Fluid Approximations for Revenue Management under High-Variance Demand: Good and Bad Formulations

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

June 14, 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. Therefore, 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! This observation may also lead one to believe that fluid approximations work well simply because a Bernoulli process gives rise to a small coefficient of variation for the demand when the mean of the demand is large.

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 good fluid approximation for our revenue management model with random number of customer arrivals.

We look for three properties in a good fluid approximation. 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 when we have a system with large resource availability. Third, the relative optimality gap of policies driven by the fluid approximation should vanish 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 that satisfies 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, $2k$ and $k\sqrt{k-2}$, yielding a coefficient of variation of $\frac{1}{2}\sqrt{k-2}$. Therefore, 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 $k\sqrt{k}$. 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 $\frac{2}{1+\sqrt{k}}$, which does not approach one as the resource capacity becomes large. In contrary, the ratio converges to zero as the resource capacity becomes large, so the natural fluid approximation becomes especially poor, as opposed to being especially good, when the resource capacity becomes 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 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 2). 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 yield inferior policies under random number of customer arrivals.

Studying fluid approximations under arrival processes other than a Bernoulli process is far from pure intellectual curiosity. We give a numerical study to check the practical benefits from getting the fluid approximation right under high-variance demand. Our results indicate that the policies driven by our fluid approximation perform up to 12% better than those driven by the naive fluid

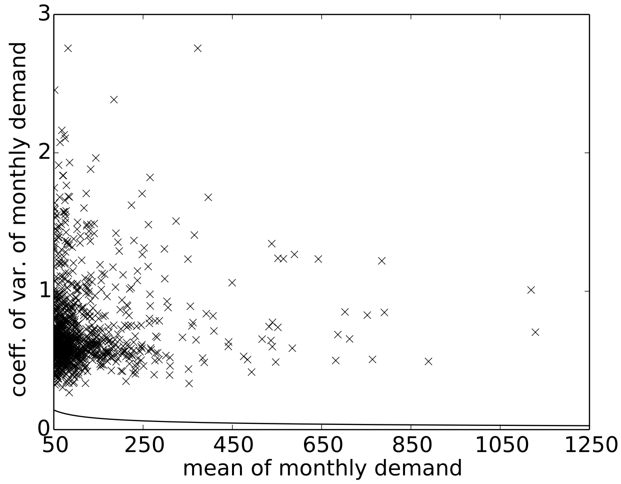

Figure 1 Mean and coefficient of variation of the monthly demand for different products.

approximation. Furthermore, the demand data can display variance that is significantly larger than what is implied by a discrete-time approximation to a Poisson process. We studied a publicly available dataset from an online appliance and electronics retailer; see Kaggle (2021). The dataset includes order transactions from January 5, 2020 to November 21, 2020. Each row corresponds to an order transaction, providing the unique product code, brand and category of the product purchased, date of the transaction and price of the product. To avoid truncation effects, we focus on the nine month period from February 1 to October 31 and calculate the monthly sales for each product. Using these nine data points for each product, we estimate the mean and variance of the monthly demand under the assumption that sales equal demand. We drop the products with mean monthly demand less than 50. In Figure 1, each cross corresponds to a product, plotting the mean and coefficient variation of the monthly demand for the product. Observe that there are products with mean monthly demand of about 500 and coefficient of variation exceeding one. If the demand were driven by a Poisson process, then a mean of 500 for the demand would imply a coefficient of variation of $1/\sqrt{500}$, which is only about 0.045. 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. All of 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. In our analysis, we dropped seasonal products such as air conditioners and space heaters and checked that the prices of the products do not vary significantly from month to month. Although our findings do not represent a full statistical analysis, they suggest that the demand data can have significantly larger variance than what is implied by a discrete-time approximation to a Poisson process.

Lastly, we note that we give all of 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. Nevertheless, our results can easily be extended to the case where the demand for each product depends on the prices or the assortment of available products.

Related Literature. Fluid approximations have traditionally been studied under demand models based on a Bernoulli process. Gallego and van Ryzin (1994) consider an asymptotic regime where the expected number of customer arrivals and the resource capacity scale with k and show that the ratio between the optimal total expected revenue and the optimal objective value of the fluid approximation is $\Omega\Big(1-\frac{1}{\sqrt{k}}\Big)$. The authors also give a policy from the fluid approximation that attains the same optimality gap. 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 multiple resources. Talluri and van Ryzin (1998) show the asymptotic optimality of a policy driven by a dual solution to the fluid approximation. Liu and van Ryzin (2008) and Gallego et al. (2004) generalize these results to models with customer choice. 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. Note that the total demand divided by k does not converge to a fixed number in distribution in the problem instance discussed earlier in this section. 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 $\Omega\Big(1-\frac{1}{\sqrt{k}}\Big)$. 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 $\Omega\Big(1-\frac{1}{\sqrt[3]{c_{\min}}}\Big)$, 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 $\Omega\Big(1-\frac{1}{\sqrt{c_{\min}}}\Big)$. The papers discussed so far focus on a Bernoulli process. Similar to us, Walczak (2006) observes that a Bernoulli process limits the demand variability and incorporates high variance by using dynamic programs with batch arrivals, which are difficult to solve when the number of resources is large.

