POMS PRODUCTION AND OPERATIONS MANAGEMENT

PRODUCTION AND OPERATIONS MANAGEMENT POMS

Vol. 26, No. 11, November 2017, pp. 1971–1988
ISSN 1059-1478 | EISSN 1937-5956 | 17 | 2611 | 1971
DOI 10.1111/poms.12750

Mitigating Inventory Overstocking: Optimal Order-up-to Level to Achieve a Target Fill Rate over a Finite Horizon

Yinliang (Ricky) Tan
A. B. Freeman School of Business, Tulane University, New Orleans, Louisiana 70118, USA, ytan2@tulane.edu

Anand A. Paul
Department of Information Systems and Operations Management, Warrington College of Business Administration, University of Florida, Gainesville, Florida 32611, USA, paulaa@ufl.edu

Qi Deng
Research Institute for Interdisciplinary Sciences, School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China, qideng@sufe.edu.cn

Lai Wei
Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 200030, China, laiwei@sjtu.edu.cn

Service level agreements (SLAs) are widely adopted performance-based contracts in operations management practice, and fill rate is the most common performance metric among all the measurements in SLAs. Traditional procedures characterizing the order-up-to level satisfying a specified fill rate implicitly assume an *infinite* performance review horizon. However, in practice, inventory managers are liable to maintain and report fill rates over a *finite* performance review horizon. This horizon discrepancy leads to deviation between the target fill rate and actual achieved fill rate. In this study, we first examine the behavior of the fill rate distribution over a finite horizon with positive lead time. We analytically prove that the expected fill rate assuming an *infinite* performance review horizon exceeds the expected fill rate assuming a *finite* performance review horizon, implying that there exists some inventory "waste" (i.e., overstocking) when the traditional procedure is used. Based on this observation and the complexity of the problem, we propose a simulation-based algorithm to reduce excess inventory while maintaining the contractual target fill rate. When the lead time is significant relative to the length of the contract horizon, we show that the improvement in the inventory system can be over 5%. Further, we extend our basic setting to incorporate the penalty for failing to meet a target, and show how one can solve large-scale problems via stochastic approximation. The primary managerial implication of our study is that ignoring the performance review horizon in an SLA will cause overstocking, especially when the lead time is large.

Key words: service level agreement; fill rate; positive lead time; base-stock policy; simulation-based optimization
History: Received: October 2016; Accepted: June 2017 by Chelliah Sriskandarajah, after 3 revisions.

1. Introduction

> "Inventory, a fundamental evil, declines in value by 1% to 2% a week in normal times, faster in tough times like the present. You want to manage it like you're in the dairy business. If it gets past its freshness date, you have a problem."
>
> Tim Cook, CEO of Apple Inc.

Inventory is one of the largest investments made by most businesses. It is also well recognized that inventory management is one of the most challenging business functions. According to a monthly survey by the U.S. Census Bureau,[1] in November 2016, the value of manufacturers' and trade inventories (including retailers and merchant wholesalers) was estimated at $1827.5 billion, which accounts for more than 10% of the annual gross domestic product (GDP) of the United States. Against this backdrop, even slight improvements in inventory management will result in dramatic savings due to the size of the gross volume. In this study, we first demonstrate a common problem that afflicts service level agreements (SLAs), then propose an innovative solution which can be easily implemented by a wide range of practitioners to reduce inventory levels while achieving target service levels.

Fill rate is defined as the average fraction of demand that is immediately satisfied from stock. An earlier noteworthy study has shown that using order-up-to level determined by current commercial software/algorithms will lead to substantially higher achieved fill rates as compared to the fill rates specified by contract over finite performance review horizon with zero-lead time (Thomas 2005). This is to say that the current formula used in textbooks and prevalent commercial software[2] results in excess inventory, which translates to unnecessarily high inventory holding costs. The root cause of this overestimation is that the traditional formula always assumes the performance review horizon to be *infinite*, whereas in practice the SLA requires the supplier to meet the target fill rate over a specific, *finite*, review period (e.g., a month, week or quarter). Considering the enormous gross inventory levels ($1827.5 billion in the United States), the possible savings from improving the inventory system are substantial. In Table 1 below, we give an example of our results. When inventory managers check the fill rate biweekly, on average they actually achieve a 92.73% fill rate, while the contractual target fill rate is only 90%. Now consider what happens in terms of the order-up-to level. This 2.73% difference translates into overstocking by 4.04%.

Despite several research papers having identified and described this interesting overestimation phenomenon from different perspectives (Banerjee and Paul 2005, Chen et al. 2003, Thomas 2005), little has been published to tackle this very important but overlooked issue. One possible reason that previous studies have not resolved this overestimation issue is the fact that the fill rate is a random variable over a finite review horizon, and as a result, the problem of determining stock levels that deliver a given fill rate is analytically intractable. In this research, we study the fill rate behavior in the setting of order-up-to policy with a finite review horizon and positive lead time. More importantly, we provide a practical tool that can be readily implemented by inventory managers and/or commercial software packages. To achieve this, we first prove structural results for expected fill rate over a finite horizon that lead to upper and lower bounds. We use these bounds in a simulation-based optimization algorithm to solve the problem. Simulation-based optimization is a viable tool when facing analytically intractable models like the one presented in this study (Fu et al. 2005). Another explanation for past neglect of this overestimation problem may be due to the tactic of using overstock to avoid invoking the penalty clause in the SLA. However, we show that as long as the penalty rate is moderate, the firm still faces a serious overstocking problem because of the variability generated by the probability distribution of fill rate over a finite horizon.

Table 1 Illustration of the Results when Target Fill Rate Equals 90%

	Achieved expected fill rate	Order-up-to level, s
Traditional formula	92.73%	23.9157 (0.00%)
Simulation-based optimization	90.01%	22.9493 (−4.04%)

Notes. This result is based on the scenario when performance review horizon $T = 15$ (i.e., roughly biweekly), lead time $L = 3$, demand is distributed as Erlang (5,1) in each time period and the initial on-hand inventory is at the order-up-to level. The percentage in parentheses represents the improvement of order-up-to level.

The problem formulation, as we demonstrate later in this study, requires the computation of the expectation of a rational function of dependent random variables, which is a formidable analytical problem. The problem is further exacerbated by the underlying distributions being high dimensional and nonfactorizable. The only recourse is to compute the expectation through Monte-Carlo methods. We propose two such methods, a vanilla technique followed by a more efficient stochastic optimization. While results obtained are qualitatively similar in the two cases, the second method demonstrates faster convergence.

The operations management literature has been rather casual about tying the formula for expected fill rate to an infinite horizon, while in fact applying it under finite horizons of various lengths. Our paper subjects the implicit assumption that this abuse is innocuous to close scrutiny, and finds that it is not as innocent as has been tacitly assumed in the literature. In essence, we find that ignoring lead time and using an infinite horizon formula in a finite horizon context together conspire to inflate inventory levels significantly.

To summarize, there are several unique contributions of this study to the academic literature as well as to practice. Firstly, we investigate how the performance review horizon, lead time, and demand distribution affect the achieved fill rate in a finite horizon. Previous studies (Banerjee and Paul 2005, Chen et al. 2003, Thomas 2005, Zhong et al. 2017) have focused only on inventory systems with zero lead time. In this study, we incorporate lead time into our model, analyze it theoretically, and show empirically that lead time worsens the overstocking problem in the sense that the higher the lead time, the higher the amount of overstocking. Secondly, we analytically prove that the achieved fill rate in the finite horizon is higher than the target fill rate, which provides the theoretical foundation of our proposed algorithms. Note that this result applies when the initial on-hand inventory is a random

variable following the steady-state on-hand inventory distribution, and also when the initial on-hand inventory is equal to the order-up-to level. Further, we find that the state of the initial inventory system is critical to the magnitude of the overstocking problem. The overstocking problem is much more serious for the inventory system when the initial on-hand inventory is equal to the order-up-to level. Finally, we develop a practical tool for inventory managers to set up the optimal inventory level needed to achieve a contractually specified fill rate. Such practical tools and software could help firms in a wide range of industries achieve inventory cost savings while simultaneously providing customers the contractually committed service level.

The rest of the study is organized as follows: In the next section, we review the related literature. Section 3 describes the theoretical setting and discusses the behavior of the fill rate random variable with positive lead time. In section 4, we first prove structural properties of average fill rate over a finite horizon and then design a simulation-based algorithm based in part on the properties established earlier. In section 5, we conduct extensive numerical analysis to compare against the traditional formula and provide managerial implications. In section 6, we extend our basic model to incorporate the penalty into consideration and discuss the performance of an alternative algorithm. The study concludes with a summary and avenues for future research.

2. Literature Review

An SLA is a type of performance-based contract in which the supplier commits to achieving a specified service level over a number of time periods defined as the performance review horizon (Chen and Thomas 2015). Service level metrics can be classified into the following three categories: α-service-level, β-service-level, and γ-service-level. α-service-level, commonly known as Type 1 service level, is defined as the fraction of cycles in which there is no stockout. Ready rate, a variation of α-service-level, measures the probability that arriving customer orders will be completely delivered from stock. β-service-level, the focus of the current study, also called Type 2 service level or fill rate, denotes the expected fraction of demand served immediately from stock. Note that the key difference between Type 1 and Type 2 service level hinges on whether or not to take the magnitude of stock-out into account. Specifically, Type 1 service level is an event-oriented performance criterion which only counts the number and ignores the magnitude of stock-out events, while Type 2 service level is a quantity-oriented performance measure which captures the magnitude of stock-outs. We focus on fill rate in this study because it is the metric that most managers associate with service level (Nahmias and Olsen 2015). γ-service-level incorporates expected cumulative backorders per time period into the service level calculation, which captures the duration of the stockout. Schneider (1981) has provided a comprehensive early review of these three service measurements under different inventory policies. Silver et al. (1998) present an insightful discussion of the different service level metrics mentioned above.

In the operations management literature and textbooks, *fill rate* is defined as the average fraction of demand that can be immediately fulfilled from on-hand inventory (Axsater 2006, Cachon and Terwiesch 2008, Song 1998). Methods presented in textbooks and used in many commercial software packages often calculate fill rate as expected demand satisfied per cycle divided by expected demand per cycle. Note that this formula is equivalent to one minus the ratio of expected back-order per cycle to expected demand per cycle. Nevertheless, Chen et al. (2003) note that this formula only holds when the demand is stationary and serially independent over an infinite horizon. Assuming zero lead time, Chen et al. (2003) also show that with a common fixed stocking quantity, the expected achieved fill rate over a finite performance review horizon is always greater than the expected fill rate over an infinite performance review horizon. In a sequel, Banerjee and Paul (2005) extend the work of Chen et al. (2003)'s work by proving that the expected fill rate is monotonically decreasing in the number of review periods. Another noteworthy study in this stream of literature is Thomas (2005), which extensively studies how the achieved fill rate behaves over a range of different demand distributions and review horizons through Monte Carlo simulation. This study not only consolidates the earlier theoretical findings but also generates managerial implications. Katok et al. (2008) further explore this issue through a controlled lab experiment and finds that applying a longer performance review period is more effective in terms of inducing service improvements. In essence, the aforementioned works all suggest that the current methods lead to a higher inventory level than necessary to achieve a specified fill rate agreement over a finite review horizon. It is also worth noting that these papers share the common underlying assumption of zero lead time, which significantly simplifies the technical analysis.

