbol{y} \in \mathbb{R}_+^{|\mathcal{J}||\mathcal{T}|}} \left\{ \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} f_j \, \mathbb{P}\{D \geq t\} \, y_{jt} \ : \ \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} \mathbf{1}_{(i \in A_j)} \, y_{jt} \leq c_i \ \ \forall i \in \mathcal{L} \right.$$
$$\left. y_{jt} \leq \lambda_{jt} \ \ \forall j \in \mathcal{J},\ t \in \mathcal{T} \right\}. \qquad \text{(Universal Fluid)}$$

We refer to the linear program above as the Universal Fluid approximation because it will provide an asymptotically tight upper bound on the optimal total expected revenue, even when we have a random number of customer arrivals. We comment on the unexpected form of the Universal Fluid approximation. The objective function accounts for the total expected revenue over the selling horizon. Given that the length of the selling horizon reaches beyond time period t, the last constraint ensures that the expected number of purchases for product j at time period t does not exceed the expected number of requests for the product. However, it is difficult to interpret the left side of the first constraint as the expected capacity consumption of resource i, because the distribution of the length of the selling horizon does not appear in this constraint. We may believe that the first constraint is tighter than it needs to be, because the sales at time period t consume the capacity of a resource in this constraint irrespective of whether the length of the selling horizon reaches beyond time period t. Thus, it is not immediately clear that the Universal Fluid approximation yields an upper bound on the optimal total expected revenue.

Upper Bound on the Optimal Total Expected Revenue:

In the next theorem, we show that the optimal objective value of the Universal Fluid approximation is indeed an upper bound on the optimal total expected revenue.

Theorem 4.1 (Upper Bound) *Letting $\boldsymbol{c} = (c_i : i \in \mathcal{L})$, noting that $J_1(\boldsymbol{c})$ is the optimal total expected revenue, we have $Z_{\mathsf{LP}}^* \geq J_1(\boldsymbol{c})$.*

The proof of the theorem is in Appendix A. We give an overview of the proof. Considering the dynamic program in Section 2, we can accept a request for product j only if the remaining capacities of the resources satisfy $\boldsymbol{x} - \sum_{i \in A_j} \boldsymbol{e}_i \geq 0$. Expressing these capacity constraints at time period t as $x_i - \mathbf{1}_{(i \in A_j)} \geq 0$ for all $i \in \mathcal{L}$ and $j \in \mathcal{J}$, we relax them by associating the Lagrange multipliers $\boldsymbol{\theta} = (\theta_{ijt} : i \in \mathcal{L},\ j \in \mathcal{J},\ t \in \mathcal{T}) \in \mathbb{R}_+^{|\mathcal{L}||\mathcal{J}||\mathcal{T}|}$ with them, yielding a relaxed dynamic program. Let $\{\hat{J}_t^{\boldsymbol{\theta}} : t \in \mathcal{T}\}$ be the value functions of the relaxed dynamic program, where we make it explicit that the value functions of the relaxed dynamic program depend on the Lagrange multipliers. We can establish two results. First, the value functions of the relaxed dynamic program provide upper

bounds on the value functions of the dynamic program in Section 2. Thus, for any $\boldsymbol{\theta} \in \mathbb{R}_+^{|\mathcal{L}||\mathcal{J}||\mathcal{T}|}$, we have $\hat{J}_1^{\boldsymbol{\theta}}(\boldsymbol{c}) \geq J_1(\boldsymbol{c})$. Second, we give a closed form expression for the value functions of the relaxed dynamic program under a specific choice of the Lagrange multipliers. In particular, letting $R_t = \rho_t \, \rho_{t+1} \ldots \rho_{T-1}$ with $R_T = 1$ for notational brevity, for any $\boldsymbol{\eta} \in \mathbb{R}_+^{|\mathcal{L}|}$, if we choose the Lagrange multipliers as $\theta_{ijT} = \eta_i$ for all $i \in \mathcal{L}$ and $j \in \mathcal{J}$, whereas $\theta_{ijt} = 0$ for all $i \in \mathcal{L}$, $j \in \mathcal{J}$ and $t \in \mathcal{T} \setminus \{T\}$, then we have $\hat{J}_1^{\boldsymbol{\theta}}(\boldsymbol{c}) = \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} \lambda_{jt} \big[\frac{R_1}{R_t} f_j - R_1 \sum_{i \in A_j} \eta_i \big]^+ + R_1 \sum_{i \in \mathcal{L}} \eta_i \, c_i$. Thus, we can minimize the last expression over all $\boldsymbol{\eta} \in \mathbb{R}_+^{|\mathcal{L}|}$ to obtain an upper bound on $J_1(\boldsymbol{c})$. We formulate the problem of minimizing the last expression over all $\boldsymbol{\eta} \in \mathbb{R}_+^{|\mathcal{L}|}$ as a linear program and the dual of this linear program is the Universal Fluid approximation. Note that the constraints in the Universal Fluid are at least as tight as those in the Traditional Fluid, so by Theorem 4.1, the optimal objective value of the Traditional Fluid is also an upper bound on the optimal total expected revenue. However, as discussed in the previous section, the Traditional Fluid approximation is not asymptotically tight as the resource capacities get large. Our proof of Theorem 4.1 through Lagrangian relaxation in Appendix A provides a constructive approach for establishing that the Universal Fluid provides an upper bound on the optimal total expected revenue. Next, we give a discussion on how the pieces of the Universal Fluid fit together to provide an upper bound on the optimal total expected revenue, but this discussion does not derive the form of the Universal Fluid approximation.

Let $\nu_{jt} : \mathbb{Z}_+^{|\mathcal{L}|} \to \{0, 1\}$ be the decision function of the optimal policy, where $\nu_{jt}(\boldsymbol{x}) = 1$ if and only if the optimal policy accepts a request for product j at time period t when the capacities of the resources are $\boldsymbol{x}$. In particular, noting the dynamic program in Section 2, the decision function of the optimal policy is given by $\nu_{jt}(\boldsymbol{x}) = \mathbf{1}_{(f_j \geq \rho_t \, (J_{t+1}(\boldsymbol{x}) - J_{t+1}(\boldsymbol{x} - \sum_{i \in A_j} \boldsymbol{e}_i)))}$. We use the random variable P_t to capture the product requested at time period t, so $P_t = j$ with probability λ_{jt}. For all $t = 1, \ldots, T$, we use the random variable $\boldsymbol{X}_t$ to capture the resource capacities at time period t under the optimal policy. Because the product requests are random, the resource capacities at each time period under the optimal policy are also random. The random variables $\{\boldsymbol{X}_t : t \in \mathcal{T}\}$ are recursively defined as $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{J}} \mathbf{1}_{(i \in A_j)} \, \mathbf{1}_{(P_t = j)} \, \nu_{jt}(\boldsymbol{X}_t) \, \boldsymbol{e}_i$ with the boundary condition that $\boldsymbol{X}_1 = \boldsymbol{c}$. The random variables $\{\boldsymbol{X}_t : t \in \mathcal{T}\}$ and $\{P_t : t \in \mathcal{T}\}$ are independent of D. In this case, the total revenue of the optimal policy is given by the random variable $\sum_{t \in \mathcal{T}} \mathbf{1}_{(D \geq t)} \sum_{j \in \mathcal{J}} f_j \, \mathbf{1}_{(P_t = j)} \, \nu_{jt}(\boldsymbol{X}_t)$, where we use the fact that the number of purchases for product j at time period t under the optimal policy is given by $\mathbf{1}_{(P_t = j)} \, \nu_{jt}(\boldsymbol{X}_t)$ and the optimal policy makes a sale at time period t only if $D \geq t$. Thus, letting $\bar{y}_{jt} = \mathbb{E}\{\mathbf{1}_{(P_t = j)} \, \nu_{jt}(\boldsymbol{X}_t)\}$ for notational brevity, taking the expectation of the last expression, the optimal total expected revenue is $J_1(\boldsymbol{c}) = \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} f_j \, \mathbb{P}\{D \geq t\} \, \bar{y}_{jt}$. The decisions of the optimal policy have to satisfy the capacity constraints even when the number of customers take on their largest possible value of T, so $\sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} \mathbf{1}_{(i \in A_j)} \, \mathbf{1}_{(P_t = j)} \, \nu_{jt}(\boldsymbol{X}_t) \leq c_i$ for all $i \in \mathcal{L}$. Taking the

expectation of both sides of the last inequality, we get $\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\mathbf{1}_{(i\in A_j)}\,\bar{y}_{jt}\le c_i$. Because $\mathbf{1}_{(P_t=j)}$ is a Bernoulli random variable with parameter λ_{jt}, we have $\bar{y}_{jt}=\mathbb{E}\{\mathbf{1}_{(P_t=j)}\,\nu_{jt}(\boldsymbol{X}_t)\}\le\lambda_{jt}$. By the last two inequalities, the solution $(\bar{y}_{jt}:j\in\mathcal{J},\ t\in\mathcal{T})$ is feasible to the Universal Fluid approximation. Furthermore, noting that $J_1(\boldsymbol{c})=\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}f_j\,\mathbb{P}\{D\ge t\}\,\bar{y}_{jt}$, this solution provides an objective value of $J_1(\boldsymbol{c})$ for the Universal Fluid approximation. Thus, the optimal objective value of the Universal Fluid approximation must be at least as large as $J_1(\boldsymbol{c})$. The preceding discussion can be viewed as an alternative proof for Theorem 4.1, but our proof in Appendix A is constructive, deriving the form of the Universal Fluid approximation, whereas the discussion in this paragraph verifies that the optimal objective value of the Universal Fluid approximation is an upper bound on the optimal total expected revenue, only after guessing the form of this approximation.