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 and the ratio between the optimal total expected and its optimal objective value is $\Omega\Big(1-\sqrt{\frac{\log c_{\min}}{c_{\min}}}\Big)$. In Section 5, we give a numerical study.

2. Revenue Management with Random Number of Arrivals

We give a natural 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 $\rho_t = \mathbb{P}\{D \geq t+1 \mid D \geq t\}$. 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, we can divide the selling horizon into subintervals, say weeks, considering the case where we have a random number of customer arrivals in each subinterval. This extension can be useful in practice, but it brings notational overhead without making our results any 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 fluid approximation where 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. This approach is arguably the most natural way to formulate a fluid approximation that corresponds to the dynamic program in Section 2, but we will see that the relative gap between the optimal objective value of this fluid approximation and the optimal total expected revenue does not necessarily 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, we consider the linear program

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

where the objective keeps the total expected revenue over the selling horizon and the first constraint ensures that the expected capacity consumption of resource i does not exceed its capacity.

Although the Bad Fluid approximation is rather natural, we will give a problem instance to show that the relative gap between its optimal objective value and the optimal total expected revenue does not necessarily vanish as the resource capacities get large. This observation may lead one to believe that it is not possible to give a fluid approximation that satisfies this property when the number of customer arrivals has an arbitrary distribution, but this belief is not correct. In the next section, we will give an alternative fluid approximation such that the ratio between the optimal objective value of the fluid approximation and the optimal total expected revenue gets arbitrarily close to one as the resource capacities get large. Furthermore, we will extract a policy from this fluid approximation that is asymptotically optimal as the resource capacities get large.

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 $\mathbb{P}\{D = k\} = 1 - \frac{1}{k-1}$ and $\mathbb{P}\{D = k^2\} = \frac{1}{k-1}$. So, $\mathbb{E}\{D\} = 2k$ and $\mathrm{Var}(D) = k^2(k-2)$. The capacity of the resource is $k\sqrt{k}$. At each of the first k time periods, an arriving customer requests the product with probability $\frac{1}{\sqrt{k}}$. At each of the last $k^2 - k$ time periods, an arriving customer requests the product with probability 1. We choose $k \geq 3$ so that $k^2 - k \geq k\sqrt{k}$, so if there happens to be k^2 time periods in the selling horizon, then we can sell all of the capacity of the resource by using the customer arrivals at the last $k^2 - k$ time periods. 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 $D=k$, then the capacity of $k\sqrt{k}$ allows us to accept all customer requests and we obtain a total expected revenue of $k\,\frac{1}{\sqrt{k}}$, where we use the fact that the customers arriving at each of the first k time periods request the product with probability $\frac{1}{\sqrt{k}}$. If $D=k^2$, then we can sell all of the capacity by using the customer arrivals at the last k^2-k time periods, so noting that the capacity of the resource is $k\sqrt{k}$, we obtain a total expected revenue of $k\sqrt{k}$. Thus, the optimal total expected revenue is $\mathbb{P}\{D=k\}\sqrt{k}+\mathbb{P}\{D=k^2\}\,k\sqrt{k}=\left(1-\frac{1}{k-1}\right)\sqrt{k}+\frac{1}{k-1}\,k\sqrt{k}=2\sqrt{k}$. In contrast, dropping the indices for the single resource and product, the Bad Fluid approximation is

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

Setting all of the decision variables to their upper bounds provides a feasible, as well as an optimal, solution with the objective value $k\,\frac{1}{\sqrt{k}}+\frac{1}{k-1}\,(k^2-k)=k+\sqrt{k}$, as $k+\sqrt{k}\le k\sqrt{k}$ for $k\ge 3$.