There are several studies that investigate the fill rate in periodic review systems with positive lead time. Johnson et al. (1995) examine the problem of estimating the fill rate. They not only provide the details of the derivation of the classical fill rate formula but also propose an exact fill rate expression

under the normally distributed demand. Sobel (2004) derives the formulas for the fill rate under general demand distributions for both single-stage and multiple-stage supply chain systems that use base-stock policies. Zhang and Zhang (2007), Zhang et al. (2010), Zhang (2012) extend Sobel's (2004) work to the general periodic review policy in which the inventory position is reviewed once every R periods for single-stage and two-stage inventory systems. Note that if $R = 1$, the general periodic review policy is equivalent to the traditional periodic-review order up to policy. In a follow-up study, Teunter (2009) derived the same expression for the fill rate in Zhang and Zhang (2007) using an alternative approach and generalized to (R, Q) policies. Guijarro et al. (2012) develop a general method to compute the fill rate for discrete demand distribution under the setting of lost sales. Paul et al. (2015) have studied the inventory planning problem for modular products with individual and aggregate fill rate constraints. The focus of the aforementioned works is to characterize the fill rate in a single or multistage inventory system over an infinite review horizon, while the review of the inventory system is often conducted in a finite performance horizon.

Thus, there is a clear gap in the extant literature; fill rate over a finite horizon with positive lead time, which is what transpires in practice, has not been well studied. In this study, we first look into the impact of the interaction of positive lead time and finite review horizon. After observing the behavior of the fill rate distribution, we analytically show that expected fill rate over a finite review horizon is always greater than the expected fill rate over an infinite review horizon, which complements the previous literature (Chen et al. 2003, Thomas 2005). Correspondingly, we develop a simulation-based optimization algorithm to help inventory managers set up the optimal inventory level, which is crucial but overlooked by the previous literature. Simulation-based optimization has been developed rapidly in the past few years due to the advances in computing power and memory. This approach provides solutions to many important practical problems previously beyond reach. A number of applications have been published in the literature. Glasserman and Tayur (1995) develop a simulation-based method to estimate the sensitivities of inventory costs with respect to the policy parameters in a multiechelon inventory systems. Kapuscinski and Tayur (1998) study a capacitated production inventory system. Spieckermann et al. (2000) apply simulation optimization to optimize an automobile manufacturing production system. Readers interested in simulation optimization may refer to Fu et al. (2005) for a comprehensive review of the current literature.

3. Problem Definition and Fill Rate Distribution

In this section, we first introduce the notations and settings of our study, and then make observations on the distribution of fill rate through the lens of Monte Carlo simulation studies.

3.1. Problem Setting

We study fill rate over a finite horizon in a periodic review model with positive lead time. Although the widespread implementation of sophisticated information systems makes continuous review systems practical, replenishment goods are shipped only periodically in many settings. Therefore, the shipment dates determine the points in time at which orders are placed (Sobel 2004). It is well known that an order-up-to policy (i.e., base stock policy) will minimize holding and shortage costs when the horizon is infinite (Zipkin 2000), but such a policy is not necessarily optimal over a finite horizon under a service-level agreement. We restrict our space of inventory policies to the space of stationary, order-up-to policies, since these policies are commonly used and easy to implement.

Let OH_t represent the on-hand inventory level at the start of period t. D_t denotes the demand random variable in period t; OO_t denotes the on-order inventory of period t, BO_t denotes total back-orders at the end period t, and IN_t denotes net inventory, which is defined as the difference between on-hand inventory and back-orders. Depending on the scenario, the initial on-hand inventory level, OH_1 can take the form of either the order-up-to level, $OH_1 = s$ (i.e., initial state hereafter) or $OH_1 = (s - \Sigma_{i=1-L}^{0} D_i)^+$ (i.e., steady-state hereafter).[3] On the one hand, if the inventory manager is dealing with a relatively new product or a newly signed contract, then it is reasonable to assume that the on-hand inventory level starts at the order-up-to level. On the other hand, if the inventory manager is facing a recurring fill rate contact, then it is appropriate to assume that the initial on-hand inventory is equal to the steady-state on-hand inventory level. For convenience, we refer to the fill rate under the first scenario as the *initial state fill rate* and the latter as the *steady-state fill rate* for the remainder of the paper.

The chronology of events is as follows. At the beginning of period t, the replenishment order placed in period $t - L - 1$ (where lead time L is a non-negative integer) arrives, then random demand is observed and satisfied from on-hand inventory. If demand exceeds on-hand inventory, excess demand is backlogged. Finally, a replenishment order is submitted to bring the inventory position (i.e., total

quantity on-order + net inventory) back up to the order-up-to level s. Note that the replenishment order quantity always equals the immediately preceding demand. As a result, we can rewrite OO_t as $\Sigma_{i=t-L}^{t-1} D_i$. Let $(x)^+$ denote $\max(x, 0)$. According to the results on page 180 of Zipkin (2000), we have

$$OH_t = (IN_t)^+ = (s - OO_t)^+ = (s - \sum_{i=t-L}^{t-1} D_i)^+. \quad (1)$$

The fill rate random variable with review horizon T and lead time L, $\beta_{T,L}(s)$, is defined as the average fraction of demand that can be satisfied from stock immediately. Hence,

$$\beta_{T,L}(s) \equiv E\left[\frac{\sum_{i=1}^{T} min(OH_i, D_i)}{\sum_{i=1}^{T} D_i}\right] \quad (2)$$

Sobel (2004) has derived the analytical solution to Equation (2) when the performance review horizon is infinite (i.e., $T \to \infty$) for arbitrary demand distributions. The following formula gives us the benchmark to compare against when the performance review horizon is finite.

$$\begin{aligned}\beta_{\infty,L}(s) &= 1 - \frac{E[D_{L+1} - (s - \sum_{i=1}^{L} D_i)^+]^+}{\mu} \\ &= \int_0^s [G^{(L)}(b) - G^{(L+1)}(b)]db/\mu\end{aligned} \quad (3)$$

where $G(\cdot)$ denotes the demand distribution function with finite expectation μ and $G^{(k)}(\cdot)$ stands for the k-fold convolution of $G(\cdot)$. Based on formula (3), we can derive the closed-form formula for average fill rate for arbitrary demand distribution. If lead time $L = 0$, formula (3) corresponds to the well-known result used to define fill rate in textbooks; that is, the fill rate equals one minus the average fraction of backordered demand over the expected demand in one period. For expositional purposes, we derive the closed-form formula for Gamma distributed demand. Let the D_t's be independent and identically distributed (i.i.d.) random variables drawn from a Gamma distribution with shape parameters γ and rate λ, so that the probability density function is given by $g(x) = \frac{\lambda^\gamma}{\Gamma(\gamma)} x^{\gamma-1} e^{-\lambda x}$. Also, the L-fold convolution of Gamma distribution is $G^{(L)}(x) = 1 - e^{-\lambda x} \Sigma_{j=0}^{L\gamma-1} \frac{(\lambda x)^j}{j!}$. As a result,

$$\beta_{\infty,L}(s) = \sum_{L\gamma+1}^{(L+1)\gamma+1} \frac{P(\Gamma(j,\lambda) \le s)}{\gamma} \quad (4)$$

Having described the inventory system and derived the fill rate formula with a positive lead time,[4] we next explore the behavior of the fill rate distribution in a finite performance review horizon. We characterize the order-up-to level s from Equation (4) and use it as the benchmark to study the fill rate in a finite review horizon by Equation (2).

3.2. Fill Rate Distribution

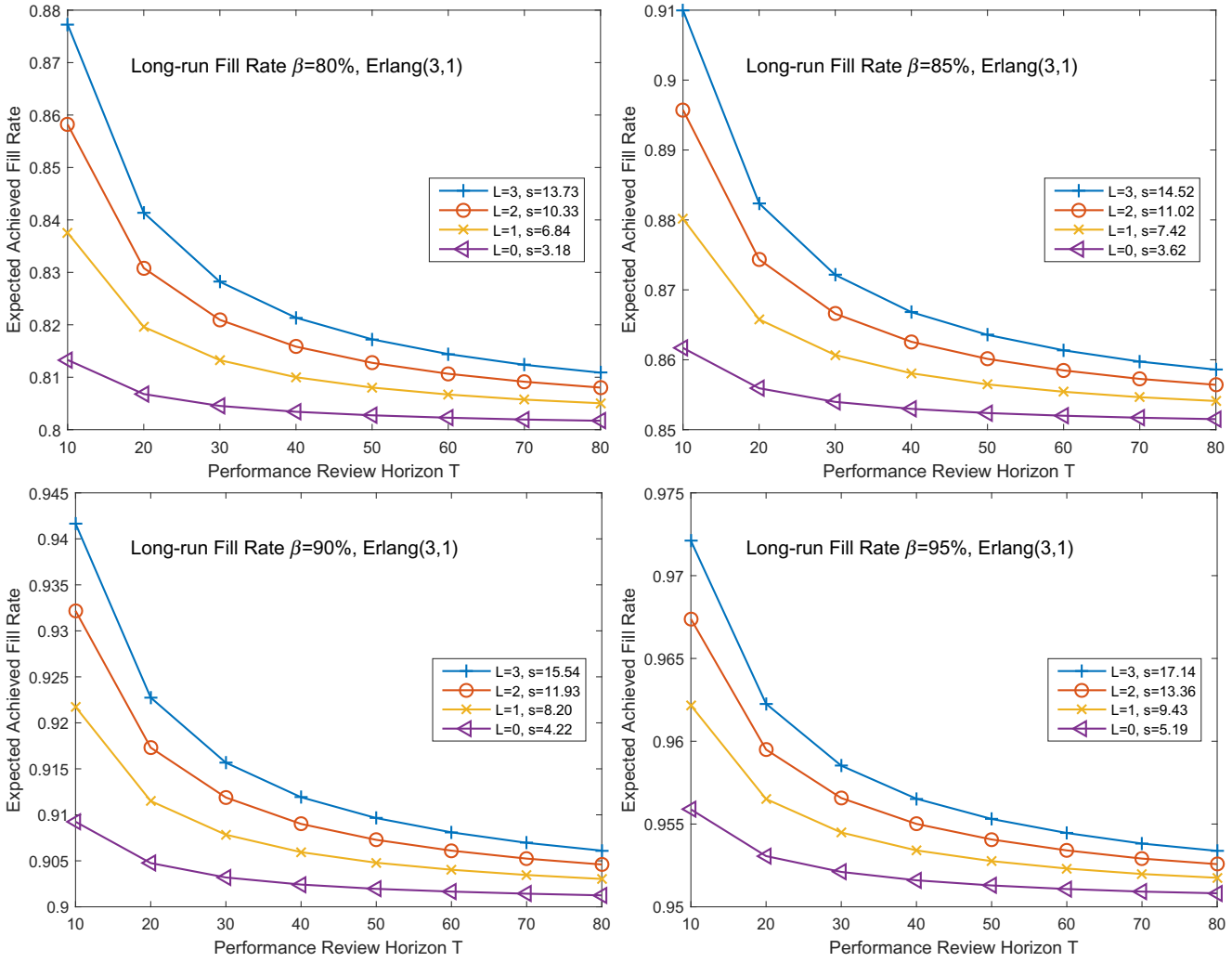
We now examine how the review horizon, lead time, and different demand distributions affect the behavior of the random variable $\beta_{T,L}$. This will not only reveal the pitfalls of the current method to determine the order-up-to level with fill rate constraint, but also facilitate the subsequent algorithm development. As we pointed out earlier, analytically characterizing $\beta_{T,L}$ appears to be an intractable task when the performance review horizon T is finite and the lead time L is positive. Instead, motivated by Thomas (2005), we estimate the distribution of $\beta_{T,L}$, using Monte Carlo simulation. Our parameters represent a wide range of plausible values, chosen to represent realistic scenarios from current industry practice. For example, $T = 4$ and $L = 1$ stands for the situation where the inventory manager makes weekly stocking decisions with a monthly performance review, and the time interval between placing and receiving an order is one week. We have varied different distributions (i.e., Erlang, Normal, and Poisson) with different parameter settings of our numerical analysis and find that our results are robust to a variety of settings.[5] Further, the numerical results for the steady-state scenario are qualitatively similar to those for the initial state scenario. Without loss of generality, we next illustrate the details when the demand is Erlang distributed for the initial state scenario.