Relationship Between the Fluid Approximations:

There is an interesting relationship between the two fluid approximations that we discussed so far. In particular, we can obtain the Traditional Fluid approximation by aggregating the constraints in the Universal Fluid approximation. Therefore, we can view the Traditional Fluid approximation as a crude version of the Universal Fluid approximation obtained by aggregating the constraints in the latter. To see this relationship, we write the first constraint in the Universal Fluid approximation as $\sum_{t=1}^{\kappa}\sum_{j\in\mathcal{J}}\mathbf{1}_{(i\in A_j)}\,y_{jt}\le c_i$ for all $\kappa=1,\ldots,T$ and $i\in\mathcal{L}$. In the last set of constraints, the ones with $\kappa=T$ are the tightest, which are the constraints that we impose in the Universal Fluid. The constraints with $\kappa=1,\ldots,T-1$ are redundant, but adding them to the Universal Fluid approximation does not change the optimal objective value. For fixed $i\in\mathcal{L}$, multiplying the constraints in this paragraph with $\mathbb{P}\{D=\kappa\}$ and adding them over all $\kappa=1,\ldots,\tau$, we obtain the constraint

$$\begin{aligned} c_i \;=\; \sum_{\kappa=1}^{T}\mathbb{P}\{D=\kappa\}\,c_i \;&\ge\; \sum_{\kappa=1}^{T}\mathbb{P}\{D=\kappa\}\sum_{t=1}^{\kappa}\sum_{j\in\mathcal{J}}\mathbf{1}_{(i\in A_j)}\,y_{jt} \\ &=\; \sum_{t=1}^{T}\sum_{\kappa=t}^{T}\mathbb{P}\{D=\kappa\}\sum_{j\in\mathcal{J}}\mathbf{1}_{(i\in A_j)}\,y_{jt} \;=\; \sum_{t=1}^{T}\mathbb{P}\{D\ge t\}\sum_{j\in\mathcal{J}}\mathbf{1}_{(i\in A_j)}\,y_{jt}. \end{aligned}$$

In the chain of inequalities above, the second equality follows by interchanging the order of the sums and the last equality holds because $\sum_{\kappa=t}^{T}\mathbb{P}\{D=\kappa\}=\mathbb{P}\{D\ge t\}$. The chain of inequalities above corresponds to the first constraint in the Traditional Fluid approximation. Therefore, we can obtain the first constraint in the Traditional Fluid approximation by aggregating the first constraint in the Universal Fluid approximation. This discussion provides some support for the Traditional Fluid approximation, but it also demonstrates that aggregating the first constraint in the Universal Fluid approximation can yield a fluid approximation with an optimal objective value that is not asymptotically tight as the resource capacities get large. Also, we can reformulate the Traditional Fluid as the Compact problem, so this approximation only uses the total expected demand for each product, but we are not aware of a similar reformulation for the Universal Fluid.

5. Asymptotic Tightness

The problem instance that we give in Section 3 demonstrates that the relative gap between the optimal objective value of the Traditional Fluid approximation and the optimal total expected revenue does not vanish as the resource capacities get large. In this section, we show that the same relative gap for the optimal objective value of the Universal Fluid approximation vanishes as the resource capacities get large. To show this result, letting $\boldsymbol{y}^* = (y^*_{jt} : j \in \mathcal{J}\ t \in \mathcal{T})$ be an optimal solution to the Universal Fluid approximation, we consider an approximate policy that makes its decisions as follows. For some tuning parameter $\theta \in (0,1)$, if there is enough resource capacity to serve a request for product j at time period t, then the approximate policy makes product j available for purchase with probability $\theta\, y^*_{jt}/\lambda_{jt}$. If we do not have enough resource capacity to serve a request for product j at time period t, then the approximate policy does not make product j available. Letting $c_{\min} = \min_{i\in\mathcal{L}} c_i$, we choose the tuning parameter as $\theta = 1 - \sqrt{\frac{2\log c_{\min}}{c_{\min}}}$. Letting $J_{\mathsf{App}}(\boldsymbol{c})$ be the total expected revenue of the approximate policy, noting that $J_1(\boldsymbol{c})$ is the optimal total expected revenue, we have $J_1(\boldsymbol{c}) \geq J_{\mathsf{App}}(\boldsymbol{c})$. In the next theorem, we lower bound the total expected revenue of the approximate policy.

Theorem 5.1 (Asymptotic Tightness) *Letting $c_{\min} = \min_{i\in\mathcal{L}} c_i$ and $L = \max_{j\in\mathcal{J}} |A_j|$, the total expected revenue of the approximate policy and the optimal total expected revenue satisfy*

$$\frac{J_1(\boldsymbol{c})}{Z^*_{\mathsf{LP}}} \;\geq\; \frac{J_{\mathsf{App}}(\boldsymbol{c})}{Z^*_{\mathsf{LP}}} \;\geq\; 1 - \sqrt{\frac{2\log c_{\min}}{c_{\min}}} - \frac{L}{c_{\min}}.$$

We give the proof of the theorem in Appendix B. Because $1 \geq \frac{J_1(\boldsymbol{c})}{Z^*_{\mathsf{LP}}}$ by Theorem 4.1, the theorem above implies that $\frac{J_1(\boldsymbol{c})}{Z^*_{\mathsf{LP}}}$ converges to one as $c_{\min}$ gets large. Thus, the relative gap between the optimal objective value of the Universal Fluid approximation and the optimal total expected revenue vanishes as the resource capacities get large, which implies that the Universal Fluid satisfies the second property in the introduction that we would expect from a sound fluid approximation. Furthermore, because $1 \geq \frac{J_{\mathsf{App}}(\boldsymbol{c})}{J_1(\boldsymbol{c})} \geq \frac{J_{\mathsf{App}}(\boldsymbol{c})}{Z^*_{\mathsf{LP}}}$, the theorem above also implies that $\frac{J_{\mathsf{App}}(\boldsymbol{c})}{J_1(\boldsymbol{c})}$ converges to one as $c_{\min}$ gets large. Thus, the relative gap between the total expected revenue of the approximate policy and the optimal total expected revenue vanishes as the resource capacities get large. In this case, the Universal Fluid satisfies the third property in the introduction that we would expect from a sound fluid approximation. Note that when L is fixed, the theorem above implies that $\frac{J_1(\boldsymbol{c})}{Z^*_{\mathsf{LP}}} = \Omega\Big(1-\sqrt{\frac{\log c_{\min}}{c_{\min}}}\Big)$ and $\frac{J_{\mathsf{App}}(\boldsymbol{c})}{Z^*_{\mathsf{LP}}} = \Omega\Big(1-\sqrt{\frac{\log c_{\min}}{c_{\min}}}\Big)$. The tuning parameter θ trades off the probability that the approximate policy runs out of resource capacities to serve a request for a product with the total expected revenue collected by the approximate policy. In the proof of Theorem 5.1, we choose $\theta = 1 - \sqrt{\frac{2\log c_{\min}}{c_{\min}}}$. Building on Lemma E.1 in Bai et al. (2022), we can also choose $\theta = 1$ to

obtain a similar relative performance guarantee. This guarantee would still converge to one as $c_{\min}$ gets large but it would depend on the ratio $\max_{j\in\mathcal{J}} f_j / \min_{j\in\mathcal{J}} f_j$, as well as $c_{\min}$ and L.

Theorem 5.1 also implies that the Universal Fluid approximation is asymptotically tight in a regime that proportionally scales the resource capacities and the total expected demand by an integer parameter k. Consider a sequence of problems $\{\mathcal{P}^k : k \in \mathbb{Z}_+\}$. In problem $\mathcal{P}^k$, the capacity of resource i is $c_i^k = k\,c_i$. We use the random variable D^k to capture the number of customer arrivals in problem $\mathcal{P}^k$. The support of D^k is given by $\mathcal{T}^k = \{1,\ldots,k\,T\}$, whereas the distribution of D^k is given by $\mathbb{P}\{D^k \geq t\} = \mathbb{P}\{D \geq \lceil t/k \rceil\}$ for all $t = 1,\ldots,k\,T$. The probability that we have a request for product j at time period t is $\lambda_{jt}^k = \lambda_{j,\lceil t/k \rceil}$. Problem $\mathcal{P}^1$ corresponds to the problem formulated in Section 2. We can view problem $\mathcal{P}^k$ as a version of problem $\mathcal{P}^1$, where we divide each time period in problem $\mathcal{P}^1$ into k micro periods and multiply the capacity of each resource in problem $\mathcal{P}^1$ by k. Thus, the capacity of each resource and the total expected demand for each product in problem $\mathcal{P}^k$ are k times those in problem $\mathcal{P}^1$. Considering problem $\mathcal{P}^k$, letting $J^k_{\mathsf{App}}(k\boldsymbol{c})$ be the total expected revenue of the approximate policy and $J^k_1(k\boldsymbol{c})$ be the optimal total expected revenue, Theorem 5.1 implies that $\lim_{k\to\infty} \frac{J^k_{\mathsf{App}}(k\boldsymbol{c})}{J^k_1(k\boldsymbol{c})} = 1$. Nevertheless, Theorem 5.1, as it is stated, is more general because it does not require the expected demand to be scaled in any particular fashion.