Thus, the ratio between the optimal total expected revenue and the optimal objective value of the Bad Fluid approximation is $\frac{2}{1+\sqrt{k}}$, which does not approach one as k gets large. Quite the contrary, this ratio converges to zero, so the Bad 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 Bad Fluid approximation. Careful reader will see that there is a product request with probability $\frac{1}{\sqrt{k}}$ at the first k time periods of our problem instance, even though we have $\sum_{j\in\mathcal{J}}\lambda_{jt}=1$ for all $t\in\mathcal{T}$ in Section 2, but we can easily avoid this discrepancy by introducing an additional product with infinitesimal revenue and a request probability of $1-\frac{1}{\sqrt{k}}$ at the first k time periods.

Closing this section, making the change of variables $w_j=\sum_{t\in\mathcal{T}}\mathbb{P}\{D\ge t\}\,y_{jt}$ and using the decision variables $\boldsymbol{w}=(w_j : j\in\mathcal{J})$, we can equivalently write the Bad Fluid approximation as

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

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 Bad Fluid approximation. The Compact problem deceptively appears to be the bona fide fluid approximation! Note that the total expected number of requests for product j is $\sum_{t\in\mathcal{T}}\mathbb{P}\{D\ge t\}\,\lambda_{jt}$. Thus, viewing the decision variable w_j as the total expected sales for product j, the first constraint ensures that the total expected capacity consumption of resource i does not exceed its capacity. The second constraint ensures that the total expected sales for product j does not exceed its total expected demand. It is surprising that a fluid approximation as natural as the one in the Compact problem is not the right fluid approximation. Next, we fix this approximation.

4. A Fluid Approximation that Checks the Boxes

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, consider the linear program

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

We comment on the unexpected form of the problem above. 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 even be led to 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. Consequently, it is not immediately clear that the Good Fluid approximation provides an upper bound on the optimal total expected revenue. In the next theorem, we show that the optimal objective value of the Good 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. In the proof, we relax the capacity constraints for the resources in the dynamic program in Section 2 by associating Lagrange multipliers with them. The value functions of the relaxed dynamic program provide upper bounds on the value functions $\{J_t : t\in\mathcal{T}\}$. Therefore, we can use the value functions of the relaxed dynamic program at the first time period to get an upper bound on $J_1(\boldsymbol{c})$. We show that the problem of choosing the Lagrange multipliers to tighten the upper bound on $J_1(\boldsymbol{c})$ surprisingly reduces to the Good Fluid approximation instead of the Bad Fluid one. Thus, the Good Fluid approximation satisfies the first property in the introduction that we would expect from a good fluid approximation. The constraints in the Good Fluid approximation are at least as tight as those in the Bad Fluid one, so by Theorem 4.1, the optimal objective value of the Bad Fluid approximation is also an upper bound on the optimal total expected revenue. However, as discussed in the previous section, the Bad Fluid approximation is not asymptotically tight in the sense that the ratio between the optimal objective value of the

$\mathsf{Bad\ Fluid}$ approximation and the optimal total expected revenue does not necessarily get arbitrarily close to one as the resource capacities get large. In the next theorem, we show that the $\mathsf{Good\ Fluid}$ approximation is indeed asymptotically tight as the resource capacities get large.

Theorem 4.2 (Asymptotic Tightness of the Good Fluid Approximation) *Letting* $c_{\min} = \min_{i \in \mathcal{L}} c_i$ *and* $L = \max_{j \in \mathcal{J}} |A_j|$, *we have*