We itemize the steps involved in conducting the simulation study,

1. Firstly, we assume that demand is Erlang distributed and set the target fill rate $\beta = 80\%$, 85%, 90%, and 95%. Specifically, the shape parameter of the Erlang distribution is set at 3 with a rate parameter equal to 1. We also vary the lead time L ranging over the values 0, 1, 2, and 3, respectively.
2. Secondly, we characterize the order-up-to level s based on the specified long-run fill rate by solving Equation (4).
3. Thirdly, we generate random variables and compute empirical fill-rate for different review horizons T based on the inventory dynamics and fill rate definition in Equation (2). Following Thomas (2005), we replicate each experiment 10^7 times to get an accurate estimation of the fill rate distribution in the finite performance review horizon.

We summarize the highlights of our simulation study in Figure 1. Figure 1 shows the average of the achieved fill rate with various long-run fill rate

Figure 1 Expected Achieved Fill Rate of Meeting a 80%, 85%, 90%, and 95% Target Fill Rate with Erlang Distribution (3,1) [Color figure can be viewed at wileyonlinelibrary.com]

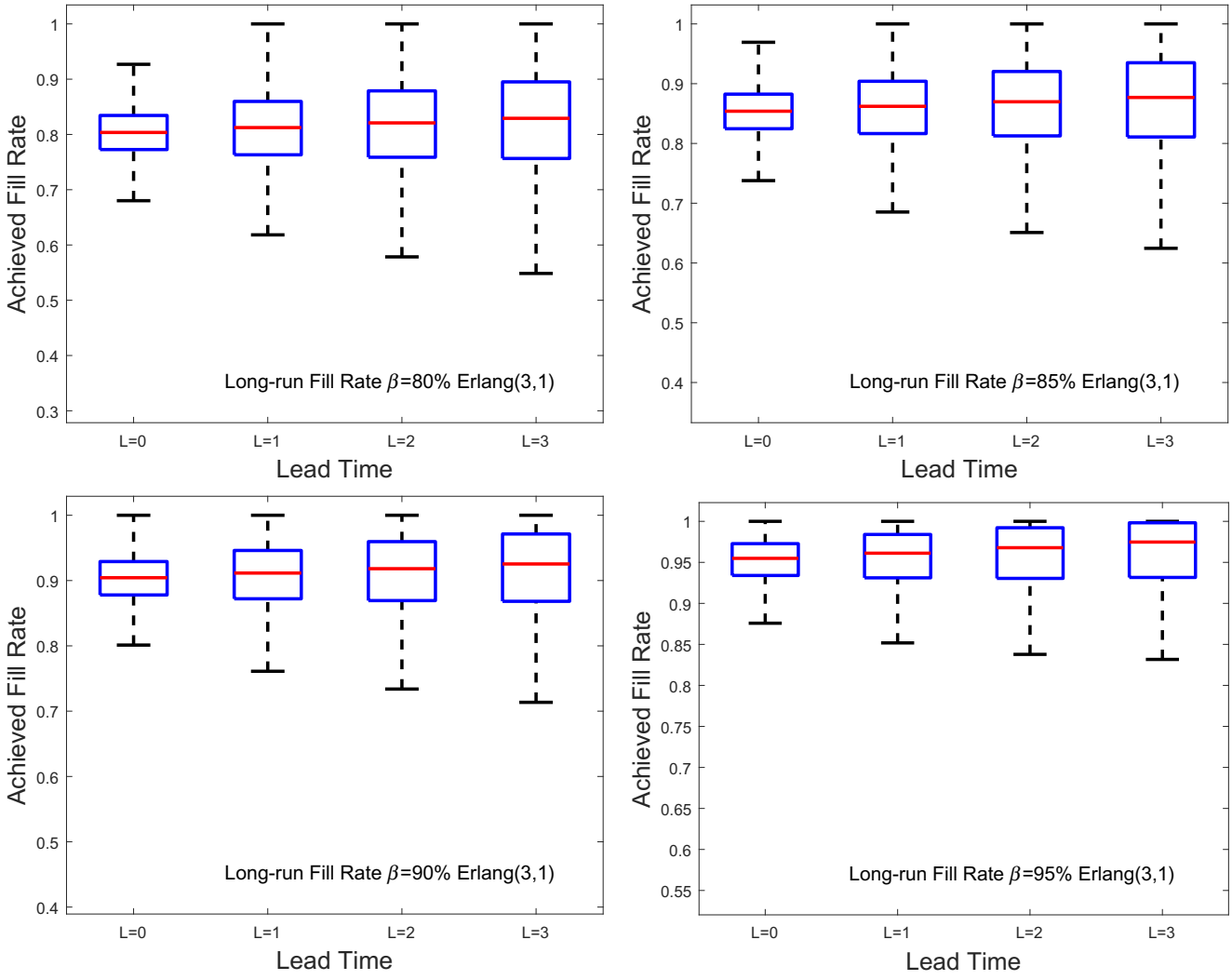


targets. By comparing the four sub-figures in Figure 1, we first find that the expected achieved fill rate is always above the long-run fill rate target and gets closer to the target when the performance review horizon T becomes larger. Further, we observe that the overestimation is more severe when the target fill rate is low and/or the performance review horizon is short. For example, if an inventory manager employs the traditional fill rate formula to maintain an 80% fill rate in a daily stocking decision with weekly performance review (i.e., $T = 10$), what she/he actually achieves is a fill rate of 88% when the lead time L is three days. Another very important observation from Figure 1 is that the achieved fill rate varies systematically with lead time L. This observation is new to the literature, as previous studies have not incorporated the lead time (Banerjee and Paul 2005, Thomas 2005) in the fill rate over the finite review horizon. From the figure, we observe that the higher the lead time is, the higher the expected achieved fill rate will be, which suggests a more severe overstocking problem. To further examine the impact of lead time on fill rate distribution, we conduct further analysis to isolate the lead time by allowing the lead time to vary while fixing the other parameters.

We plot the boxplot of the fill rate distribution with different lead times (L = 0, 1, 2, and 3) and different target fill rate while keeping all other variables fixed in Figure 2.

Immediately, we notice that the fill rate distribution is very sensitive to the change in lead time. A close examination of the boxplots shows that when the lead time is relatively large, the median and upper quartile of the fill rate distribution both increase, while the lower quartile and the minimum both decrease as lead time increases; as a result, the fill rate distribution as a whole spreads out as lead time increases. This suggests that both the magnitude and the

Figure 2 **Boxplot of Achieved Fill Rate of Meeting a 80%, 85%, 90%, and 95% Target Fill Rate with Erlang Distribution (3,1) and Performance Review Horizon $T = 40$ [Color figure can be viewed at wileyonlinelibrary.com]**



variability of the fill rate distribution increases as lead time increases. Thus, we conclude that the overestimation problem (i.e., the achieved fill rate is higher than the target fill rate) is more severe than previously thought as the lead time plays a significant role in aggravating this problem.

4. Theoretical Results and a Simulation-based Optimization Algorithm

In this section, we show how to characterize the order-up-to level subject to a target fill rate. The underlying mechanism is as follows. Because of the monotonicity of the expected fill rate in the order-up-to level, we are able to search the optimal order-up-to level from the lower bound to the upper bound. Due to the complexity of the fill rate distribution behavior, we resort to simulation to calculate the expected achieved fill rate.

In the next sub-section, we prove that given an order-up-to level, s, the achieved fill rate in a finite horizon is greater than or equal to the fill rate in an infinite review horizon for both steady state and initial state fill rate. This theoretical result not only consolidates our finding in the previous section, but also establishes the upper bound of the search region. Essentially, we prove that this upper bound is equal to the order-up-to level characterized from the traditional formula assuming the review horizon is infinite.

4.1. Theoretical Results

Here, we prove our main structural result $\beta_{T,L}(s) \geq \beta_{\infty,L}(s)$. The immediate implication of this result is as follows: If we use the optimal order-up-to level assuming an infinite horizon, then our actual achieved fill rate will become higher than the target level. Thus, this provides us an upper bound of our

search region in the later algorithm. Note that the difficulty in proving this result stems from the positive lead time, which was assumed to be zero in the previous literature (Banerjee and Paul 2005, Chen et al. 2003). Relaxing the assumption of zero lead time requires a novel approach to proving the result, as we will show.

Let $\{D_i: i = 1, 2, \ldots\}$ be independent and identically distributed (i.i.d.) non-negative demand random variables. Following Chen et al. (2003) and without loss of generality, we assume $E[D_i] = 1$. We focus on the discussion of the realistic case where $T \geq L + 1$; the proof for the case $T \leq L$ follows along the same lines.

There are two different ways of defining fill rate over a multiple-period horizon, depending on whether the initial on-hand inventory level is assumed to have attained a steady-state distribution or not. We will prove that the result $\beta_{T,L}(s) \geq \beta_{\infty,L}(s)$ holds true, using either definition. On the one hand, the steady-state definition is appropriate when the horizon over which fill rate is measured consists of periods $L + 1$ through $L + T$; periods 1 through L are not accounted for in the measurement of average fill rate. Alternatively, it applies when we transition from one T-period horizon to another T-period horizon with the same demand distributions; in this case, the leftover inventory from the first horizon serves as the starting on-hand inventory for the second horizon. On the other hand, the initial state definition applies when average fill rate is measured over periods 1 though T; the system is started "cold" with an on-hand inventory level equal to the order-up-to level. The initial state definition of average fill rate accommodates a transient effect, and this may better capture the true fill rate when the inventory system begins operations from scratch. For instance, consider a situation in which a firm plans to sell a product over a horizon of 3 months with weekly replenishments; the firm enters into a service-level agreement for this selling season. That is to say that $T = 12$ weeks and $L = 1$ week. In the next selling season of 3 months, demand conditions may change and the product itself may be upgraded. The firm signs a fresh SLA for the next 3 months. In this example, we suggest that the initial state definition of fill rate is more appropriate than the steady-state definition.