6. Numerical Study

We give a numerical study to show the benefits from using the Universal Fluid approximation instead of the Traditional Fluid one when we have a random number of customer arrivals. In our experimental setup, we generate a number of test problems. For each test problem, we check the upper bound on the optimal total expected revenue from the two fluid approximations, as well as the performance of the policies driven by the two fluid approximations. Our test problems are on an airline network, where a resource corresponds to a flight leg and a product corresponds to an itinerary. There is one hub and six spokes. We have a flight that connects each spoke to the hub and the hub to each spoke, so the number of resources is 12. We have a high-fare and a low-fare itinerary that connect every origin-destination pair. Thus, the number of products is $2 \times 7 \times 6 = 84$. The itineraries that connect a spoke to the hub or the hub to a spoke are direct, including one flight leg, whereas the itineraries that connect a spoke to a spoke connect at the hub, including two flight legs. The number of customer arrivals is discretized and truncated log-normal with mean μ and coefficient of variation v. We vary $\mu \in \{400, 800, 1600, 3200\}$ and $v \in \{\frac{1}{128}, \frac{1}{64}, \frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1\}$. In Appendix C, we give the details of our approach for generating our test problems.

For each test problem, we solve the Universal Fluid and Traditional Fluid, as well as simulate the performance of the policies driven by the two approximations. Letting $\boldsymbol{y}^* = (y^*_{jt} : j \in \mathcal{J},\ t \in \mathcal{T})$ be

an optimal solution to the Universal Fluid or Traditional Fluid, the policy makes product j available at time period t with probability $\frac{y^*_{jt}}{\lambda_{jt}}$, which corresponds to using $\theta = 1$. Both policies had the best practical performance with this choice of θ. Our results are in Table 1. The first column gives the value of (v, μ) for each test problem. Letting $c_{\max} = \max_{i \in \mathcal{L}} c_i$, the second column gives the value of $(c_{\min}, c_{\max})$. The third, fourth and fifth columns focus on the Universal Fluid approximation and give the optimal objective value of the Universal Fluid, the total expected revenue of the policy from the Universal Fluid and the ratio between the performance of the policy and the optimal objective value of the fluid approximation. By the discussion just after Theorem 5.1, the entries of the fifth column approach one as $c_{\min}$ gets large. The sixth, seventh and eighth columns focus on the Traditional Fluid approximation and their interpretation is analogous to that of the third, fourth and fifth columns. The ninth column gives the percent gap between the optimal objective values of the two fluid approximations. The tenth column gives the percent gap between the performance of the two policies. The eleventh column gives the CPU seconds to solve the Universal Fluid on a 2.4 Ghz 8-core Intel i9 CPU with 64 GB of RAM using Java 19 and Gurobi 9.5.2. Letting $\overline{D}_p$ be the p-th percentile of the random variable D, the twelfth column gives the value of $(\overline{D}_5, \overline{D}_{95})$. Thus, the number of customer arrivals takes values in the interval $[\overline{D}_5, \overline{D}_{95}]$ with a probability of 0.9.

Considering the Universal Fluid approximation, our results in the table indicate that the ratio between the total expected revenue of its corresponding policy and the optimal objective value of the Universal Fluid approximation, as expected, gets close to one as $c_{\min}$ gets large. For the largest value of $c_{\min}$ with $c_{\min} = 173$, the ratio exceeds 0.96 in all of our test problems. On the other hand, considering the Traditional Fluid approximation, the ratio between the total expected revenue of its corresponding policy and the optimal objective value of the Traditional Fluid approximation does not necessarily get close to one as $c_{\min}$ gets large. For the test problems with a coefficient of variation of 1, the ratio is only 0.63 even when $c_{\min} = 173$. Thus, the phenomenon that we observed in the problem instance in Section 3 holds even for randomly generated test problems. Furthermore, the upper bound on the optimal total expected revenue provided by the Universal Fluid approximation can be substantially tighter than the one provided by the Traditional Fluid approximation. The upper bounds from the two fluid approximations can differ by as much as 36.26%. Getting the fluid approximation right also makes a noticeable impact on the performance of the corresponding policy. Total expected revenues obtained by the policies driven by the two fluid approximations can differ by as much as 13.38%, in favor of the Universal Fluid approximation. Overall, the Universal Fluid can provide significant improvements over the Traditional Fluid.

If the number of customer arrivals is fixed, so that $D = T$ with probability one, then the Traditional Fluid reduces to the standard fluid approximation under a Bernoulli process. It is known

Params.		Universal Fluid			Traditional Fluid			Obj.	Policy	CPU	
(v,μ)	$[c_{\min},c_{\max}]$	Obj.	Policy	Ratio	Obj.	Policy	Ratio	Gap	Gap	Secs.	$[\overline{D}_5,\overline{D}_{95}]$
$(\frac{1}{128},\ 400)$	[22, 44]	36,036	31,283	0.87	36,250	31,280	0.86	0.59%	0.01%	0.02	[394, 404]
$(\frac{1}{128},\ 800)$	[43, 88]	72,076	65,303	0.91	72,503	65,155	0.90	0.59%	0.23%	0.05	[789, 808]
$(\frac{1}{128},1600)$	[86,177]	143,886	134,532	0.93	144,807	134,237	0.93	0.64%	0.22%	0.11	[1578,1619]
$(\frac{1}{128},3200)$	[173,353]	287,983	275,310	0.96	289,763	274,495	0.95	0.62%	0.30%	0.26	[3158,3238]
$(\frac{1}{64},\ 400)$	[22, 44]	34,909	30,395	0.87	35,330	30,301	0.86	1.21%	0.31%	0.02	[389, 408]
$(\frac{1}{64},\ 800)$	[43, 88]	69,370	63,243	0.91	70,282	63,001	0.90	1.31%	0.38%	0.05	[778, 819]
$(\frac{1}{64},1600)$	[86,177]	139,262	130,934	0.94	141,015	130,417	0.92	1.26%	0.39%	0.11	[1558,1638]
$(\frac{1}{64},3200)$	[173,353]	278,520	268,012	0.96	282,027	267,023	0.95	1.26%	0.37%	0.51	[3117,3277]
$(\frac{1}{32},\ 400)$	[22, 44]	32,330	28,319	0.88	33,212	28,256	0.85	2.73%	0.22%	0.03	[378, 420]
$(\frac{1}{32},\ 800)$	[43, 88]	64,968	59,541	0.92	66,670	59,243	0.89	2.62%	0.50%	0.05	[758, 839]
$(\frac{1}{32},1600)$	[86,177]	130,010	123,358	0.95	133,413	122,494	0.92	2.62%	0.70%	0.33	[1518,1678]
$(\frac{1}{32},3200)$	[173,353]	260,015	251,379	0.97	266,822	249,569	0.94	2.62%	0.72%	0.81	[3037,3357]
$(\frac{1}{16},\ 400)$	[22, 44]	28,202	24,899	0.88	29,782	24,891	0.84	5.60%	0.03%	0.03	[359, 439]
$(\frac{1}{16},\ 800)$	[43, 88]	56,501	52,461	0.93	59,656	52,013	0.87	5.59%	0.85%	0.23	[719, 879]
$(\frac{1}{16},1600)$	[86,177]	112,979	107,413	0.95	119,291	106,317	0.89	5.58%	1.02%	0.58	[1440,1759]
$(\frac{1}{16},3200)$	[173,353]	226,049	219,567	0.97	238,678	217,010	0.91	5.59%	1.16%	1.33	[2880,3519]
$(\frac{1}{8},\ 400)$	[22, 44]	21,807	19,502	0.89	24,346	19,488	0.80	11.64%	0.08%	0.14	[322, 481]
$(\frac{1}{8},\ 800)$	[43, 88]	43,557	40,666	0.93	48,628	40,139	0.83	11.63%	1.30%	0.29	[645, 963]
$(\frac{1}{8},1600)$	[86,177]	87,240	83,480	0.96	97,391	81,629	0.84	11.62%	2.22%	0.63	[1292,1927]
$(\frac{1}{8},3200)$	[173,353]	174,479	169,217	0.97	194,782	164,295	0.84	11.64%	2.91%	1.43	[2585,3855]
$(\frac{1}{4},\ 400)$	[22, 44]	17,160	15,823	0.92	19,892	15,426	0.78	15.92%	2.51%	0.31	[258, 569]
$(\frac{1}{4},\ 800)$	[43, 88]	34,325	32,260	0.94	39,782	30,748	0.77	15.90%	4.69%	0.65	[516,1138]
$(\frac{1}{4},1600)$	[86,177]	68,695	66,032	0.96	79,637	62,495	0.78	15.93%	5.36%	1.43	[1033,2277]
$(\frac{1}{4},3200)$	[173,353]	137,377	132,793	0.97	159,256	123,493	0.78	15.93%	7.00%	3.25	[2067,4556]
$(\frac{1}{2},\ 400)$	[22, 44]	15,540	14,505	0.93	18,419	13,562	0.74	18.52%	6.50%	0.37	[163, 746]
$(\frac{1}{2},\ 800)$	[43, 88]	31,070	29,956	0.96	36,839	27,300	0.74	18.57%	8.87%	0.86	[327,1493]
$(\frac{1}{2},1600)$	[86,177]	62,168	59,743	0.96	73,747	53,657	0.73	18.63%	10.19%	2.14	[655,2988]
$(\frac{1}{2},3200)$	[173,353]	124,325	122,410	0.98	147,496	110,722	0.75	18.64%	9.55%	5.37	[1312,5977]
$(1,\ 400)$	[22, 44]	13,270	12,384	0.93	18,054	11,096	0.61	36.05%	10.40%	0.68	[71,1035]
$(1,\ 800)$	[43, 88]	26,523	25,092	0.95	36,105	21,846	0.61	36.13%	12.94%	1.51	[142,2070]
$(1,1600)$	[86,177]	53,057	51,176	0.96	72,282	44,331	0.61	36.24%	13.38%	3.45	[285,4142]
$(1,3200)$	[173,353]	106,098	105,015	0.99	144,564	90,975	0.63	36.26%	13.37%	9.42	[572,8285]