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

We give the proof of the theorem in Appendix B. Because $1 \geq J_1(\boldsymbol{c})/Z^*_{\mathsf{LP}}$ by Theorem 4.1, the theorem above implies that the ratio between the optimal total expected revenue and the optimal objective value of the $\mathsf{Good\ Fluid}$ approximation converges to one as the resource capacities get large. Thus, the $\mathsf{Good\ Fluid}$ approximation satisfies the second property in the introduction that we would expect from a good fluid approximation. In the proof of Theorem 4.2, letting $\boldsymbol{y}^* = (y^*_{jt} : j \in \mathcal{J}\ t \in \mathcal{T})$ be an optimal solution to the $\mathsf{Good\ Fluid}$ approximation, we consider an approximate policy. For some $\theta \in (0,1)$, if there is enough capacity to serve a request for product j at time period t, then the approximate policy makes product j available for purchase with probability $\theta\, y^*_{jt}/\lambda_{jt}$. We show that the total expected revenue of the approximate policy is lower bounded by $\left(1 - \sqrt{\frac{2 \log c_{\min}}{c_{\min}}} - \frac{L}{c_{\min}}\right) Z^*_{\mathsf{LP}}$ for a choice of θ, in which case, the result follows by noting that the total expected revenue of the approximate policy is also upper bounded by $J_1(\boldsymbol{c})$. Once we get the form of the fluid approximation in the $\mathsf{Good\ Fluid}$ approximation right, it is not difficult to get the lower bound on the total expected revenue of the approximate policy. However, it is critical to use the right fluid approximation and we cannot get the analogous result by using the $\mathsf{Bad\ Fluid}$ approximation. In the proof of the theorem, we use the one-sided Bernstein inequality to lower bound the total expected revenue of the approximate policy. Using the Markov inequality instead, we can also show that the total expected revenue of the approximate policy is lower bounded by $\frac{1}{4L} Z^*_{\mathsf{LP}}$. The latter lower bound is independent of the capacities of the resources, but does not get arbitrarily close to one as the resource capacities get large. When L is fixed, the theorem above implies that $\frac{J_1(\boldsymbol{c})}{Z^*_{\mathsf{LP}}} = \Omega\left(1 - \sqrt{\frac{\log c_{\min}}{c_{\min}}}\right)$. Lastly, because $J_1(\boldsymbol{c}) \leq Z^*_{\mathsf{LP}}$ by Theorem 4.1, the total expected revenue of the approximate policy is also lower bounded by $\left(1 - \sqrt{\frac{2 \log c_{\min}}{c_{\min}}} - \frac{L}{c_{\min}}\right) J_1(\boldsymbol{c})$, so the relative optimality gap of the approximate policy driven by the $\mathsf{Good\ Fluid}$ approximation vanishes as the resource capacities get large. Thus, the $\mathsf{Good\ Fluid}$ approximation satisfies the third property in the introduction that we would expect from a good fluid approximation, so it checks all the boxes.

Note that the $\mathsf{Bad\ Fluid}$ approximation has an equivalent reformulation as in the $\mathsf{Compact}$ problem, but the $\mathsf{Good\ Fluid}$ approximation does not admit a similar equivalent reformulation.

5. Numerical Study

We give a numerical study to show the benefits from using the Good Fluid approximation instead of the Bad Fluid one when we have a random number of customer arrivals. We generate a number of test problems. We check the upper bound on the optimal total expected revenue from the two fluid approximations, as well as the performance of 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 $v \in \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1\}$. In Appendix C, we give the details of our approach for generating our test problems.

For each test problem, we solve the Good Fluid and Bad Fluid approximations, 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 Good Fluid or Bad Fluid approximation, the policy makes product j available at time period t with probability $\frac{y^*_{jt}}{\lambda_{jt}}$. Our results are in Table 1. The first and second columns give the values of (v, μ) and $c_{\min}$ for the test problem. The third, fourth and fifth columns focus on the Good Fluid approximation and give the optimal objective value of the Good Fluid approximation, the total expected revenue of the policy from the Good Fluid approximation and the ratio between the performance of the policy and the optimal objective value of the fluid approximation. By the discussion just after Theorem 4.2, the entries of the fifth column approach one as $c_{\min}$ gets large. The sixth, seventh and eighth columns focus on the Bad 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.

Our results indicate that the ratio between the optimal objective value of the Good Fluid approximation and the total expected revenue of its corresponding policy, as expected, gets close to one as $c_{\min}$ gets large. When $c_{\min} = 173$, the ratio exceeds 0.96 in all test problems. On the other hand, the ratio between the optimal objective value of the Bad Fluid approximation and its corresponding policy does not get close to one as $c_{\min}$ gets large. For the test problems with the smallest coefficient of variation of 1/8, the ratio does not reach above 0.88 even when $c_{\min} =$