Initial state case: We now commence the proof of the theorem $\beta_{T,L}(s) \geq \beta_{\infty,L}(s)$ for the initial state case. We first prove the result using the initial state definition of average fill rate, after which we prove it using the steady-state definition. We remark that the proof using the steady-state definition is similar to the first proof, but requires some finessing; We provide complete details of both proofs.

From the initial state definition of fill rate, we have

$$\begin{aligned}\beta_{T,L}(s) &= E\left[\frac{\min(D_1, s)}{\sum_{i=1}^{T} D_i}\right] \\ &\quad + \sum_{k=2}^{L} E\left[\frac{\min\left(D_k, \left(s - \sum_{i=1}^{k-1} D_i\right)^+\right)}{\sum_{i=1}^{T} D_i}\right] \\ &\quad + \sum_{k=L+1}^{T} E\left[\frac{\min\left(D_k, \left(s - \sum_{i=1}^{L} D_{k-i}\right)^+\right)}{\sum_{i=1}^{T} D_i}\right] \\ &= E\left[\frac{\min(D_1, s)}{\sum_{i=1}^{T} D_i}\right] \\ &\quad + \sum_{k=2}^{L} E\left[\frac{\min\left(D_k, \left(s - \sum_{i=1}^{k-1} D_i\right)^+\right)}{\sum_{i=1}^{T} D_i}\right] \\ &\quad + (T - L)E\left[\frac{\min\left(D_{L+1}, \left(s - \sum_{i=1}^{L} D_i\right)^+\right)}{\sum_{i=1}^{T} D_i}\right],\end{aligned} \tag{5}$$

where the second equation holds because D_i are i.i.d. for $i = 1, 2, \ldots, T$.

The following lemma provides the average fill rate under infinite planning horizon.

LEMMA 1. *When $E[D_i] = 1$, we have*

$$\begin{aligned}\beta_{\infty,L}(s) &:= \lim_{T \to +\infty} \beta_{T,L}(s) \\ &= E\left[\min\left(D_{L+1}, \left(s - \sum_{i=1}^{L} D_i\right)^+\right)\right].\end{aligned} \tag{6}$$

PROOF. The proof is given in Online Appendix S1.□

In the following theorem, we show that for a given based stock level s, the average fill rate under the finite review horizon is no lower than that under the infinite time horizon. Accordingly, to reach a certain fill rate level β, the based stock level under the finite time horizon $s_{T,L}(\beta)$ is no higher than that under the infinite time horizon $s_{\infty,L}(\beta)$.

THEOREM 1. $\beta_{T,L}(s) \geq \beta_{\infty,L}(s)$.

Without loss of generality, we assume $E[D_i] = 1$ for $i = 1, 2, \ldots$. Before proceeding to the proof of the theorem, we first state several lemmas that we shall use in the proof.

LEMMA 2. *When* $E[D_i] = 1$ *for* $i = 1, 2, \ldots$, *there exists a* $v_L > 0$ *such that*

$$E\left[\left(\frac{1}{z+\sum_{i=L+2}^{T} D_i} - \frac{1}{T}\right)\right] \begin{cases} > 0 & \text{for } 0 \le z < v_L \\ = 0 & \text{for } z = v_L \\ < 0 & \text{for } z > v_L. \end{cases} \quad (7)$$

PROOF. The proof is given in Online Appendix S2.□

Define

$$h_L(s) := E\left[\left(\frac{1}{\sum_{i=1}^{T} D_i} - \frac{1}{T}\right) \min\left(D_{L+1}, \left(s - \sum_{i=1}^{L} D_i\right)^+\right)\right]. \quad (8)$$

LEMMA 3. *Suppose* D_i *has absolutely continuous distribution with density function* $f(y) > 0$ *for all* $y > 0$ *and* $E[D_i] = 1$ *for* $i = 1, 2, \ldots$, *we have*

$$h_L(s) \ge 0 \text{ for } s \ge v_L, \quad (9)$$

where v_L *is defined in Lemma 2.*

PROOF. The proof is given in Online Appendix S3.□

Define

$$g(s,x) = E\left[\left(\frac{1}{x + D_{L+1} + \sum_{i=L+2}^{T} D_i} - \frac{1}{T}\right) \min(D_{L+1}, s - x)\right]. \quad (10)$$

We have

$$g(s,x) = \int_0^{s-x} E\left[\left(\frac{1}{x + y + \sum_{i=L+2}^{T} D_i} - \frac{1}{T}\right) y\right] f(y)dy + \int_{s-x}^{\infty} E\left[\left(\frac{1}{x + y + \sum_{i=L+2}^{T} D_i} - \frac{1}{T}\right)(s-x)\right] f(y)dy. \quad (11)$$

In the following lemma, we establish some properties of $g(s, x)$ when $s = v_L$, where v_L is defined in Lemma 2.

LEMMA 4. *Suppose* D_i *has absolutely continuous distribution with density function* $f(y) > 0$ *for all* $y > 0$ *and* $E[D_i] = 1$ *for* $i = 1, 2, \ldots$, *there exists a* $u_L \in (0, v_L)$ *such that*

$$g(v_L, x) \begin{cases} > 0 & \text{for } 0 \le x < u_L \\ = 0 & \text{for } x = u_L \\ < 0 & \text{for } u_L < x < v_L \\ = 0 & \text{for } x = v_L \end{cases}, \quad (12)$$

where v_L *is defined in Lemma 2.*

PROOF. The proof is given in Online Appendix S4.□

LEMMA 5. *Suppose* D_i *has absolutely continuous distribution with density function* $f(y) > 0$ *for all* $y > 0$ *and* $E[D_i] = 1$ *for* $i = 1, 2, \ldots$, *we have*

$$h_L(s) \ge 0 \text{ for } 0 \le s < v_L, \quad (13)$$

where v_L *is defined in Lemma 2.*

PROOF. The proof is given in Online Appendix S5.□

We now proceed to the proof of the theorem.

PROOF OF THEOREM 1. Since D_i are non-negative and i.i.d. for $i = 1, 2, \ldots, T$, we have

$$\begin{aligned} E\left[\frac{\min(D_1, s)}{\sum_{i=1}^{T} D_i}\right] &= E\left[\frac{\min(D_2, s)}{\sum_{i=1}^{T} D_i}\right] \\ &\ge E\left[\frac{\min\left(D_2, (s - D_1)^+\right)}{\sum_{i=1}^{T} D_i}\right] \\ &\ge \ldots \\ &\ge E\left[\frac{\min\left(D_{L+1}, \left(s - \sum_{i=1}^{L} D_i\right)^+\right)}{\sum_{i=1}^{T} D_i}\right]. \end{aligned}$$

Hence, based on Equation (5), we have

$$\beta_{T,L}(s) \ge TE\left[\frac{\min\left(D_{L+1}, \left(s - \sum_{i=1}^{L} D_i\right)^+\right)}{\sum_{i=1}^{T} D_i}\right]. \quad (14)$$

Thus, in order to prove $\beta_{T,L}(s) \ge \beta_{\infty,L}(s)$, based on Equations (6) and (14), it suffices to prove

$$TE\left[\frac{\min\left(D_{L+1}, \left(s - \sum_{i=1}^{L} D_i\right)^+\right)}{\sum_{i=1}^{T} D_i}\right] \ge E\left[\min\left(D_{L+1}, \left(s - \sum_{i=1}^{L} D_i\right)^+\right)\right]. \quad (15)$$

□

Based on Equation (8), we have $h_L(s) = E[(\frac{1}{\Sigma_{i=1}^{T} D_i} - \frac{1}{T}) \cdot \min(D_{L+1}, (s - \Sigma_{i=1}^{L} D_i)^+)]$. Hence, it suffices to prove

$$h_L(s) \geq 0 \text{ for all } s \geq 0. \tag{16}$$

We first assume that D_i has absolutely continuous distribution with density function $f(y) > 0$ for all $y > 0$. Under this assumption, $h_L(s) \geq 0$ for $s \geq v_L$ is proved in Lemma 3 and $h_L(s) \geq 0$ for $0 \leq s < v_L$ is proved in Lemma 5. Thus, we have $h_L(s) \geq 0$ for all $s \geq 0$ and our conclusion holds.

Next, we discuss the general case where the distribution of D_i might not be absolutely continuous. In this case, let $D_i^* = \frac{D_i + \delta\varepsilon_i}{1+\delta}$ for $i = 1,2,\ldots$, where δ is a positive constant, and ε_i are i.i.d. exponential random variables with mean 1 and are independent of D_i. Hence, D_i^* are absolutely continuous random variables with positive density functions and $E[D_i^*] = 1$, and so the above proof applies to D_i^*. Letting $\delta \to 0$, the claim holds for the general case.

Steady-state case: Next, we prove that the previous theorem also holds under the steady-state definition of fill rate. The focal difference between the steady-state and initial state definition of fill rate lies in the initial on-hand inventory level. Let D_{-L+1}, D_{-L+2}, …, D_{-2}, D_{-1}, D_0 be i.i.d. copies of D_i. Under the steady state, the average fill rate over periods 1, 2, …, T is

$$\begin{aligned}
\tilde{\beta}_{T,L}(s) &= \sum_{k=1}^{T} E\left[\frac{\min\left(D_k, \left(s - \sum_{i=k-L}^{k-1} D_i\right)^+\right)}{\sum_{i=1}^{T} D_i}\right] \\
&= \sum_{k=1}^{L} E\left[\frac{\min\left(D_k, \left(s - \sum_{i=k-L}^{k-1} D_i\right)^+\right)}{\sum_{i=1}^{T} D_i}\right] \\
&\quad + \sum_{k=L+1}^{T} E\left[\frac{\min\left(D_k, \left(s - \sum_{i=k-L}^{k-1} D_i\right)^+\right)}{\sum_{i=1}^{T} D_i}\right] \\
&= \sum_{k=1}^{L} E\left[\frac{\min\left(D_k, \left(s - \sum_{i=k-L}^{0} D_i - \sum_{i=1}^{k-1} D_i\right)^+\right)}{\sum_{i=1}^{T} D_i}\right] \\
&\quad + (T-L)E\left[\frac{\min\left(D_{L+1}, \left(s - \sum_{i=1}^{L} D_i\right)^+\right)}{\sum_{i=1}^{T} D_i}\right],
\end{aligned} \tag{17}$$

where the second equation holds because D_i are i.i.d. for $i = 1, 2, \ldots, T$. These terms involve random variables D_t from the previous history $t = \ldots, -3, -2, -1$. It is straightforward to see that this is equivalent to measuring the average fill rate over periods L through $L + T$ and ignoring periods 1 through L for measurement.

Lemma 6. *When $E[D_i] = 1$, we have*

$$\begin{aligned}
\tilde{\beta}_{\infty,L}(s) &:= \lim_{T \to +\infty} \tilde{\beta}_{T,L}(s) \\
&= E\left[\min\left(D_{L+1}, \left(s - \sum_{i=1}^{L} D_i\right)^+\right)\right] = \beta_{\infty,L}(s).
\end{aligned} \tag{18}$$

Proof. The proof is similar to that of Lemma 1 and so is omitted. The crucial ingredient of the proof is the dominated convergence theorem. □

Theorem 2. $\tilde{\beta}_{T,L}(s) \geq \tilde{\beta}_{\infty,L}(s)$.