Table 1 Comparison of the two fluid approximations.

that this fluid approximation is asymptotically tight under a Bernoulli process as the resource capacities get large. Thus, we expect the Traditional Fluid approximation to provide tighter upper bounds and stronger policies when the coefficient of variation for the number of customer arrivals is smaller and the resource capacities are larger. Considering the test problems with the smallest of coefficient of variation of 1/128, when $c_{\min} = 173$, the ratio between the total expected revenue of the policy from the Traditional Fluid and the optimal objective value of the same fluid approximation is 0.95, so the performance of the policy is within 5% of the optimal total expected revenue. For the same problem instance, the corresponding ratio for the Universal Fluid is 0.96. Furthermore, if $D = T$ with probability one, then $\mathbb{P}\{D \geq t\} = 1$ for all $t = 1, \ldots, T$, in which case, the Traditional Fluid and Universal Fluid approximations become equivalent to each other. Thus, we expect the two fluid approximations to behave similarly when the coefficient of variation is small. For the test problems with the smallest coefficient of variation of 1/128, the optimal objective values of the two fluid approximations differ by at most 0.64% and the total expected revenues of the policies driven by the

two fluid approximations differ by at most 0.30%. Nevertheless, when the coefficient of variation of the number of customer arrivals is 1/8 or more, we observe significant gaps in the optimal objective values of the two fluid approximations and the performance of the policies. For the test problems with a coefficient of variation of 1/8, the upper bounds provided by the Universal Fluid improve those provided by the Traditional Fluid by as much as 11.64%. The total expected revenues of the policies driven by the two fluid approximations differ by as much as 2.91%.

7. Conclusions

We made three contributions. First, the form of the fluid approximation we propose under a random number of customer arrivals is somewhat unexpected because the distribution of the number of customer arrivals does not appear in the capacity constraints. A naive fluid approximation that uses the expected capacity consumption of the resources on the left side of the capacity constraints is not asymptotically tight. Second, our work shows that we can formulate asymptotically tight fluid approximations when the number of customer arrivals has arbitrary distributions. The fact that the coefficient of variation of the demand under the Bernoulli arrival process gets smaller as the mean demand gets larger is not a requirement to formulate asymptotically tight fluid approximations. Third, getting the fluid approximation right is practically important. The policy driven by the right fluid approximation can perform significantly better.

Working with richer customer arrival processes is an interesting research area. In our model, we have a random number of customer arrivals with a finite upper bound. This finite upper bound is not a huge practical concern, but our fluid approximation can also work with number of customer arrivals without a finite upper bound. In Appendix D, we give one possible approach to address the case where no such finite upper bound is available. There is a host of other approximation strategies, beside fluid approximations, for large-scale revenue management problems. It would be interesting to study whether they can be extended to incorporate high-variance demand. As discussed in the introduction, there is also work on improving the performance of the policy by periodically solving the fluid approximation. It would be useful to explore the analogues of these results under random number of customer arrivals.

References

Adelman, D., A. J. Mersereau. 2008. Relaxations of weakly coupled stochastic dynamic programs. *Operations Research* **56**(3) 712–727.

Aouad, A., W. Ma. 2022. A nonparametric framework for online stochastic matching with correlated arrivals. Tech. rep., London Business School, London, UK.

Bai, Y., O. El Housni, P. Rusmevichientong, H. Topaloglu. 2022. Coordinated inventory stocking and assortment personalization. Tech. rep., Cornell University, New York, NY.

Balseiro, S. R., O. Besbes, D. Pizarro. 2021. Survey of dynamic resource constrained reward collection problems: Unified model and analysis. Tech. rep., Columbia University, New York, NY.

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

Cooper, W. L. 2002. Asymptotic behavior of an allocation policy for revenue management. *Operations Research* **50**(4) 720–727.

Feng, Y., R. Niazadeh, A. Saberi. 2022. Near-optimal Bayesian online assortment of reusable resources. Tech. rep., University of Chicago, Chicago, IL.

Gallego, G., G. Iyengar, R. Phillips, A. Dubey. 2004. Managing flexible products on a network. CORC Technical Report TR-2004-01.

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

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

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

Kaggle. 2021. E-commerce purchase history from an electronics store. Last checked: May 30, 2022. URL `https://www.kaggle.com/datasets/mkechinov/ecommerce-purchase-history-from-electronics-store`.

Liu, Q., G. J. van Ryzin. 2008. On the choice-based linear programming model for network revenue management. *Manufacturing & Service Operations Management* **10**(2) 288–310.

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. 1998. An analysis of bid-price controls for network revenue management. *Management Science* **44**(11) 1577–1593.

Talluri, K. T., G. J. van Ryzin. 2005. *The Theory and Practice of Revenue Management*. Kluwer Academic Publishers, Boston, MA.

Walczak, D. 2006. Modeling high demand variance in dynamic programming. *Journal of Revenue and Pricing Management* **5**(2) 94–101.

Online Supplement
Fluid Approximations for Revenue Management under High-Variance Demand

Appendix A: An Upper Bound on the Optimal Total Expected Revenue

In this section, we give a proof for Theorem 4.1. We will use an equivalent reformulation of the dynamic program in Section 2 that is more suitable for Lagrangian relaxation. In our equivalent reformulation, we use the decision variables $\boldsymbol{u} = (u_j : j \in \mathcal{J}) \in \{0,1\}^{|\mathcal{J}|}$, where $u_j = 1$ if and only if we make product j available at a generic time period. If the remaining capacities of the resources are given by the vector $\boldsymbol{x} \in \mathbb{Z}_+^{|\mathcal{L}|}$, then the set of feasible decisions is given by $\mathcal{U}(\boldsymbol{x}) = \{\boldsymbol{u} \in \{0,1\}^{|\mathcal{J}|} : \mathbf{1}_{(i \in A_j)}\, u_j \leq x_i \;\forall\, i \in \mathcal{L},\; j \in \mathcal{J}\}$, which ensures that we can make product j available only when there is at least one unit of remaining capacity for all resources that are used by product j. In this case, the dynamic program in Section 2 is equivalent to

$$J_t(\boldsymbol{x}) = \max_{\boldsymbol{u} \in \mathcal{U}(\boldsymbol{x})} \left\{ \sum_{j \in \mathcal{J}} \lambda_{jt} \left\{ f_j\, u_j + \rho_t\, J_{t+1}\Big(\boldsymbol{x} - u_j \sum_{i \in A_j} \boldsymbol{e}_i\Big) \right\} \right\}, \tag{1}$$

with the boundary condition that $J_{T+1} = 0$. The value functions computed through the dynamic program in (1) are identical to those computed through the dynamic program in Section 2.

Considering (1), for each $i \in \mathcal{L}$ and $j \in \mathcal{J}$, we relax the constraint $\mathbf{1}_{(i \in A_j)}\, u_j \leq x_i$ at time period t by associating the Lagrange multiplier θ_{ijt} with it to obtain the dynamic program

$$\begin{aligned}
\hat{J}_t^{\boldsymbol{\theta}}(\boldsymbol{x}) &= \max_{\boldsymbol{u} \in \{0,1\}^{|\mathcal{J}|}} \left\{ \sum_{j \in \mathcal{J}} \lambda_{jt} \left\{ f_j\, u_j + \rho_t\, \hat{J}_{t+1}^{\boldsymbol{\theta}}\Big(\boldsymbol{x} - u_j \sum_{i \in A_j} \boldsymbol{e}_i\Big) \right\} + \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{J}} \lambda_{jt}\, \theta_{ijt} \Big[x_i - \mathbf{1}_{(i \in A_j)}\, u_j \Big] \right\} \\
&= \max_{\boldsymbol{u} \in \{0,1\}^{|\mathcal{J}|}} \left\{ \sum_{j \in \mathcal{J}} \lambda_{jt} \left\{ \Big[f_j - \sum_{i \in \mathcal{L}} \mathbf{1}_{(i \in A_j)}\, \theta_{ijt} \Big] u_j + \rho_t\, \hat{J}_{t+1}^{\boldsymbol{\theta}}\Big(\boldsymbol{x} - u_j \sum_{i \in A_j} \boldsymbol{e}_i\Big) \right\} \right\} + \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{J}} \lambda_{jt}\, \theta_{ijt}\, x_i, \quad (2)
\end{aligned}$$

with the boundary condition that $\hat{J}_{T+1}^{\boldsymbol{\theta}} = 0$. In the first equality above, we scaled the Lagrange multiplier θ_{ijt} with λ_{jt}, which will simplify our notation. The second equality follows by arranging the terms. We refer to the dynamic program in (2) as the relaxed dynamic program. In the relaxed dynamic program, we make it explicit that the value functions depend on our choice of the Lagrange multipliers $\boldsymbol{\theta} = (\theta_{ijt} : i \in \mathcal{L},\; j \in \mathcal{J},\; t \in \mathcal{T}) \in \mathbb{R}_+^{|\mathcal{L}||\mathcal{J}||\mathcal{T}|}$. It is a standard result that the value functions of the relaxed dynamic program provide upper bounds on the value functions in (1); see, for example, Adelman and Mersereau (2008). Thus, for any choice of the Lagrange multipliers $\boldsymbol{\theta} \in \mathbb{R}_+^{|\mathcal{L}||\mathcal{J}||\mathcal{T}|}$, we have $\hat{J}_t^{\boldsymbol{\theta}}(\boldsymbol{x}) \geq J_t(\boldsymbol{x})$ for all $t \in \mathcal{T}$.