Params. (v, μ)	$c_{\min}$	Good Fluid Approx. Obj.	Policy	Ratio	Bad Fluid Approx. Obj.	Policy	Ratio	Obj. Gap	Policy Gap
$(\frac{1}{8}, 400)$	22	17,034	15,520	0.91	17,278	14,776	0.86	1.43%	4.79%
$(\frac{1}{8}, 800)$	43	34,072	31,826	0.93	34,554	30,089	0.87	1.41%	5.46%
$(\frac{1}{8}, 1600)$	86	68,203	64,720	0.95	69,180	60,689	0.88	1.43%	6.23%
$(\frac{1}{8}, 3200)$	173	136,406	130,983	0.96	138,359	121,988	0.88	1.43%	6.87%
$(\frac{1}{4}, 400)$	22	16,647	15,416	0.93	17,268	14,938	0.87	3.73%	3.10%
$(\frac{1}{4}, 800)$	43	33,299	31,541	0.95	34,534	30,034	0.87	3.71%	4.78%
$(\frac{1}{4}, 1600)$	86	66,638	64,428	0.97	69,140	61,243	0.89	3.75%	4.94%
$(\frac{1}{4}, 3200)$	173	133,273	129,803	0.97	138,280	121,540	0.88	3.76%	6.37%
$(\frac{1}{2}, 400)$	22	15,540	14,502	0.93	17,242	13,558	0.79	10.95%	6.51%
$(\frac{1}{2}, 800)$	43	31,069	29,954	0.96	34,481	27,340	0.79	10.98%	8.73%
$(\frac{1}{2}, 1600)$	86	62,165	59,745	0.96	69,034	53,690	0.78	11.05%	10.13%
$(\frac{1}{2}, 3200)$	173	124,321	122,401	0.98	138,068	110,811	0.80	11.06%	9.47%
$(1, 400)$	22	13,270	12,384	0.93	17,170	11,149	0.65	29.39%	9.98%
$(1, 800)$	43	26,523	25,092	0.95	34,337	22,060	0.64	29.46%	12.09%
$(1, 1600)$	86	53,057	51,176	0.96	68,746	44,702	0.65	29.57%	12.65%
$(1, 3200)$	173	106,098	105,015	0.99	137,491	91,682	0.67	29.59%	12.70%

Table 1 Comparison of the two fluid approximations.

173. The ratio can be even smaller, so even further from one, for the test problems with larger coefficients of variation. For the test problems with the largest coefficient of variation of one, the ratio never reaches above 0.67. Therefore, 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 Good Fluid approximation can be substantially tighter than the one provided by the Bad Fluid approximation. In particular, the upper bounds from the two fluid approximations can differ by as much as 29.59%. Lastly, 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 differ by as much as 12.70%, in favor of the Good Fluid approximation.

6. Conclusions

We made three contributions. First, the form of our fluid approximation 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. It is hard to interpret the left side of the capacity constraints in our fluid approximation as the expected capacity consumption of resources, so it is not immediately clear that the optimal objective value of our fluid approximation is an upper bound on the optimal total expected revenue. A naive fluid approximation that uses the expected capacity consumption of 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, far from being an intellectual curiosity. 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 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. One can explore the analogues of these results under random number of customer arrivals. Lastly, our numerical study showed that our fluid approximation can make a significant impact in the performance of the corresponding policy. We can check the benefits from getting the fluid approximation right when the demand for the products depends on the prices or the assortment of available products.

References

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.

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: Good and Bad Formulations

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

We give a proof for Theorem 4.1. We will use an equivalent reformulation of the dynamic program in Section 2. In our equivalent reformulation, we use the decision variables $\boldsymbol{u}=(u_j : j\in\mathcal{J})\in\{0,1\}^n$, 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 $\mathcal{U}(\boldsymbol{x})=\{\boldsymbol{u}\in\{0,1\}^{|\mathcal{J}|} : \mathbf{1}_{(i\in A_j)}\, u_j\le x_i\ \forall\, i\in\mathcal{L},\ j\in\mathcal{J}\}$, 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 $J_{T+1}=0$. In the next lemma, we give an upper bound on the value functions $\{J_t : t\in\mathcal{T}\}$ computed through the dynamic program above.

Lemma A.1 *Letting $R_t=\rho_t\,\rho_{t+1}\ldots\rho_{T-1}$ with $R_T=1$, for any $\boldsymbol{\theta}=(\theta_i : i\in\mathcal{L})\in\mathbb{R}_+^{|\mathcal{L}|}$, $\boldsymbol{x}\in\mathbb{Z}_+^{|\mathcal{L}|}$ and $t\in\mathcal{T}$, the value functions $\{J_t : t\in\mathcal{T}\}$ in (1) satisfy*

$$J_t(\boldsymbol{x})\ \le\ \sum_{k=t}^{T}\sum_{j\in\mathcal{J}}\lambda_{jk}\left[\frac{R_t}{R_k}f_j-R_t\sum_{i\in A_j}\theta_i\right]^+ +R_t\sum_{i\in\mathcal{L}}\theta_i\,x_i.$$