Proof. The proof is given in Online Appendix S6.□

From the proof of Theorems 1 and 1, it is straight forward to show that our results also hold when the fill rate is counted over periods $[t, t-1+T]$ for $t = 2, 3, L$. The detailed discussion is provided in Online Appendix S7.

Based on Theorems 1 and 2, we prove that given an order-up-to level, s, the achieved fill rate in a finite horizon is greater than or equal to the fill rate in an in finite review horizon for both steady-state and initial state fill rate. This theoretical result defines the upper bound of the search region of the proposed algorithm in section 4.2. That is, we prove that this upper bound is equal to the order-up-to level characterized from the traditional formula assuming the review horizon is infinite. Next we prove the lower bound of the search region of the order-up-to level.

Here we provide an upper bound for both $\tilde{\beta}_{T,L}(s)$ and $\beta_{T,L}(s)$, which we use to provide the lower bound of our search region of the order-up-to level.

Theorem 3. $\tilde{\beta}_{T,L}(s) \leq \beta_{T,L}(s) \leq \beta_{1,0}(s)$.

Proof. The proof is given in Online Appendix S8. Based on Theorem 3, it is straightforward to observe that the order-up-to level s in the $\beta_{1,0}(s)$ can be used to provide a lower bound in searching the optimal order-up-to level, s. □

4.2. A Simulation-based Optimization Algorithm

Based on our previous theoretical results, we present a simulation based optimization algorithm to search

for the optimal order-up-to quantity and characterize the convergence properties of the algorithm. To facilitate the development, we introduce some notations. First of all, for simplicity we omit the subscript L and use $\beta_{T,L}$ and β_T interchangeably. Let $Z^N = \{Z_n : n = 1, 2, \ldots, N\}$ be the i.i.d. samples of the random demand sequence, where each $Z_n = \langle D_1^{(n)}, D_2^{(n)}, \ldots, D_T^{(n)}\rangle$ is the i-th sample of demand sequence in T periods. For each period t: $1 \le t \le T$, let $OH_t^{(n)}(s)$ be the on hand inventory, $BO_t^{(n)}(s)$ be the back order level and $OO_t^{(n)}(s)$ be the on order inventory, associated with the order-up-to level s, respectively. $\hat{\beta}_T^Z$ indicates the empirical fill rate computed from sample path Z, and we let $\hat{\beta}_T^{(n)}$ be short for $\hat{\beta}_T^{Z_n}$. Replacing the population mean with the sample mean, we seek a solution of the approximated problem as follows

$$\frac{1}{N}\sum_{n=1}^{N}\hat{\beta}_T^{(n)}(s) = \beta. \tag{19}$$

To begin with, we present some useful properties of fill rate in the following proposition.

PROPOSITION 1. *Let $Z = \{\langle D_1, D_2, \ldots, D_T\rangle\}$ be any random demand sequence. $\hat{\beta}_T^{(Z)}(s)$ is the empirical fill rate computed from the sequence Z. For each $t = 1, 2, \ldots, T$; OH_t is a monotonically increasing function of s; BO_t is a monotonically decreasing function of s. As a result, $\hat{\beta}_T^{(Z)}(s)$ is a monotonically increasing function of s.*

PROOF. From the definition in the previous section, we can rewrite $BO_t(s) = \max(D_{t-1} - D_{t-L-1} + OO_{t-1} - s, 0)$. The monotonicity of BO_t, OH_t is an immediate consequence. □

It can be observed from Proposition 1 that $\beta_T(s) = E_Z[\hat{\beta}_T^Z(s)]$ is a monotonically increasing function of the order-up-to level s. In other words, we know the range and the search direction of the optimal order-up-to level s^*. Based on the theoretical development established above, we develop a bisection method based on Monte Carlo simulation, to find the optimal order-up-to level s in the following Algorithm. The algorithm starts by finding an upper bound s_u through repeated doubling until a value is found such that the expected fill rate is larger than that specified. A lower bound s_l is similarly obtained by repeated shrinkage. The algorithm then repeatedly divides and updates the interval $[s_l, s_u]$, selecting the subinterval containing the optimal value. Monotonicity of $\beta_T(s)$ guarantees the uniqueness of the solution and its global optimality.

To start the bisection method, we must find the appropriate lower bound and upper bound for choosing the initial point. A naive method is to set zero as lower bound and a sufficiently large number as upper bound. However, setting a fixed initial point does not take into account the problem context, in particular, the distribution structure. Thus, it may be inefficient, due to the wide range of optimal values given demand distribution parameters. In contrast, our theoretical results suggest a better way to initialize the bisection algorithm. Based on Theorem 3, it is straightforward to see that the order-up-to level in the $\beta_{1,0}(s)$ can be used to provide a lower bound, and we can use the order-up-to level in the $\beta_{\infty,L}(s)$ to provide an upper bound. Due to the page limit, we discuss the simulation results and running time with various kinds of initialization in the appendix. Essentially, we find that the computation time of our algorithm is greatly reduced by adopting the initialization informed by our theoretical analysis.

5. Numerical Results

In this section, we illustrate and discuss the outcomes of the algorithm described in the previous section and compare the order-up-to level, s, of the proposed algorithm to that given by the traditional formula. All models and algorithms in this section are implemented in MATLAB R2015b, on a Thinkpad workstation with Intel Xeon E3-1505M CPU 2.80 GHz and 16GB of memory.

In the numerical experiments, we illustrate multiple demand distributions with different parameter settings. For expositional purposes, we report the results from the Erlang distribution and leave the results of other distributions in the appendix as the insights are qualitatively similar across different distributions. Specifically, we choose Erlang (3,1) as the demand distribution, vary the lead time L from 0 to 3, the target fill rate from 75% to 95%, and the performance review horizon from $T = 10$ to $T = 60$. We also report the order-up-to level from the traditional procedure for comparison. The results are summarized in Tables 2 and 3 at the end of this document, where Table 2 shows the results from the initial state and the Table 3 illustrates the results from the steady state. We summarize several important observations from the numerical experimental results below.

One interesting observation is that the inventory savings are smaller when the steady-state definition of fill rate is used, as opposed to the initial state definition. That is to say that the traditional method performs much better under the steady state than the initial state fill rate. This finding may be explained as follows. It is clear that the traditional procedure is exactly correct when the horizon length is infinite and

Table 2 The Comparison of Order-up-to Level between Traditional Formula and Proposed Algorithm for Initial State with Erlang (3,1)

Lead time	Target fill rate	Performance review horizon						
		∞	10	20	30	40	50	60
$L = 0$	75%	2.824	2.735 (3.17%)	2.779 (1.61%)	2.794 (1.07%)	2.802 (0.78%)	2.806 (0.63%)	2.809 (0.54%)
	80%	3.179	3.079 (3.15%)	3.127 (1.61%)	3.145 (1.07%)	3.154 (0.78%)	3.159 (0.63%)	3.162 (0.54%)
	85%	3.619	3.506 (3.12%)	3.561 (1.61%)	3.580 (1.07%)	3.591 (0.78%)	3.596 (0.63%)	3.600 (0.54%)
	90%	4.215	4.086 (3.08%)	4.149 (1.56%)	4.170 (1.07%)	4.182 (0.78%)	4.189 (0.63%)	4.193 (0.54%)
	95%	5.186	5.024 (3.13%)	5.105 (1.56%)	5.131 (1.07%)	5.146 (0.78%)	5.151 (0.68%)	5.156 (0.59%)
$L = 1$	75%	6.364	5.960 (6.35%)	6.160 (3.20%)	6.227 (2.15%)	6.263 (1.59%)	6.283 (1.27%)	6.295 (1.07%)
	80%	6.841	6.430 (6.01%)	6.634 (3.03%)	6.702 (2.03%)	6.739 (1.49%)	6.759 (1.20%)	6.770 (1.03%)
	85%	7.423	7.004 (5.64%)	7.209 (2.88%)	7.282 (1.90%)	7.318 (1.42%)	7.340 (1.12%)	7.350 (0.98%)
	90%	8.196	7.764 (5.27%)	7.976 (2.69%)	8.048 (1.81%)	8.088 (1.32%)	8.108 (1.07%)	8.124 (0.88%)
	95%	9.426	8.970 (4.83%)	9.191 (2.49%)	9.270 (1.66%)	9.311 (1.22%)	9.334 (0.98%)	9.348 (0.83%)
$L = 2$	75%	9.757	8.924 (8.53%)	9.347 (4.20%)	9.485 (2.78%)	9.557 (2.05%)	9.596 (1.65%)	9.621 (1.39%)
	80%	10.328	9.506 (7.96%)	9.922 (3.93%)	10.058 (2.61%)	10.129 (1.93%)	10.169 (1.54%)	10.194 (1.29%)
	85%	11.019	10.204 (7.40%)	10.616 (3.66%)	10.750 (2.44%)	10.820 (1.81%)	10.861 (1.44%)	10.885 (1.22%)
	90%	11.929	11.120 (6.79%)	11.522 (3.42%)	11.661 (2.25%)	11.731 (1.66%)	11.772 (1.32%)	11.795 (1.12%)
	95%	13.360	12.545 (6.10%)	12.949 (3.08%)	13.086 (2.05%)	13.158 (1.51%)	13.197 (1.22%)	13.223 (1.03%)
$L = 3$	75%	13.082	11.701 (10.56%)	12.426 (5.02%)	12.651 (3.30%)	12.764 (2.43%)	12.828 (1.94%)	12.870 (1.62%)
	80%	13.733	12.390 (9.78%)	13.092 (4.66%)	13.310 (3.08%)	13.421 (2.27%)	13.485 (1.81%)	13.525 (1.51%)
	85%	14.516	13.208 (9.01%)	13.885 (4.35%)	14.102 (2.86%)	14.208 (2.12%)	14.272 (1.68%)	14.311 (1.42%)
	90%	15.541	14.266 (8.20%)	14.919 (4.00%)	15.131 (2.64%)	15.238 (1.95%)	15.298 (1.56%)	15.336 (1.32%)
	95%	17.142	15.895 (7.28%)	16.522 (3.61%)	16.740 (2.34%)	16.840 (1.76%)	16.899 (1.42%)	16.941 (1.17%)

Table 3 The Comparison of Order-up-to Level between Traditional Formula and Proposed Algorithm for Steady State with Erlang (3,1)