Noting that $J_1(\boldsymbol{c})$ is the optimal total expected revenue, by the discussion in the previous paragraph, for any $\boldsymbol{\theta} \in \mathbb{R}_+^{|\mathcal{L}||\mathcal{J}||\mathcal{T}|}$, $\hat{J}_1^{\boldsymbol{\theta}}(\boldsymbol{c})$ is an upper bound on the optimal total expected revenue. To

show Theorem 4.1, we focus on a specific choice of the Lagrange multipliers, where the Lagrange multipliers at all time periods except for the last one are zero, whereas the Lagrange multipliers at the last time period depend on the resources but not on the products. For any $\boldsymbol{\eta} = (\eta_i : i \in \mathcal{L}) \in \mathbb{R}_+^{|\mathcal{L}|}$, define $\mathcal{F}(\boldsymbol{\eta}) = \{\boldsymbol{\theta} \in \mathbb{R}_+^{|\mathcal{L}||\mathcal{J}||\mathcal{T}|} : \theta_{ijT} = \eta_i \ \forall i \in \mathcal{L},\ j \in \mathcal{J},\ \theta_{ijt} = 0 \ \forall i \in \mathcal{L},\ j \in \mathcal{J},\ t \in \mathcal{T} \setminus \{T\}\}$, so if $\boldsymbol{\theta} \in \mathcal{F}(\boldsymbol{\eta})$, then the Lagrange multipliers $(\theta_{ijT} : j \in \mathcal{J})$ take the common value of η_i, but the Lagrange multipliers $(\theta_{ijt} : i \in \mathcal{L}\ , j \in \mathcal{J},\ t \in \mathcal{T} \setminus \{T\})$ take the value of zero. Focusing on a specific choice of the Lagrange multipliers does not take full advantage of the flexibility provided by the possibility of using a different Lagrange multiplier for each resource, product and time period, but it will be enough to show that the optimal objective value of the Universal Fluid approximation is an upper bound on the optimal total expected revenue. In the next lemma, we give a closed form expression for the value functions $\{\hat{J}_t^{\boldsymbol{\theta}} : t \in \mathcal{T}\}$ when we have $\boldsymbol{\theta} \in \mathcal{F}(\boldsymbol{\eta})$.

Lemma A.1 *Letting $R_t = \rho_t \, \rho_{t+1} \ldots \rho_{T-1}$ with $R_T = 1$, for any $\boldsymbol{\eta} \in \mathbb{R}_+^{|\mathcal{L}|}$, if the Lagrange multipliers satisfy $\boldsymbol{\theta} \in \mathcal{F}(\boldsymbol{\eta})$, then we have*

$$\hat{J}_t^{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=t}^{T} \sum_{j \in \mathcal{J}} \lambda_{jk} \Big[\frac{R_t}{R_k} f_j - R_t \sum_{i \in A_j} \eta_i \Big]^+ + R_t \sum_{i \in \mathcal{L}} \eta_i \, x_i.$$

Proof: We show the result by using induction over the time periods. At time period T, noting that $\hat{J}_{T+1}^{\boldsymbol{\theta}} = 0$ and $\theta_{ijT} = \eta_i$, as well as using the fact that $\sum_{j \in \mathcal{J}} \lambda_{jT} = 1$, by (2), we have

$$\hat{J}_T^{\boldsymbol{\theta}}(\boldsymbol{x}) = \max_{\boldsymbol{u} \in \{0,1\}^{|\mathcal{J}|}} \left\{ \sum_{j \in \mathcal{J}} \lambda_{jT} \Big[f_j - \sum_{i \in \mathcal{L}} \mathbf{1}_{(i \in A_j)} \, \eta_i \Big] u_j \right\} + \sum_{i \in \mathcal{L}} \sum_{j \in \mathcal{J}} \lambda_{jT} \, \eta_i \, x_i = \sum_{j \in \mathcal{J}} \lambda_{jT} \Big[f_j - \sum_{i \in A_j} \eta_i \Big]^+ + \sum_{i \in \mathcal{L}} \eta_i \, x_i.$$

Thus, the result holds at time period T. Assuming that the result holds at time period $t+1$, we show that the result holds at time period t as well.

Letting $K_t = \sum_{k=t}^{T} \sum_{j \in \mathcal{J}} \lambda_{jk} \, [\frac{R_t}{R_k} f_j - R_t \sum_{i \in A_j} \eta_i]^+$, by the induction assumption, we have $\hat{J}_{t+1}^{\boldsymbol{\theta}}(\boldsymbol{x}) = K_{t+1} + R_{t+1} \sum_{i \in \mathcal{L}} \eta_i \, x_i$. Noting that $\theta_{ijt} = 0$ for $t \in \mathcal{T} \setminus \{T\}$, by (2), we have

$$\begin{aligned}
\hat{J}_t^{\boldsymbol{\theta}}(\boldsymbol{x}) &\overset{(a)}{=} \max_{\boldsymbol{u} \in \{0,1\}^{|\mathcal{J}|}} \left\{ \sum_{j \in \mathcal{J}} \lambda_{jt} \left\{ f_j \, u_j + \rho_t \, K_{t+1} + \rho_t \, R_{t+1} \sum_{i \in \mathcal{L}} \eta_i \, (x_i - \mathbf{1}_{(i \in A_j)} \, u_j) \right\} \right\} \\
&\overset{(b)}{=} \max_{\boldsymbol{u} \in \{0,1\}^{|\mathcal{J}|}} \left\{ \sum_{j \in \mathcal{J}} \lambda_{jt} \Big[f_j - \rho_t \, R_{t+1} \sum_{i \in A_j} \eta_i \Big] u_j \right\} + \rho_t \, K_{t+1} + \rho_t \, R_{t+1} \sum_{i \in \mathcal{L}} \eta_i \, x_i \\
&= \sum_{j \in \mathcal{J}} \lambda_{jt} \Big[f_j - \rho_t \, R_{t+1} \sum_{i \in A_j} \eta_i \Big]^+ + \rho_t \, K_{t+1} + \rho_t \, R_{t+1} \sum_{i \in \mathcal{L}} \eta_i \, x_i \\
&\overset{(c)}{=} K_t + R_t \sum_{i \in \mathcal{L}} \eta_i \, x_i,
\end{aligned}$$

where (a) holds because we have $\hat{J}_{t+1}^{\boldsymbol{\theta}}(\boldsymbol{x} - u_j \sum_{i \in A_j} \boldsymbol{e}_i) = K_{t+1} + R_{t+1} \sum_{i \in \mathcal{L}} \eta_i \, (x_i - \mathbf{1}_{(i \in A_j)} \, u_j)$ by the induction assumption, (b) follows by arranging the terms and using the fact that $\sum_{j \in \mathcal{J}} \lambda_{jt} = 1$

and (c) holds because we have $R_t = \rho_t\, R_{t+1}$ and $K_t = \sum_{j\in\mathcal{J}} \lambda_{jt}\,[f_j - R_t \sum_{i\in A_j} \theta_i]^+ + \rho_t\, K_{t+1}$ by the definitions of R_t and K_t. By the chain of equalities above, the result holds at time period t. ■

We have $R_t = \rho_t\, \rho_{t+1} \cdots \rho_{T-1} = \frac{\mathbb{P}\{D\geq t+1\}}{\mathbb{P}\{D\geq t\}}\, \frac{\mathbb{P}\{D\geq t+2\}}{\mathbb{P}\{D\geq t+1\}} \cdots \frac{\mathbb{P}\{D\geq T\}}{\mathbb{P}\{D\geq T-1\}} = \frac{\mathbb{P}\{D\geq T\}}{\mathbb{P}\{D\geq t\}}$. Since $\mathbb{P}\{D\geq 1\} = 1$, the last equality also yields $R_1 = \mathbb{P}\{D\geq T\}$. By Lemma A.1, for any $\boldsymbol{\eta}\in\mathbb{R}_+^{|\mathcal{L}|}$ and $\boldsymbol{\theta}\in\mathcal{F}(\boldsymbol{\eta})$, we get

$$\begin{aligned}\hat{J}_1^{\boldsymbol{\theta}}(\boldsymbol{c}) \;&=\; \sum_{t=1}^{T}\sum_{j\in\mathcal{J}} \lambda_{jt}\left[\frac{R_1}{R_t} f_j - R_1 \sum_{i\in A_j}\eta_i\right]^+ + R_1 \sum_{i\in\mathcal{L}} \eta_i\, c_i \\ &=\; \sum_{t=1}^{T}\sum_{j\in\mathcal{J}} \lambda_{jt}\left[\mathbb{P}\{D\geq t\}\, f_j - \mathbb{P}\{D\geq T\}\sum_{i\in\mathcal{L}} \mathbf{1}_{(i\in A_j)}\,\eta_i\right]^+ + \mathbb{P}\{D\geq T\}\sum_{i\in\mathcal{L}}\eta_i\, c_i. \quad (3)\end{aligned}$$