Proof: We show the result by using induction over the time periods. At time period T, by (1), we have $J_T(\boldsymbol{x})=\sum_{j\in\mathcal{J}}\lambda_{jT}\,f_j\,u_j^*$ for some $\boldsymbol{u}^*\in\mathcal{U}(\boldsymbol{x})$. Noting that $\mathbf{1}_{(i\in A_j)}\,u_j^*\le x_i$ and $\theta_i\ge 0$, we get

$$\begin{aligned}J_T(\boldsymbol{x})\ &=\ \sum_{j\in\mathcal{J}}\lambda_{jT}\,f_j\,u_j^*\ \le\ \sum_{j\in\mathcal{J}}\lambda_{jT}\,f_j\,u_j^*+\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{L}}\lambda_{jT}\,\theta_i\left[x_i-\mathbf{1}_{(i\in A_j)}\,u_j^*\right]\\ &\overset{(a)}{=}\ \sum_{j\in\mathcal{J}}\lambda_{jT}\left[f_j-\sum_{i\in\mathcal{L}}\mathbf{1}_{(i\in A_j)}\,\theta_i\right]u_j^*+\sum_{i\in\mathcal{L}}\theta_i\,x_i\ \overset{(b)}{\le}\ \sum_{j\in\mathcal{J}}\lambda_{jT}\left[f_j-\sum_{i\in A_j}\theta_i\right]^+ +\sum_{i\in\mathcal{L}}\theta_i\,x_i,\end{aligned}$$

where (a) is by arranging the terms and noting that $\sum_{j\in\mathcal{J}}\lambda_{jT}=1$ and (b) holds since $\sum_{j\in\mathcal{J}}\alpha_j\,u_j^*\le\sum_{j\in\mathcal{J}}[\alpha_j]^+$ for any $(\alpha_j : j\in\mathcal{J})\in\mathbb{R}^{|\mathcal{J}|}$ and $\boldsymbol{u}^*\in\{0,1\}^n$. 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. For notational brevity, let $K_t=\sum_{k=t}^{T}\sum_{j\in\mathcal{J}}\lambda_{jk}\,[\frac{R_t}{R_k}f_j-R_t\sum_{i\in A_j}\theta_i]^+$. By the induction assumption, we have $J_{t+1}(\boldsymbol{x}) \le K_{t+1} + R_{t+1} \sum_{i\in\mathcal{L}} \theta_i\, x_i$ for any $\boldsymbol{x} \in \mathbb{Z}_+^{|\mathcal{L}|}$. By (1), we have $J_t(\boldsymbol{x}) = \sum_{j\in\mathcal{J}} \lambda_{jt} \{f_j\, u_j^* + \rho_t\, J_{t+1}(\boldsymbol{x} - u_j^* \sum_{i\in A_j} \boldsymbol{e}_i)\}$ for some $\boldsymbol{u}^* \in \mathcal{U}(\boldsymbol{x})$. In this case, we get

$$\begin{aligned}
J_t(\boldsymbol{x}) &= \sum_{j\in\mathcal{J}} \lambda_{jt} \left\{ f_j\, u_j^* + \rho_t\, J_{t+1}\Big(\boldsymbol{x} - u_j^* \sum_{i\in A_j} \boldsymbol{e}_i\Big)\right\} \\
&\overset{(c)}{\le} \sum_{j\in\mathcal{J}} \lambda_{jt} \left\{ f_j\, u_j^* + \rho_t\, K_{t+1} + \rho_t\, R_{t+1} \sum_{i\in\mathcal{L}} \theta_i\, (x_i - \mathbf{1}_{(i\in A_j)}\, u_j^*)\right\} \\
&\overset{(d)}{=} \sum_{j\in\mathcal{J}} \lambda_{jt} \Big[f_j - \rho_t\, R_{t+1} \sum_{i\in A_j} \theta_i \Big]\, u_j^* + \rho_t\, K_{t+1} + \rho_t\, R_{t+1} \sum_{i\in\mathcal{L}} \theta_i\, x_i \\
&\le \sum_{j\in\mathcal{J}} \lambda_{jt} \Big[f_j - \rho_t\, R_{t+1} \sum_{i\in A_j} \theta_i \Big]^+ + \rho_t\, K_{t+1} + \rho_t\, R_{t+1} \sum_{i\in\mathcal{L}} \theta_i\, x_i \overset{(e)}{=} K_t + R_t \sum_{i\in\mathcal{L}} \theta_i\, x_i,
\end{aligned}$$

where (c) is by the induction assumption, (d) uses the fact that $\sum_{j\in\mathcal{J}} \lambda_{jt} = 1$ and (e) holds because $R_t = \rho_t\, R_{t+1}$ 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 definition of K_t. ■