Lead time	Target fill rate	Performance review horizon						
		∞	10	20	30	40	50	60
$L = 0$	75%	2.824	2.735 (3.17%)	2.779 (1.61%)	2.794 (1.07%)	2.802 (0.78%)	2.806 (0.63%)	2.809 (0.54%)
	80%	3.179	3.079 (3.15%)	3.127 (1.61%)	3.145 (1.07%)	3.154 (0.78%)	3.159 (0.63%)	3.162 (0.54%)
	85%	3.619	3.506 (3.12%)	3.561 (1.61%)	3.580 (1.07%)	3.591 (0.78%)	3.596 (0.63%)	3.600 (0.54%)
	90%	4.215	4.086 (3.08%)	4.149 (1.56%)	4.170 (1.07%)	4.182 (0.78%)	4.189 (0.63%)	4.193 (0.54%)
	95%	5.186	5.024 (3.13%)	5.105 (1.56%)	5.131 (1.07%)	5.146 (0.78%)	5.151 (0.68%)	5.156 (0.59%)
$L = 1$	75%	6.364	6.177 (2.93%)	6.264 (1.56%)	6.297 (1.05%)	6.314 (0.78%)	6.323 (0.63%)	6.330 (0.54%)
	80%	6.841	6.642 (2.91%)	6.734 (1.56%)	6.769 (1.05%)	6.787 (0.78%)	6.797 (0.63%)	6.804 (0.54%)
	85%	7.423	7.207 (2.91%)	7.307 (1.56%)	7.345 (1.05%)	7.365 (0.78%)	7.376 (0.63%)	7.383 (0.54%)
	90%	8.196	7.960 (2.88%)	8.068 (1.56%)	8.112 (1.03%)	8.132 (0.78%)	8.144 (0.63%)	8.152 (0.54%)
	95%	9.426	9.159 (2.83%)	9.279 (1.56%)	9.329 (1.03%)	9.352 (0.78%)	9.371 (0.59%)	9.375 (0.54%)
$L = 2$	75%	9.757	9.485 (2.78%)	9.607 (1.54%)	9.656 (1.04%)	9.681 (0.78%)	9.695 (0.63%)	9.705 (0.54%)
	80%	10.328	10.043 (2.76%)	10.169 (1.54%)	10.222 (1.03%)	10.247 (0.78%)	10.262 (0.63%)	10.273 (0.54%)
	85%	11.019	10.718 (2.73%)	10.850 (1.54%)	10.906 (1.03%)	10.933 (0.78%)	10.949 (0.63%)	10.960 (0.54%)
	90%	11.929	11.603 (2.73%)	11.749 (1.51%)	11.807 (1.03%)	11.836 (0.78%)	11.854 (0.63%)	11.865 (0.54%)
	95%	13.360	13.001 (2.69%)	13.158 (1.51%)	13.223 (1.03%)	13.256 (0.78%)	13.275 (0.63%)	13.289 (0.54%)
$L = 3$	75%	13.082	12.742 (2.60%)	12.889 (1.48%)	12.951 (1.00%)	12.983 (0.76%)	13.001 (0.62%)	13.012 (0.54%)
	80%	13.733	13.379 (2.58%)	13.530 (1.48%)	13.594 (1.01%)	13.629 (0.76%)	13.646 (0.63%)	13.659 (0.54%)
	85%	14.516	14.148 (2.54%)	14.300 (1.49%)	14.371 (1.00%)	14.406 (0.76%)	14.424 (0.63%)	14.438 (0.54%)
	90%	15.541	15.150 (2.51%)	15.310 (1.49%)	15.382 (1.03%)	15.420 (0.78%)	15.442 (0.63%)	15.458 (0.54%)
	95%	17.142	16.715 (2.49%)	16.891 (1.46%)	16.974 (0.98%)	17.008 (0.78%)	17.033 (0.63%)	17.050 (0.54%)

the initial on-hand inventory distribution follows the steady-state on-hand inventory distribution. The initial state fill rate formulation differs from this set-up in the following two respects: The initial on-hand inventory distribution is not the steady-state distribution, and the horizon length is finite rather than infinite. On the other hand, the steady-state fill rate formulation differs from the set-up that would make the traditional formulation exact in one respect rather than two: the horizon length is finite rather than infinite. Therefore, one would expect the steady-state fill rate formulation to result in a smaller discrepancy than the initial state fill rate formulation. This is precisely what we observe in all of our numerical observations, for a range of demand distributions and parameter settings. From a practical perspective, the above observation suggests that the inventory manager should be more cautious regarding the overstocking issue if the product is relatively new, or in the case of a newly signed contract in which the

inventory system starts at the order-up-to level when they face a fill rate contract.

Next we observe that the order-up-to level required to deliver a given target expected fill rate is less than the order-up-to level derived using the traditional formula. *It is worth noting that each unit decrease in order-up-to level translates into a unit decrease in average inventory level*. This difference is most acute—and the savings in average inventory cost most significant—with long lead times and short planning horizons. For example, using initial state average fill rate,

- When the lead time is three periods, the planning horizon is 10 weeks, and the target fill rate is 75% to 80%, our algorithm results in an average savings in inventory cost of around 8%.
- When the target fill rate is 90%, the average inventory cost saving is 6%.

When the lead time is short and/or the review horizon is long, the difference is less pronounced. The difference also declines as the target fill rate increases. To explain this, we note again that the initial state fill rate formulation differs from the set-up that would make the traditional formulation exact in two ways: The initial on-hand inventory distribution is not the steady-state distribution, and the horizon length is finite rather than infinite. Therefore, the longer the horizon, the closer the model gets to an infinite horizon model, and the smaller is the discrepancy between an infinite horizon model and the finite horizon model. For a fixed horizon length, the longer the lead time, the greater the number of periods required to attain the steady-state distribution of initial inventory at the start of a period. Hence the greater the lead time, the greater the discrepancy between the exact finite horizon model and the infinite horizon approximation.

It is clear, then, that our algorithm can deliver significant inventory cost savings to firms that operate a base-stock policy in a supply chain characterized by significant lead times, uncertain demand, and different demand conditions from one selling season to another, as is the case with fashion goods when production is offshore. We further illustrate the robustness of our finding when incorporating the penalty for not meeting the contractual fill rate in the next section.

6. Extensions

In this section, we extend our basic setting by considering a SLA with penalties and an alternative stochastic algorithm. The first extension illustrates that our result is robust even with conservative behavior, while the second extension shows how to improve the efficiency of the algorithm.

6.1. Service Level Agreement with Penalty

An SLA is a widely adopted form of performance-based contract in operation management practice. One feature of this contract is that the supplier may suffer a penalty for not meeting the specified target fill rate. In practice, the penalty for failing to meet a target might be a specific financial penalty and/or loss of goodwill. But since it is hard to quantify goodwill costs, we focus on a model with a financial penalty. There are two common forms of financial penalty: lump-sum penalty and proportional penalty. As pointed out by Liang and Atkins (2013), both the proportional and the lump-sum penalty SLA can induce first-best investment if the supplier adopts a static policy. Hence, without loss of generality, we assume a proportional SLA where the supplier incurs a penalty proportional to the deviation from the contracted target fill rate.

The proportional form of the penalty implies that the supplier facing the SLA may be interested in a particular probability of meeting the target service level with a specific quantity of stock, rather than targeting the expected fill rate *per se* (Thomas 2005). The unit understocking penalty and unit overstocking cost in the SLA model are equivalent to the unit underage cost and unit overage cost, respectively, in the newsvendor model; the supplier needs to balance these two costs by choosing the appropriate probability of meeting the performance threshold. The novelty here—compared with the standard newsvendor model—is that the underlying distribution from which the supplier needs to choose the critical percentile is the fill-rate distribution rather than the demand distribution (recall that fill rate is a random variable under a finite performance review horizon). Once the critical percentile is chosen, the corresponding stock level is induced by the demand distribution. The calculation of the critical stock level is difficult to do explicitly via a formula because we are in a multiple period setting with positive lead time.

Define η^* to be the probability that the service level target will be met based on demand distribution D, and let $\hat{\eta}_T(s)$ denote its approximation using empirical samples Z^N, namely $\hat{\eta}_T(s) = \frac{1}{N}\Sigma_{n=1}^{N}\mathbf{1}_{\hat{\beta}_T^{(n)}(s) \geq \beta}$. We note that the optimal level of η^* is determined by the familiar newsvendor model.

$$\eta^* = P_{Z\sim D}(\hat{\beta}_T(s) \geq \beta) \tag{20}$$

$$\hat{\eta}_T(s) = P_{Z\sim Z^N}(\hat{\beta}_T^Z(s) \geq \beta) \tag{21}$$

The corresponding inventory problem is as follows:

$$s^* = \text{argmin}_s \hat{\eta}_T(s) \geq \eta^* \tag{22}$$

We solve this problem using Algorithm K in the Online Appendix. The line search scheme is similar to the base algorithm described above. The major difference is an additional step to search for the lower and upper bound of optimal order-up-to level, because expected fill rate on infinite horizon $\beta_\infty(s)$ may not satisfy the probability constraint. The numerical results are summarized in Table 4 at the end of this document.

From Table 4, we observe that if the penalty fee (unit understocking cost) and inventory holding cost (unit overstocking cost) are comparable in magnitude, then the traditional formula will create significant overstocks in the inventory system. On the other hand, if the supplier faces a very high penalty, then the results from the traditional formula may lead to under-stocking. From a practical perspective, this result is consistent with our intuition that inventory managers tend to over stock expensive products for fear of violating the contractual service agreement. Nevertheless, we demonstrate that this conservative behavior will hurt the firm's performance under a wide range of parameters.

6.2. A Stochastic Algorithm For Fast Approximation

In the previous section, we propose a bisection method to find optimal order-up-to level. Although this method can provide a very accurate solution to the inventory problem, it may have some disadvantage in practice. Most importantly, the solution quality of this approach hinges on the scale of simulated samples, and it often needs to simulate a large set of samples to guarantee good solution quality. To obtain an accurate solution, the demand of computational resources poses a prohibitive challenge for inventory managers when a large number of inventory queries have to be processed in parallel on time. The purpose of this subsection is to describe a more efficient algorithm that exhibits a greatly improved performance to obtain the optimal order-up-to level.

Fill rate, as we proved earlier, is a monotonically non-decreasing function of order-up-to level s. As we demonstrate in this section, this allows us to improve the efficiency of the previously described bisection based root finding procedure. Under the framework of stochastic approximation (SA) by Robbins and Monro (1951), we describe the proposed stochastic algorithm for searching optimal order-up-to level in Algorithm 4. For the sake of simplicity, we leave the algorithm details in the Online Appendix but briefly describe its main idea. This new procedure applies a difference equation based on a dynamical system that relies on as few as a single sample to seek the optimal solution. Let β be the target fill rate, SA iteratively

Table 4 The Comparison of Order-up-to Level between Traditional Formula and Penalty Model for Initial State with Erlang (3,1)