We have $\hat{J}_1^{\boldsymbol{\theta}}(\boldsymbol{c}) \geq J_1(\boldsymbol{c})$ for all $\boldsymbol{\theta}\in\mathbb{R}_+^{|\mathcal{L}||\mathcal{J}||\mathcal{T}|}$, but we have $\boldsymbol{\theta}\in\mathbb{R}_+^{|\mathcal{L}||\mathcal{J}||\mathcal{T}|}$ for any $\boldsymbol{\eta}\in\mathbb{R}_+^{|\mathcal{L}|}$ and $\boldsymbol{\theta}\in\mathcal{F}(\boldsymbol{\eta})$. Thus, the expression on the right side of (3) is an upper bound on $J_1(\boldsymbol{c})$ for any $\boldsymbol{\eta}\in\mathbb{R}_+^{|\mathcal{L}|}$.

By the discussion in the previous paragraph, if we minimize the expression on the right side of (3) over all $\boldsymbol{\eta}\in\mathbb{R}_+^{|\mathcal{L}|}$, then we get an upper bound on $J_1(\boldsymbol{c})$. Below is the proof of Theorem 4.1.

Proof of Theorem 4.1:

Using the decision variables $\boldsymbol{\eta} = (\eta_i : i\in\mathcal{L})\in\mathbb{R}_+^{|\mathcal{L}|}$ and $\boldsymbol{z} = (z_{jt} : j\in\mathcal{J},\ t\in\mathcal{T})\in\mathbb{R}_+^{|\mathcal{J}||\mathcal{T}|}$, we can minimize the expression on the right side of (3) over all $\boldsymbol{\eta}\in\mathbb{R}_+^{|\mathcal{L}|}$ by solving the linear program

$$\min_{(\boldsymbol{\eta},\boldsymbol{z})\in\mathbb{R}_+^{|\mathcal{L}|+|\mathcal{J}||\mathcal{T}|}} \left\{ \sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}} \lambda_{jt}\, z_{jt} + \mathbb{P}\{D\geq T\}\sum_{i\in\mathcal{L}} c_i\,\eta_i \;:\; z_{jt} \;\geq\; \mathbb{P}\{D\geq t\}\, f_j \,-\, \mathbb{P}\{D\geq T\}\sum_{i\in\mathcal{L}}\mathbf{1}_{(i\in A_j)}\,\eta_i \;\;\forall\, j\in\mathcal{J},\ t\in\mathcal{T}\right\}. \quad (4)$$

We can assume that $\mathbb{P}\{D\geq T\} > 0$, because if $\mathbb{P}\{D\geq T\} = 0$, then we can choose the upper bound of the support of D as the largest value of $\tau\in\{1,\ldots,T-1\}$ such that $\mathbb{P}\{D\geq\tau\} > 0$.

Associating the dual variables $\boldsymbol{y} = (y_{jt} : j\in\mathcal{J},\ t\in\mathcal{T})\in\mathbb{R}_+^{|\mathcal{J}||\mathcal{T}|}$ with the constraints in the problem above, the dual of problem (4) is given by

$$\max_{\boldsymbol{y}\in\mathbb{R}_+^{|\mathcal{J}||\mathcal{T}|}} \left\{ \sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}} f_j\,\mathbb{P}\{D\geq t\}\, y_{jt} \;:\; \mathbb{P}\{D\geq T\}\sum_{t\in\mathcal{T}}\sum_{j\in\mathcal{J}}\mathbf{1}_{(i\in A_j)}\, y_{jt} \;\leq\; \mathbb{P}\{D\geq T\}\, c_i \;\;\forall\, i\in\mathcal{L} \right.$$
$$\left. y_{jt} \;\leq\; \lambda_{jt} \;\;\forall\, j\in\mathcal{J},\ t\in\mathcal{T}\right\}. \quad (5)$$

Because $\mathbb{P}\{D\geq T\} > 0$, problem (5) is equivalent to the Universal Fluid. The desired result follows since (4) and (5) have the same optimal objective value, which is an upper bound on $J_1(\boldsymbol{c})$. ■

In our derivation of the Universal Fluid approximation, we used the relaxed dynamic program with a specific choice of the Lagrange multipliers, where the Lagrange multipliers at all time periods

except for the last one are zero, whereas the Lagrange multipliers at the last time period depend on the resources but not on the products. We can potentially derive other fluid approximations by using more general Lagrange multipliers that exploit the possibility that we can use a different Lagrange multiplier for each resource, product and time period. The upper bounds on the optimal total expected revenue provided by such fluid approximations would be at least as tight as those provided by the Universal Fluid approximation, but such fluid approximations do not have a form that is as simple as the Universal Fluid approximation. Furthermore, the Universal Fluid approximation turns out to be enough to obtain asymptotically tight upper bounds.

Appendix B: Asymptotic Tightness of the Fluid Approximation

We give a proof for Theorem 5.1. Let $\boldsymbol{y}^* = (y^*_{jt} : j \in \mathcal{J},\ t \in \mathcal{T})$ be an optimal solution to the Universal Fluid approximation. Consider the following approximate policy for some $\theta \in (0,1)$. At time period t, we make product j available for purchase with probability $\theta \frac{y^*_{jt}}{\lambda_{jt}}$. If the customer arriving at time period t wants to purchase product j and there is capacity available to serve a request for product j, then we sell a unit of product j and consume the capacities of the resources used by the product. Define three Bernoulli random variables. The first one, denoted by A_{jt}, takes value one if the approximate policy makes product j available at time period t. We have $\mathbb{P}\{A_{jt}=1\} = \theta \frac{y^*_{jt}}{\lambda_{jt}}$. The second one, denoted by Λ_{jt}, takes value one if the customer arriving at time period t is interested in purchasing product j. We have $\mathbb{P}\{\Lambda_{jt}=1\} = \lambda_{jt}$. The third one, denoted by G_{jt}, takes value one if we have capacity to serve a request for product j at time period t under the approximate policy. In this case, the total revenue obtained by the approximate policy is given by the random variable $\sum_{t=1}^{T} \sum_{j\in\mathcal{J}} f_j \mathbf{1}_{(D\geq t,\, A_{jt}=1,\, \Lambda_{jt}=1,\, G_{jt}=1)}$, where we use the fact that the approximate policy makes a sale for product j at time period t if and only if the selling horizon reaches beyond this time period, the approximate policy makes product j available, the arriving customer is interested in purchasing product j and we have capacity to serve a request for product j. Note that G_{jt} depends on the decisions of the approximate policy at time periods $1,\ldots,t-1$ and D is independent of the decisions of the approximate policy and the products of interest to the arriving customers. Thus, letting $J_{\mathsf{App}}(\boldsymbol{c})$ be the total expected revenue of the approximate policy, we get

$$\begin{aligned} J_{\mathsf{App}}(\boldsymbol{c}) &= \sum_{t=1}^{T} \sum_{j\in\mathcal{J}} f_j\, \mathbb{P}\{D\geq t\}\, \mathbb{P}\{A_{jt}=1\}\, \mathbb{P}\{\Lambda_{jt}=1\}\, \mathbb{P}\{G_{jt}=1\} \\ &= \sum_{t=1}^{T} \sum_{j\in\mathcal{J}} f_j\, \mathbb{P}\{D\geq t\}\, \theta\, \frac{y^*_{jt}}{\lambda_{jt}}\, \lambda_{jt}\, \mathbb{P}\{G_{jt}=1\}. \quad (6) \end{aligned}$$

We lower bound the probability $\mathbb{P}\{G_{jt}=1\}$. At time period t, the approximate policy makes product j available with probability $\theta \frac{y^*_{jt}}{\lambda_{jt}}$, whereas we have a request for product j with

probability λ_{jt}. Thus, under the approximate policy, there is a unit of demand for capacity of resource i at time period t with probability $\sum_{j\in\mathcal{J}} \mathbf{1}_{(i\in A_j)}\, \theta\, \frac{y^*_{jt}}{\lambda_{jt}}\, \lambda_{jt} = \theta \sum_{j\in\mathcal{J}} \mathbf{1}_{(i\in A_j)}\, y^*_{jt}$. However, having demand for capacity of resource i at time period t does not mean that the approximate policy depletes the capacity of the resource at time period t. In particular, considering some product j that uses the capacity of resource i, even if the approximate policy makes product j available at time period t and the customer arriving at time period t is interested in product j, we may not have capacity on some other resource used by product j, in which case, we would not be serving the demand for the product. Thus, letting $\{N_{it} : t \in \mathcal{T}\}$ be a collection of independent Bernoulli random variables, where N_{it} takes value one with probability $\theta \sum_{j\in\mathcal{J}} \mathbf{1}_{(i\in A_j)}\, y^*_{jt}$, under the approximate policy, the total capacity consumption of resource i over time periods $1,\ldots,t-1$ is upper bounded by $\sum_{k=1}^{T} N_{ik}$. Thus, having $\sum_{k=1}^{T} N_{ik} < c_i$ for all $i \in A_j$ implies that $G_{jt} = 1$, so $\mathbb{P}\{\sum_{k=1}^{T} N_{ik} < c_i \;\; \forall\, i \in A_j\} \leq \mathbb{P}\{G_{jt} = 1\}$. We need the concentration bound in the next lemma.