We have $R_t = \rho_t\, \rho_{t+1} \ldots \rho_{T-1} = \frac{\mathbb{P}\{D\ge t+1\}}{\mathbb{P}\{D\ge t\}} \frac{\mathbb{P}\{D\ge t+2\}}{\mathbb{P}\{D\ge t+1\}} \cdots \frac{\mathbb{P}\{D\ge T\}}{\mathbb{P}\{D\ge T-1\}} = \frac{\mathbb{P}\{D\ge T\}}{\mathbb{P}\{D\ge t\}}$. Since $\mathbb{P}\{D \ge 1\} = 1$, the last equality also yields $R_1 = \mathbb{P}\{D \ge T\}$. Using Lemma A.1 with $t = 1$ and $\boldsymbol{x} = \boldsymbol{c}$, we obtain

$$\begin{aligned}
J_1(\boldsymbol{c}) &\le \sum_{t=1}^{T} \sum_{j\in\mathcal{J}} \lambda_{jt} \Big[\frac{R_1}{R_t} f_j - R_1 \sum_{i\in A_j} \theta_i \Big]^+ + R_1 \sum_{i\in\mathcal{L}} \theta_i\, c_i \\
&= \sum_{t=1}^{T} \sum_{j\in\mathcal{J}} \lambda_{jt} \Big[\mathbb{P}\{D \ge t\}\, f_j - \mathbb{P}\{D\ge T\} \sum_{i\in\mathcal{L}} \mathbf{1}_{(i\in A_j)}\, \theta_i \Big]^+ + \mathbb{P}\{D\ge T\} \sum_{i\in\mathcal{L}} \theta_i\, c_i. \qquad (2)
\end{aligned}$$

The inequality above holds for any $\boldsymbol{\theta} \in \mathbb{R}_+^{|\mathcal{L}|}$. Thus, if we minimize the expression on the right side of the inequality above over all $\boldsymbol{\theta} \in \mathbb{R}_+^{|\mathcal{L}|}$, then we still 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 $\boldsymbol{\theta} = (\theta_i : i \in \mathcal{L}) \in \mathbb{R}_+^{|\mathcal{L}|}$ 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 (2) over all $\boldsymbol{\theta} \in \mathbb{R}_+^{|\mathcal{L}|}$ by solving the linear program

$$\min_{(\boldsymbol{\theta},\boldsymbol{z})\in\mathbb{R}_+^{|\mathcal{L}|+|\mathcal{J}||\mathcal{T}|}} \left\{ \sum_{t\in\mathcal{T}} \sum_{j\in\mathcal{J}} \lambda_{jt}\, z_{jt} + \mathbb{P}\{D\ge T\} \sum_{i\in\mathcal{L}} c_i\, \theta_i \ : \ z_{jt} \ \ge\ \mathbb{P}\{D\ge t\}\, f_j - \mathbb{P}\{D\ge T\} \sum_{i\in\mathcal{L}} \mathbf{1}_{(i\in A_j)}\, \theta_i \quad \forall\, j \in \mathcal{J},\ t\in\mathcal{T} \right\}. \qquad (3)$$

By the discussion right after (2), the optimal objective value of the linear program above is an upper bound on the optimal total expected revenue.

Without loss of generality, we have $\mathbb{P}\{D \ge T\} > 0$. If $\mathbb{P}\{D \ge T\} = 0$, then we can choose the upper bound of the support of the number of customer arrivals 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, the dual of (3) is given by

$$\max_{\boldsymbol{y} \in \mathbb{R}_+^{|\mathcal{J}||\mathcal{T}|}} \left\{ \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} f_j \, \mathbb{P}\{D \geq t\} \, y_{jt} \ \ : \ \ \mathbb{P}\{D \geq T\} \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{J}} \mathbf{1}_{(i \in A_j)} \, y_{jt} \ \leq \ \mathbb{P}\{D \geq T\} \, c_i \quad \forall \, i \in \mathcal{L} \right.$$
$$\left. y_{jt} \ \leq \ \lambda_{jt} \quad \forall \, j \in \mathcal{J},\ t \in \mathcal{T} \right\}. \tag{4}$$

Because $\mathbb{P}\{D \geq T\} > 0$, (4) is equivalent to the Good Fluid approximation. The desired result follows since (3) and (4) have the same optimal objective value, which is an upper bound on $J_1(\boldsymbol{c})$. ■

Appendix B: Asymptotic Tightness of the Fluid Approximation

We give a proof for Theorem 4.2. Let $\boldsymbol{y}^* = (y_{jt} : j \in \mathcal{J},\ t \in \mathcal{T})$ be an optimal solution to the Good 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 App be the total expected revenue of the approximate policy, we get

$$\mathsf{App} = \sum_{t=1}^{T} \sum_{j \in \mathcal{J}} f_j \, \mathbb{P}\{D \geq t\} \, \mathbb{P}\{A_{jt} = 1\} \, \mathbb{P}\{\Lambda_{jt} = 1\} \, \mathbb{P}\{G_{jt} = 1\}$$
$$= \sum_{t=1}^{T} \sum_{j \in \mathcal{J}} f_j \, \mathbb{P}\{D \geq t\} \, \theta \, \frac{y^*_{jt}}{\lambda_{jt}} \, \lambda_{jt} \, \mathbb{P}\{G_{jt} = 1\}. \tag{5}$$

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 $\mathbb{P}\{\sum_{k=1}^{T} N_{ik} < c_i \;\; \forall\, i \in A_j\} \le \mathbb{P}\{G_{jt} = 1\}$. 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*

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

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

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

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

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

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

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

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 4.2.

