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 performance review horizon. However, in practice, inventory managers are liable to maintain and report fill rates over a 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 performance review horizon exceeds the expected fill rate assuming a 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.

K *:* service level agreement; fill rate; positive lead time; base-stock policy; simulation-based optimization
H : 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.

F 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 - , whereas in practice the SLA requires the supplier to meet the target fill rate over a specific, , 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.

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 non-factorizable. 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 foundatiotos,videsrurpropose

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 model 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

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 periods for single-stage and two-stage inventory systems. Note that if = 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 (,) 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

quantity on-order + net inventory) back up to the order-up-to level . Note that the replenishment order quantity always equals the immediately preceding demand. As a result, we can rewrite as $\Sigma_{=-L}^{-1} D$

targets. By comparing the four sub-figures in Figure 1, we first find that the expected achieved fill rate is

LEMMA 2. W $E[D] = 1$ $= 1, 2, \ldots,$
$_L > 0$

$$E \quad \frac{1}{}$$

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

L • 4B	•	•						
	•	∞	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)

L • 4B	•	•						
	•	∞	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

generates a solution sequence $\{ _1, _2, \ldots, , \ldots\}$ as follows:

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: γ = 1, 3, 5, 7, = 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 $\pm$0.005. On the other hand, while preserving solution quality, the stochastic algorithm obtains much faster empirical convergence with up to 7$\times$ 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,

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., (,) 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

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

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

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

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

Zipkin, P. 2000. *F I M* , 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 Algorithmochastic Algoritrd online