Lemma B.1 *Letting $\{N_{it} : t \in \mathcal{T}\}$ be a collection of independent Bernoulli random variables, where N_{it} takes value one with probability $\theta \sum_{j\in\mathcal{J}} \mathbf{1}_{(i\in A_j)}\, y^*_{jt}$, we have*

$$\mathbb{P}\left\{\sum_{t=1}^{T} N_{it} \geq c_i\right\} \;\leq\; \exp\left(-\frac{\frac{3}{2}\,(1-\theta)^2 c_{\min}}{2\,\theta+1}\right).$$

Proof: Letting $\rho_{it} = \theta \sum_{j\in\mathcal{J}} \mathbf{1}_{(i\in A_j)}\, y^*_{jt}$ for notational brevity, so that we have $\mathbb{E}\{N_{it}\} = \rho_{it}$ and $\mathrm{Var}(N_{it}) = \rho_{it}\,(1-\rho_{it})$, we upper bound the expectation and variance of $\sum_{t=1}^{T} N_{it}$ as

$$\mathrm{Var}\left(\sum_{t=1}^{T} N_{it}\right) = \sum_{t=1}^{T} \rho_{it}\,(1-\rho_{it}) \leq \sum_{t=1}^{T} \rho_{it} = \mathbb{E}\left\{\sum_{t=1}^{T} N_{it}\right\} = \theta \sum_{t=1}^{T} \sum_{j\in\mathcal{J}} \mathbf{1}_{(i\in A_j)}\, y^*_{jt} \leq \theta\, c_i,$$

where the last inequality holds because $\boldsymbol{y}^* = (y^*_{jt} : j \in \mathcal{J},\; t \in \mathcal{T})$ is an optimal solution to the Universal Fluid approximation, so it satisfies the first constraint in the fluid approximation.

Noting that $\mathbb{E}\{\sum_{t=1}^{T} N_{it}\} \leq \theta\, c_i$ by the chain of inequalities above, using the one-sided Bernstein inequality, we obtain the chain of inequalities

$$\begin{aligned}\mathbb{P}\left\{\sum_{t=1}^{T} N_{it} \geq c_i\right\} &\overset{(a)}{\leq} \mathbb{P}\left\{\sum_{t=1}^{T}\Big[N_{it} - \mathbb{E}\{N_{it}\}\Big] \geq (1-\theta)\, c_i\right\} \overset{(b)}{\leq} \exp\left(-\frac{\frac{1}{2}(1-\theta)^2\, c_i^2}{\sum_{t=1}^{T}\mathrm{Var}(N_{it}) + \frac{1}{3}(1-\theta)c_i}\right)\\ &\overset{(c)}{\leq} \exp\left(-\frac{\frac{1}{2}(1-\theta)^2\, c_i^2}{\theta\, c_i + \frac{1}{3}(1-\theta)c_i}\right) = \exp\left(-\frac{\frac{3}{2}(1-\theta)^2\, c_i}{2\,\theta+1}\right) \overset{(d)}{\leq} \exp\left(-\frac{\frac{3}{2}(1-\theta)^2\, c_{\min}}{2\,\theta+1}\right),\end{aligned}$$

where (a) holds because $\mathbb{E}\{\sum_{t=1}^{T} N_{it}\} \leq \theta\, c_i$, (b) is the one-sided Bernstein inequality, (c) uses the fact that $\sum_{t=1}^{T} \mathrm{Var}(N_{it}) \leq \theta\, c_i$ and (d) uses the fact that $c_{\min} \leq c_i$. ■

We can use the bound in the lemma above along with the union bound to come up with a lower bound on the probability $\mathbb{P}\{\sum_{t=1}^{T} N_{it} < c_i \;\; \forall\, i \in A_j\}$. By the discussion right before the lemma, a

lower bound on the last probability is also a lower bound on the probability $\mathbb{P}\{G_{jt}=1\}$. Putting these observations together will yield a proof for Theorem 5.1.

Proof of Theorem 5.1:

Noting the discussion just before Lemma B.1, $\mathbb{P}\{\sum_{k=1}^{T} N_{ik} < c_i \ \forall i \in A_j\} \le \mathbb{P}\{G_{jt}=1\}$. We lower bound the probability $\mathbb{P}\{G_{jt}=1\}$ as

$$\begin{aligned}\mathbb{P}\{G_{jt}=1\} &\ge \mathbb{P}\left\{\sum_{t=1}^{T} N_{it} < c_i \ \forall i \in A_j\right\} = 1-\mathbb{P}\left\{\exists i \in A_j \text{ such that } \sum_{t=1}^{T} N_{it} \ge c_i\right\}\\ &\overset{(a)}{\ge} 1-\sum_{i\in A_j}\mathbb{P}\left\{\sum_{t=1}^{T} N_{it} \ge c_i\right\} \overset{(b)}{\ge} 1-L\exp\left(-\frac{\frac{3}{2}(1-\theta)^2 c_{\min}}{2\theta+1}\right) \overset{(c)}{\ge} 1-L\exp\left(-\frac{(1-\theta)^2 c_{\min}}{2}\right),\end{aligned}$$

where (a) is the union bound, (b) follows from Lemma B.1, as well as the fact that $|A_j| \le L$ and (c) uses the fact that $\theta \in (0,1)$, in which case, we have $2\theta+1<3$.

If we use $\theta = 1-\sqrt{\frac{2\log c_{\min}}{c_{\min}}}$ in our approximate policy, then the right of the chain of inequalities above reads $1-\frac{L}{c_{\min}}$, so $\mathbb{P}\{G_{jt}=1\} \ge 1-\frac{L}{c_{\min}}$ with this choice of θ. Thus, by (6), we get

$$\begin{aligned}J_{\mathsf{App}}(\boldsymbol{c}) \ &\ge \ \left(1-\sqrt{\frac{2\log c_{\min}}{c_{\min}}}\right)\sum_{t=1}^{T}\sum_{j\in\mathcal{J}} f_j\,\mathbb{P}\{D\ge t\}\,y^*_{jt}\left(1-\frac{L}{c_{\min}}\right)\\ &\overset{(d)}{=} \left(1-\sqrt{\frac{2\log c_{\min}}{c_{\min}}}\right)\left(1-\frac{L}{c_{\min}}\right)Z^*_{\mathsf{LP}} \ \ge \ \left(1-\sqrt{\frac{2\log c_{\min}}{c_{\min}}}-\frac{L}{c_{\min}}\right)Z^*_{\mathsf{LP}},\end{aligned}$$

where (d) holds because the solution $\boldsymbol{y}^* = (y^*_{jt} : j\in\mathcal{J},\ t\in\mathcal{T})$ is optimal to the Universal Fluid. The desired result follows because the optimal total expected revenue satisfies $J_1(\boldsymbol{c}) \ge J_{\mathsf{App}}(\boldsymbol{c})$. ■

Appendix C: Experimental Setup for the Test Problems

We give the details of our approach for generating our test problems. Letting Γ be a log-normal random variable with mean μ and standard deviation μv and k be the smallest integer such that $\mathbb{P}\{\Gamma \le k\} \ge 0.99$, we set the maximum length of the selling horizon as $T=k$. For each $t=1,\ldots,T$, letting $\gamma_t = \mathbb{P}\{t-1 \le \Gamma \le t\}$, the probability mass function of D evaluated at t is proportional to γ_t. In particular, for each $t=1,\ldots,T$, we set $\mathbb{P}\{D=t\} = \gamma_t / \sum_{s=1}^{T}\gamma_s$. We place the hub at the center of a 100×100 square and generate the locations of the spokes uniformly over the same square. The fare associated with a low-fare itinerary is the sum of the Euclidean distances traversed by the flights in the itinerary. The fare associated with a high-fare itinerary is κ times the fare of the corresponding low-fare itinerary. We set $\kappa = 4$.

To come up with the arrival probabilities for the customers interested in different itineraries, for each origin-destination pair (f,g), we generate ξ_{fg} from the uniform distribution over $[0,1]$.