Proof of Theorem 4.2:

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

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

where (a) is the union bound, (b) follows from Lemma B.1, as well as the fact that $|A_j| \leq L$ 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 $\mathbb{P}\{G_{jt}=1\} \geq 1 - \frac{L}{c_{\min}}$ with this choice of θ. Thus, by (5), we get

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

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

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 $\mathbb{P}\{\Gamma \leq k\} \geq 0.99$, we set the maximum length of the selling horizon as $T = k$. For each $t = 1, \ldots, T$, letting $\gamma_t = \mathbb{P}\{t-1 \leq \Gamma \leq t\}$, 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 $\kappa = 8$.

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 Good Fluid approximation 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 Good 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 Good Fluid approximation. 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 Good 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 $y^*_{kT} + \sum_{t=1}^{\tau} y^*_{kt} \le \sum_{t \in \mathcal{T}} \mathbf{1}_{(i \in A_k)}\, y^*_{kt} \le \sum_{j \in \mathcal{J}} \sum_{t \in \mathcal{T}} \mathbf{1}_{(i \in A_j)}\, y^*_{jt} \le c_i$, where the last inequality holds because $\boldsymbol{y}^*$ satisfies the first constraint in the Good Fluid approximation. By the definition of τ, we have $\tau \ge c_i / \overline{\lambda}$, so the last chain of inequalities yields $y^*_{kT} + \sum_{t=1}^{\tau} y^*_{kT} \le \tau\, \overline{\lambda} \le \sum_{t=1}^{\tau} \lambda_{kt}$, where we use the fact that $\lambda_{kt} \ge \overline{\lambda}$ for all $t \in \mathcal{T}$. Thus, we have $y^*_{kT} \le \sum_{t=1}^{\tau} (\lambda_{kt} - y^*_{kt})$. Noting that $\lambda_{kt} - y^*_{kt} \ge 0$ for all $t = 1, \ldots, \tau$ by the second constraint in the Good Fluid approximation, having $y^*_{kT} \le \sum_{t=1}^{\tau} (\lambda_{kt} - y^*_{kt})$ implies that there exists a collection of nonnegative numbers $\delta_1, \ldots, \delta_\tau$ such that we have $\sum_{t=1}^{\tau} \delta_t = y^*_{kT}$ and $\delta_t \le \lambda_{kt} - y^*_{kt}$ 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 Good Fluid approximation, $\widehat{\boldsymbol{y}}$ satisfies this constraint too. Because $\delta_t \le \lambda_{kt} - y^*_{kt}$, the solution $\widehat{\boldsymbol{y}}$ satisfies the second constraint in the Good Fluid approximation as well.

Thus, the solution $\widehat{\boldsymbol{y}}$ is feasible to the Good Fluid approximation. The difference between the objective function values provided by $\widehat{\boldsymbol{y}}$ and $\boldsymbol{y}^*$ is $f_k \sum_{t=1}^{\tau} \mathbb{P}\{D \ge t\}\, \delta_t - f_k\, \mathbb{P}\{D \ge T\}\, y^*_{kT} = f_k \sum_{t=1}^{\tau} [\mathbb{P}\{D \ge t\} - \mathbb{P}\{D \ge T\}]\, \delta_{kt} \ge 0$, where we use $\mathbb{P}\{D \ge t\} \ge \mathbb{P}\{D \ge T\}$ 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 Good Fluid approximation. We can extend the proposition above to the case where there exists some $\overline{\lambda} > 0$ such that $\lambda_{jt} \ge \mathbf{1}_{(\lambda_{jt} > 0)}\, \overline{\lambda}$ 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 $\sum_{t=1}^{K_j} \mathbf{1}_{(\lambda_{jt} > 0)} \ge \tau$ 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 Good Fluid approximation. On the other hand, if $\sum_{t=1}^{K_j} \mathbf{1}_{(\lambda_{jt} > 0)} \ge \tau$, 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 Good Fluid approximation.