Critical ratio	Fill rate	Traditional formula	Review horizon					
			$T = 10$	$T = 30$	$T = 50$	$T = 70$	$T = 100$	$T = 180$
0.40	75%	6.36	5.61 (11.85%)	6.02 (5.33%)	6.12 (3.75%)	6.17 (3.00%)	6.21 (2.37%)	6.26 (1.64%)
	80%	6.84	6.00 (12.33%)	6.46 (5.57%)	6.57 (3.92%)	6.63 (3.13%)	6.67 (2.49%)	6.72 (1.71%)
	85%	7.42	6.44 (13.31%)	6.98 (5.96%)	7.11 (4.20%)	7.17 (3.37%)	7.23 (2.66%)	7.29 (1.83%)
	90%	8.20	6.95 (15.17%)	7.65 (6.72%)	7.81 (4.71%)	7.89 (3.76%)	7.95 (2.97%)	8.03 (2.04%)
	95%	9.43	7.64 (18.91%)	8.61 (8.62%)	8.86 (5.96%)	8.98 (4.74%)	9.08 (3.72%)	9.19 (2.53%)
0.45	75%	6.36	5.76 (9.46%)	6.12 (3.84%)	6.20 (2.58%)	6.24 (2.00%)	6.27 (1.54%)	6.30 (1.01%)
	80%	6.84	6.16 (9.90%)	6.56 (4.03%)	6.66 (2.71%)	6.70 (2.11%)	6.73 (1.62%)	6.77 (1.06%)
	85%	7.42	6.62 (10.83%)	7.10 (4.38%)	7.20 (2.94%)	7.25 (2.28%)	7.29 (1.75%)	7.34 (1.14%)
	90%	8.20	7.16 (12.68%)	7.78 (5.04%)	7.92 (3.37%)	7.98 (2.61%)	8.03 (1.99%)	8.09 (1.29%)
	95%	9.43	7.87 (16.47%)	8.78 (6.81%)	9.01 (4.46%)	9.10 (3.43%)	9.18 (2.60%)	9.27 (1.68%)
0.50	75%	6.36	5.92 (7.02%)	6.21 (2.35%)	6.27 (1.42%)	6.30 (1.02%)	6.32 (0.71%)	6.34 (0.39%)
	80%	6.84	6.33 (7.43%)	6.67 (2.50%)	6.74 (1.50%)	6.77 (1.08%)	6.79 (0.76%)	6.81 (0.42%)
	85%	7.42	6.80 (8.33%)	7.22 (2.78%)	7.30 (1.67%)	7.33 (1.20%)	7.36 (0.85%)	7.39 (0.46%)
	90%	8.20	7.36 (10.16%)	7.92 (3.34%)	8.03 (2.01%)	8.08 (1.44%)	8.11 (1.01%)	8.15 (0.56%)
	95%	9.43	8.11 (13.96%)	8.96 (4.93%)	9.15 (2.94%)	9.23 (2.11%)	9.29 (1.48%)	9.35 (0.83%)
0.55	75%	6.36	6.07 (4.54%)	6.31 (0.84%)	6.35 (0.24%)	6.36 (0.02%)	6.37 (−0.12%)	6.38 (−0.23%)
	80%	6.84	6.51 (4.87%)	6.78 (0.95%)	6.82 (0.29%)	6.84 (0.04%)	6.85 (−0.11%)	6.86 (−0.23%)
	85%	7.42	7.00 (5.71%)	7.34 (1.16%)	7.39 (0.39%)	7.41 (0.11%)	7.43 (−0.07%)	7.44 (−0.22%)
	90%	8.20	7.58 (7.53%)	8.07 (1.59%)	8.14 (0.64%)	8.17 (0.26%)	8.19 (0.02%)	8.21(−0.18%)
	95%	9.43	8.35 (11.39%)	9.14 (3.03%)	9.30 (1.37%)	9.35 (0.76%)	9.39 (0.34%)	9.43 (−0.03%)
0.60	75%	6.36	6.24 (1.97%)	6.41 (−0.71%)	6.43 (−0.98%)	6.43 (−1.01%)	6.43 (−0.98%)	6.42 (−0.88%)
	80%	6.84	6.69 (2.23%)	6.89 (−0.66%)	6.91 (−0.98%)	6.91 (−1.03%)	6.91 (−1.00%)	6.90 (−0.90%)
	85%	7.42	7.20 (3.00%)	7.46 (−0.54%)	7.49 (−0.94%)	7.50 (−1.02%)	7.50 (−1.01%)	7.49 (−0.92%)
	90%	8.20	7.80 (4.80%)	8.21 (−0.21%)	8.26 (−0.81%)	8.27 (−0.96%)	8.28 (−1.00%)	8.27 (−0.95%)
	95%	9.43	8.61 (8.69%)	9.33 (1.01%)	9.45 (−0.27%)	9.49 (−0.66%)	9.51 (−0.84%)	9.51 (−0.93%)

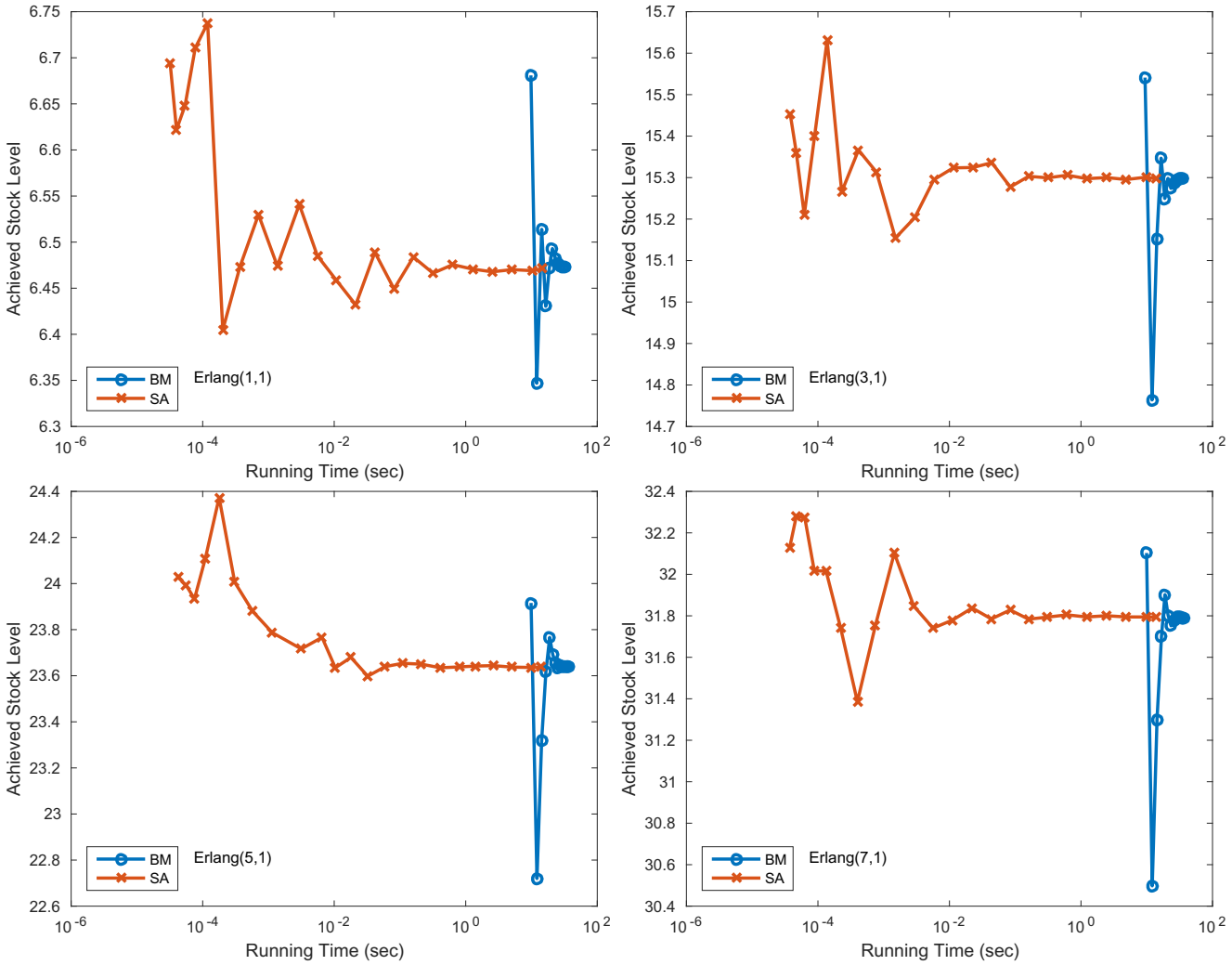
generates a solution sequence $\{s_1, s_2, \ldots, s_n, \ldots\}$ as follows:

$$s_n = s_{n-1} + \eta_n\left(\beta - \hat{\beta}_T^{(n)}\right),$$

where $\hat{\beta}_T^{(n)}$ is the empirical fill rate on the n-th sample and the scalar η_n is used to control the moving step size. We have a few comments regarding the proposed algorithm. Firstly, stochastic approximation has a significantly faster iteration when only single sample is simulated for computing $\hat{\beta}_T^{(n)}$. On the other hand, the bisection method requires all the samples to update lower or upper bound, and hence will be much slower than SA. Secondly, SA iterates in a non-monotone way, and the update direction is not always towards the optimal order-up-to level because of large variance in $\hat{\beta}_T^{(n)}$; SA requires many more iterations to ensure convergence in the long run. Based on the theory in Robbins and Monro (1951), we can employ a decaying step size to achieve asymptotic convergence to the optimal order-up-to level.

We now demonstrate the efficiency of stochastic approximation through numerical experiments. We first compare the convergence of the stochastic approximation method with the bisection method. In the following experiments, we set the parameters $T = 50$ (horizon length) and $L = 3$ (lead time). We first drew 10^7 samples, each of which was a vector in $\mathbf{R}^T$ with coordinates generated independently from a Erlang distribution, Erlang (γ, λ) with fixed rate $\lambda = 1$, and varying shapes $\gamma = 1, 3, 5, 7$. Simulations were then run using these samples. The step size η_n was chosen to be the function $\eta_n = \frac{a}{a+n^\kappa}$, where $a = 100$ and $\kappa = 1$. Figure 3 illustrates the obtained order-up-to level with

Figure 3 Illustration of Convergence Comparison between Bisection and Stochastic Approximation Algorithm [Color figure can be viewed at wileyonlinelibrary.com]



Notes. The red/blue curves represent the convergence of the stock levels over the algorithm running time, for the stochastic approximation/bisection method, respectively. Demand sequence is generated from Erlang(γ, 1) with shape $\gamma = 1, 3, 5, 7$. The target fill rate is $\beta = 0.9$ with $L = 3$, and review horizon $T = 50$.

Table 5 Efficiency Comparison between Bisection Method (BM) and Stochastic Approximation Algorithms (SA)

			$T = 20$				$T = 40$				$T = 60$			
			BM		SA		BM		SA		BM		SA	
Lead time	Erlang shape	Traditional formula	s	Time	s	Time	s	Time	s	Time	s	Time	s	Time
$L = 0$	$\gamma = 1$	2.303	2.191	4.423	2.197	2.022	2.249	7.479	2.245	3.877	2.263	11.522	2.265	5.725
	$\gamma = 3$	4.215	4.146	4.795	4.145	1.974	4.179	8.616	4.180	3.753	4.192	11.919	4.192	5.743
	$\gamma = 5$	6.064	6.003	5.086	6.003	2.198	6.031	8.424	6.034	3.756	6.041	13.013	6.045	5.513
	$\gamma = 7$	7.887	7.834	5.562	7.832	2.175	7.859	9.925	7.860	3.733	7.865	13.614	7.867	5.506
$L = 1$	$\gamma = 1$	3.890	3.662	6.262	3.663	2.027	3.771	11.045	3.771	3.876	3.808	17.771	3.812	5.725
	$\gamma = 3$	8.196	7.975	7.471	7.977	1.967	8.084	14.329	8.084	3.756	8.122	19.635	8.120	5.546
	$\gamma = 5$	12.236	12.002	7.202	12.002	2.201	12.116	14.070	12.116	3.772	12.155	20.026	12.155	5.670
	$\gamma = 7$	16.180	15.924	8.051	15.929	1.964	16.056	14.885	16.055	3.735	16.094	22.228	16.095	5.503
$L = 2$	$\gamma = 1$	5.322	4.960	7.599	4.962	2.027	5.135	15.614	5.139	4.055	5.193	22.187	5.197	5.727
	$\gamma = 3$	11.929	11.524	8.719	11.526	1.967	11.729	16.462	11.729	3.757	11.794	24.369	11.796	5.530
	$\gamma = 5$	18.146	17.689	9.786	17.690	1.961	17.923	18.426	17.923	3.735	18.001	27.440	17.998	5.519
	$\gamma = 7$	24.216	23.710	9.383	23.710	1.960	23.965	18.378	23.975	3.727	24.051	27.356	24.053	5.598
$L = 3$	$\gamma = 1$	6.681	6.172	10.356	6.168	2.029	6.423	19.744	6.415	3.885	6.506	29.362	6.500	5.748
	$\gamma = 3$	15.541	14.919	11.492	14.924	1.973	15.234	22.608	15.234	3.948	15.338	31.857	15.339	5.693
	$\gamma = 5$	23.916	23.201	11.170	23.206	1.961	23.565	22.859	23.574	3.741	23.687	32.168	23.686	5.709
	$\gamma = 7$	32.103	31.297	12.286	31.302	2.138	31.717	24.379	31.712	3.746	31.843	34.458	31.850	5.734