One of the locations in the origin-destination pair can be the hub. Letting N be the set of all locations, we normalize these samples by setting $\zeta_{fg} = \xi_{fg} / \sum_{(p,q)\in N^2, p\neq q} \xi_{pq}$ so that they add up to one. The probability that the customer arriving at any time period is interested in an itinerary that connects the origin-destination pair (f,g) is ζ_{fg}. The probability that a customer is interested in a low-fare itinerary decreases over time, whereas we have the reverse trend for the probability that a customer is interested in a high-fare itinerary. In this way, we generate test problems where the requests for high-fare itineraries tend to arrive later and we need to protect capacity for the high-fare itinerary requests that tend to arrive later. To generate test problems with this feature, for each origin-destination pair (f,g), we generate a time threshold τ_{fg} uniformly over $\{1,\ldots,T\}$. The probability of having a request for a low-fare itinerary linearly decreases over time, whereas the probability of having a request for a high-fare itinerary is zero until time period τ_{fg}, but it increases linearly after time period τ_{fg}. In particular, we define the functions $G, H_{fg} : \mathcal{T} \to \mathbb{R}_+$ as $G(t) = 1 - \frac{t-1}{T-1}$ and $H_{fg}(t) = \left[\frac{t-\tau_{fg}}{T-\tau_{fg}}\right]^+$. In this case, if itinerary j is the low-fare itinerary connecting origin-destination pair (f,g), then $\lambda_{jt} = \zeta_{fg} \frac{G(t)}{G(t)+H_{fg}(t)}$ and if itinerary j is the high-fare itinerary connecting origin-destination pair (f,g), then $\lambda_{jt} = \zeta_{fg} \frac{H_{fg}(t)}{G(t)+H_{fg}(t)}$. Once we generate the customer arrival probabilities, we set the capacities of the flight legs so that the total expected demand for the capacity on the flight leg exceeds the capacity of the flight leg by a factor of 1.6. In other words, noting that the total expected demand for the capacity on flight leg i is $\sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}} \mathbf{1}_{(i\in A_j)} \lambda_{jt}$, the capacity of flight leg i is $c_i = \lceil \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}} \mathbf{1}_{(i\in A_j)} \lambda_{jt} / 1.6 \rceil$.

We can choose the coefficient of variation of a log-normal random variable as large as we would like, which was the motivation for using this distribution for D in our experimental setup.

Appendix D: Finite Upper Bound on the Number of Customer Arrivals

We start by considering the case where there exists some $\overline{\lambda} > 0$ such that $\lambda_{jt} \geq \overline{\lambda}$ for all $j \in \mathcal{J}$ and $t \in \mathcal{T}$. Thus, the probability that an arriving customer is interested in a particular product is uniformly lower bounded by $\overline{\lambda}$. We discuss relaxations of this setup at the end of this section. Making the dependence of the Universal Fluid on the set of possible values for the length of the selling horizon explicit, we write the optimal objective value of this problem as $Z^*_{\mathsf{LP}}(\mathcal{T})$. Define the time threshold $\tau = \lceil \max_{i\in\mathcal{L}} c_i / \overline{\lambda} \rceil$. In the next proposition, we show that if $T > \tau$, then we can drop the last time period T in the Universal Fluid approximation. Therefore, we can always use a finite upper bound of τ on the possible realizations of the number of customer arrivals.

Proposition D.1 *If $T > \tau$, then we have $Z^*_{\mathsf{LP}}(\mathcal{T}) = Z^*_{\mathsf{LP}}(\mathcal{T} \setminus \{T\})$.*

Proof: Let $\boldsymbol{y}^* = (y^*_{jt} : j \in \mathcal{J},\ t \in \mathcal{T})$ be an optimal solution to the Universal Fluid. If $y^*_{jT} = 0$ for all $j \in \mathcal{J}$, then the result follows. Otherwise, there exists some product k such that $y^*_{kT} > 0$. We

will construct another optimal solution $\widehat{\boldsymbol{y}} = (\hat{y}_{jt} : j \in \mathcal{J},\ t \in \mathcal{T})$ to the Universal Fluid approximation such that $\hat{y}_{kT} = 0$ and $\hat{y}_{jT} = y^*_{jtT}$ for all $j \in \mathcal{J} \setminus \{k\}$, in which case, repeatedly applying the same construction for each $k \in \mathcal{J}$ such that $y^*_{kT} > 0$, the desired result follows. Let $k \in \mathcal{J}$ be such that $y^*_{kT} > 0$. Choose some resource i that is used by product k, so $\mathbf{1}_{(i \in A_k)} = 1$. Using the fact that $T > \tau$ we have $y^*_{kT} + \sum_{t=1}^{\tau} y^*_{kt} \leq \sum_{t \in \mathcal{T}} \mathbf{1}_{(i \in A_k)}\, y^*_{kt} \leq \sum_{j \in \mathcal{J}} \sum_{t \in \mathcal{T}} \mathbf{1}_{(i \in A_j)}\, y^*_{jt} \leq c_i$, where the last inequality holds because $\boldsymbol{y}^*$ satisfies the first constraint in the Universal Fluid approximation. By the definition of τ, we have $\tau \geq c_i / \overline{\lambda}$, so the last chain of inequalities yields $y^*_{kT} + \sum_{t=1}^{\tau} y^*_{kT} \leq \tau\, \overline{\lambda} \leq \sum_{t=1}^{\tau} \lambda_{kt}$, where we use the fact that $\lambda_{kt} \geq \overline{\lambda}$ for all $t \in \mathcal{T}$. Thus, we have $y^*_{kT} \leq \sum_{t=1}^{\tau} (\lambda_{kt} - y^*_{kt})$. Noting that $\lambda_{kt} - y^*_{kt} \geq 0$ for all $t = 1, \ldots, \tau$ by the second constraint in the Universal Fluid approximation, having $y^*_{kT} \leq \sum_{t=1}^{\tau} (\lambda_{kt} - y^*_{kt})$ implies that there exists a collection of non-negative numbers $\delta_1, \ldots, \delta_\tau$ such that we have $\sum_{t=1}^{\tau} \delta_t = y^*_{kT}$ and $\delta_t \leq \lambda_{kt} - y^*_{kt}$ for all $t = 1, \ldots, \tau$. In this case, we define the solution $\widehat{\boldsymbol{y}} = (\hat{y}_{jt} : j \in \mathcal{J},\ t \in \mathcal{T})$ as $\hat{y}_{jt} = y^*_{jt}$ for all $j \in \mathcal{J} \setminus \{k\}$, $t \in \mathcal{T}$ and

$$\hat{y}_{kt} = \begin{cases} y^*_{kt} + \delta_t & \text{if } t = 1, \ldots, \tau \\ y^*_{kt} & \text{if } t = \tau + 1, \ldots, T-1 \\ 0 & \text{if } t = T. \end{cases}$$

Because $\sum_{t=1}^{\tau} \delta_t = y^*_{kT}$, we have $\sum_{t \in \mathcal{T}} \hat{y}_{jt} = \sum_{t \in \mathcal{T}} y^*_{jt}$ for all $j \in \mathcal{J}$, so noting that $\boldsymbol{y}^*$ satisfies the first constraint in the Universal Fluid approximation, $\widehat{\boldsymbol{y}}$ satisfies this constraint too. Because $\delta_t \leq \lambda_{kt} - y^*_{kt}$, the solution $\widehat{\boldsymbol{y}}$ satisfies the second constraint in the Universal Fluid approximation as well.

Thus, the solution $\widehat{\boldsymbol{y}}$ is feasible to the Universal Fluid approximation. The difference between the objective function values provided by $\widehat{\boldsymbol{y}}$ and $\boldsymbol{y}^*$ is $f_k \sum_{t=1}^{\tau} \mathbb{P}\{D \geq t\}\, \delta_t - f_k\, \mathbb{P}\{D \geq T\}\, y^*_{kT} = f_k \sum_{t=1}^{\tau} [\mathbb{P}\{D \geq t\} - \mathbb{P}\{D \geq T\}]\, \delta_t \geq 0$, where we use $\mathbb{P}\{D \geq t\} \geq \mathbb{P}\{D \geq T\}$ for all $t = 1, \ldots, \tau$. ■

By the proposition above, we can drop all time periods in $\mathcal{T} \setminus \{1, \ldots, \tau\}$ from consideration in the Universal Fluid approximation. We can extend the proposition above to the case where there exists some $\overline{\lambda} > 0$ such that $\lambda_{jt} \geq \mathbf{1}_{(\lambda_{jt} > 0)}\, \overline{\lambda}$ for all $j \in \mathcal{J}$ and $t \in \mathcal{T}$, so that the nonzero values for the probability that an arriving customer is interested in a particular product is uniformly lower bounded by $\overline{\lambda}$. In this case, we define as τ before. Furthermore, for each $j \in \mathcal{J}$, we choose $K_j = 1, \ldots, T+1$ such that we have either $\sum_{t=1}^{K_j} \mathbf{1}_{(\lambda_{jt} > 0)} \geq \tau$ or $\sum_{K_j}^{T} \mathbf{1}_{(\lambda_{jt} > 0)} = 0$. Note that we can always choose $K_j = T+1$, so there is always a value for K_j that satisfies one of the two conditions. In this case, we can show that if $T > \max_{j \in \mathcal{J}} K_j$, then we have $Z^*_{\mathsf{LP}}(\mathcal{T}) = Z^*_{\mathsf{LP}}(\mathcal{T} \setminus \{T\})$. In particular, if $\sum_{K_j}^{T} \mathbf{1}_{(\lambda_{jt} > 0)} = 0$, then there are no requests for product j after time period K_j. Thus, we can indeed set the decision variable y_{jT} to zero in the Universal Fluid. On the other hand, if $\sum_{t=1}^{K_j} \mathbf{1}_{(\lambda_{jt} > 0)} \geq \tau$, then there are τ time periods at which there is demand for product j with a probability of at least $\overline{\lambda}$. In this case, we can use the same argument in the proof of Proposition D.1 to conclude that we can set the decision variable y_{jT} to zero in the Universal Fluid.