Notes: "s" denotes the order-up-to level, and "Time" denotes the computation time in seconds. For each method 10^7 simulated data are generated from *Erlang*(γ, 1) with shape $\gamma = 1, 3, 5, 7$. The target fill rate is $\beta = 0.90$.

respect to the logarithm of the running time in seconds, for both approaches. We make a few observations. First of all, SA is non-monotone and fluctuating due to the stochastic noise, but it immediately (in less than 0.5 s) moves to a steady phase where the sequence is converging to the optimal order-up-to level. Secondly, we observe that the stochastic algorithm converges to the optimum fairly quickly; it is able to reach close enough to the optimum without even using all the simulated data, at a point when the bisection method has not yet finished one iteration.

Next, we investigate the time cost of stochastic approximation as opposed to the bisection method on simulated data with combination of following parameters: $\gamma = 1, 3, 5, 7$, $T = 20, 40, 60$, and $L = 0, 1, 2, 3$. The experimental results are presented in Table 5. We let the bisection method run until the change was less than 0.005 and the stochastic algorithm terminates after one-pass of the simulated data. We observe that the mean value of the order-up-to levels obtained from the stochastic approximation is very close to the bisection method, with a difference of less than ± 0.005. On the other hand, while preserving solution quality, the stochastic algorithm obtains much faster empirical convergence with up to 7× speed-up compared to the bisection method.

7. Conclusions

Our study was motivated by observing the discrepancy between the traditional fill rate formula, which applies only in an infinite horizon model, and the finite-horizon service-level agreements that are implemented in practice. We find that under certain circumstances (e.g., high lead time relative to the length of the planning horizon, variations in demand conditions or product features from one SLA to the next) this discrepancy can have a significant impact on achieved fill rate over a finite performance review horizon. It is very important to note that imposing a finite-horizon, service level contract will inflate the achieved expected fill rate to a level well above the contractually specified target, which results in substantially higher inventory related costs. For instance, current commercial software suggests that the inventory manager needs to set the stocking levels at 11.52% higher than the optimal level when the performances review horizon $T = 10$ and lead time $L = 1$ with a initial state initial inventory. The potential savings from lowering inventory levels, when aggregated over a year, are substantial and can have a direct impact on a firm's balance sheet.

We briefly point out a few limitations and sketch some ideas for future research. To begin with, our model assumes that a manager faces an objective function for which the optimal solution is to select the minimum stock level that ensures that a contracted fill rate, to be measured over a finite horizon with i.i.d. demands from period to period, is met on average over the horizon. The practical significance of our modeling approach and solution is critically tied to the assumption that the operations manager is willing and able to use simulation-based optimization; this assumption underpins the algorithmic section of our paper. An extension of our research might consider a different objective function that includes expected

holding costs, expected backorder costs, and the expected cost of not meeting the fill-rate stipulated by the SLA over a finite horizon. We analyze the above stated components of the objective function separately, but not all together in a single objective function. Second, SLAs are widely applied in many different contexts. We focus on their application in inventory management systems. Future research could expand the current idea to investigate other settings. For example, the staffing decisions in a Call Center where SLA is also commonly implemented (Xia et al. 2015). Third, we restrict the inventory policy to a base stock policy. It is worthwhile to explore whether a similar overstocking problem will occur with other inventory policies (e.g., (R, Q) policy). Notwithstanding these limitations, this study closes a significant gap in the literature by investigating the role of positive lead time and provides a solution to the problem. We believe that the current research is also relevant to the practice in the sense that our results can be integrated into commercial software and generates tremendous savings for managers facing the SLAs.

Acknowledgments

Lai Wei and Qi Deng are co-corresponding authors. We thank Professor Chelliah Sriskandarajah, the Senior editor, and two anonymous reviewers for their valuable and constructive suggestions, which helped to improve the paper significantly. Lai Wei thank the generous sponsor provided by the Shanghai Pujiang Program (17PJC065). The authors also appreciate Douglas Thomas, Chun-Miin (Jimmy) Chen, and conference participants of POMS-HK, POMS and INFORMS Annual conferences for their helpful comments.

Notes

[1]Most recent statistics are available at https://www.census.gov/mtis/index.html.

[2]For example, the SAS Inventory Replenishment Planning 9.1 Users Guide provides the detailed formula used in the software, which can be retrieved at http://support.sas.com/documentation/onlinedoc/91pdf/sasdoc_91/inventory_ ug_7307.pdf.

[3]We thank the one anonymous Reviewer and Senior Editor for suggesting that we consider the initial on-hand inventory starting from the steady state.

[4]Note that there is an inconsistency between Equation (4) above and Equation (6) in Sobel (2004). Essentially, there was a typographic error in the subscript of the summation in Sobel (2004), where one should have summed from $L\gamma + 1$ instead of $L\gamma + 2$.

[5]We thank the one anonymous Reviewer and Senior Editor for suggesting this. Due to the page limit, we have attached the results when the initial on-hand inventory equals to the order-up-to level for Normal and Poisson distributed demands in the appendix. Additional results are available based upon requests to the authors.

References

Axsater, S. 2006. *Inventory Control*, 2nd edn. Springer's International Series in Operations Research & Management Science, Berlin.

Banerjee, A., A. Paul. 2005. Average fill rate and horizon length. *Oper. Res. Lett.* **33**(5): 525–530.

Cachon, G., C. Terwiesch. 2008. *Matching Supply with Demand: An Introduction to Operations Management*, 2nd edn. McGraw-Hill/Irwin, New York.

Chen, C., D. Thomas. 2015. Inventory allocation in the presence of service level agreements. Working paper.

Chen, J., D. K. Lin, D. Thomas. 2003. On the single item fill rate for a finite horizon. *Oper. Res. Lett.* **31**(2): 119–123.

Fu, M. C., F. W. Glover, J. April. 2005. Simulation optimization: A review, new developments, and applications. *Proceedings of Winter Simulation Conference*. IEEE, Orlando, 83–95.

Glasserman, P., S. Tayur. 1995. Sensitivity analysis for base-stock levels in multiechelon production-inventory systems. *Management Sci.* **41**(2): 263–281.

Guijarro, E., M. Cardos, E. Babiloni. 2012. On the exact calculation of the fill rate in a periodic review inventory policy under discrete demand patterns. *Eur. J. Oper. Res.* **218**(2): 442–447.

Johnson, E. M., H. L. Lee, T. Davis, R. Hall. 1995. Expressions for item fill rates in periodic inventory systems. *Nav. Res. Log.* **42**(1): 57–80.

Kapuscinski, R., S. Tayur. 1998. A capacitated production-inventory model with periodic demand. *Oper. Res.* **46**(6): 899–911.

Katok, E., D. Thomas, A. Davis. 2008. Inventory service-level agreements as coordination mechanisms: The effect of review periods. *Manuf. Serv. Oper. Manag.* **10**(4): 609–624.

Liang, L., D. Atkins. 2013. Designing service level agreements for inventory management. *Prod. Oper. Manag.* **22**(5): 1103–1117.

Nahmias, S., T. L. Olsen, 2015. *Production and Operations Analysis*, 7th edn. Waveland Press, Long Grove.

Paul, A., Y. Tan, A. J. Vakharia. 2015. Inventory planning for a modular product family. *Prod. Oper. Manag.* **24**(7): 1033–1053.

Robbins, H., S. Monro. 1951. A stochastic approximation method. *Ann. Math. Stat.* **22**(3): 400–407.

Schneider, H. 1981. Effect of service-levels on order-points or order-levels in inventory models. *Int. J. Prod. Res.* **19**(6): 615–631.

Silver, E., D. Pyke, R. Peterson. 1998. *Inventory Management and Production Planning and Scheduling*, 3rd edn. Wiley, New York.

Sobel, M. 2004. Fill rates of single-stage and multistage supply systems. *Manuf. Serv. Oper. Manag.* **6**(1): 41–52.

Song, J. 1998. On the order fill rate in a multi-item, base-stock inventory system. *Oper. Res.* **45**(6): 831–845.

Spieckermann, S., K. Gutenschwager, H. Heinzel, S. Vob, D. Maintal. 2000. Simulation-based optimization in the automotive industry: A case study on body shop design. *Simulation* **75**(5): 276–286.

Teunter, R. H. 2009. Note on the fill rate of single-stage general periodic review inventory systems. *Oper. Res. Lett.* **37**(1): 67–68.

Thomas, D. 2005. Measuring item fill-rate performance in a finite horizon. *Manuf. Serv. Oper. Manag.* **7**(1): 74–80.

Xia, Y., B. Chen, V. Jayaraman, C. L. Munson. 2015. Competition and market segmentation of the call center service supply chain. *Eur. J. Oper. Res.* **247**(21): 504–514.

Zhang, J. 2012. Analysis of fill rate in general periodic review two-stage inventory systems. *Int. J. Oper. Res.* **14**(4): 505–512.

Zhang, J., L. Bai, Y. He. 2010. Fill rate of general periodic review two-stage inventory systems. *Int. J. Oper. Res.* **8**(1): 62–84.

Zhang, J., J. Zhang 2007. Fill rate of single-stage general periodic review inventory systems. *Oper. Res. Lett.* **35**(4): 503–509.

Zhong, Y., Z. Zheng, M. C. Chou, C. Teo. 2017. Resource pooling and allocation policies to deliver differentiated service. *Management Sci.*, https://doi.org/10.1287/mnsc.2016.2674

Zipkin, P. 2000. *Foundations of Inventory Management*, 1st edn. McGraw-Hill/Irwin, Boston.

Supporting Information

Additional supporting information may be found online in the supporting information tab for this article:

Appendix S1: Proof of Lemma 1.
Appendix S2: Proof of Lemma 2.
Appendix S3: Proof of Lemma 3.
Appendix S4: Proof of Lemma 4.
Appendix S5: Proof of Lemma 5.
Appendix S6: Proof of Theorem 2.
Appendix S7: Proof of Remark 1.
Appendix S8: Proof of Theorem 3.
Appendix S9: Fill Rate Distribution.
Appendix S10: Bisection Method.
Appendix S11: Service Level Agreement with Penalty.
Appendix S12: Stochastic Algorithm.
Appendix S13: Statistical Consistency.
Appendix S14: Numerical Results.