European Journal of Operational Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Production, Manufacturing, Transportation and Logistics

On the stationary stochastic response of an order-constrained inventory system

Xun Wang[a], Stephen M. Disney[b], Borja Ponte[c,*]

[a] Logistics Systems Dynamics Group, Cardiff Business School, Cardiff University, UK
[b] Centre for Simulation, Analytics & Modelling, University of Exeter Business School, University of Exeter, UK
[c] Department of Business Administration, Polytechnic School of Engineering, University of Oviedo, Spain

A R T I C L E I N F O

Article history:
Received 29 June 2021
Accepted 14 April 2022
Available online xxx

Keywords:
Inventory
Constraint
Stochastic response
Base stock policy

A B S T R A C T

We investigate the stochastic response of a base stock inventory system where the order quantity is either upper- or lower-constrained. This system can represent many real-world settings: forbidden returns, minimum order quantities, and capacity constraints for example. We show that this problem can be translated into a stopping time problem where the distributions of orders and inventory can be represented by a countably infinite mixture of truncated and convoluted demand distributions. This result can be extended to the cases of arbitrary lead time and auto-correlated demand. A state space algorithm is developed to approximate the first- and second-order moments of the order quantity and inventory level. Via a numerical analysis, we investigate the performance of the approximation, as well as the operational and economic impact of the order constraint. In particular, the constraint impacts order and inventory variances via different combinations of the mixture and truncation effects. We show how tuning the constraint can improve the operational and financial performance of the inventory system by acting as a smoothing mechanism.

1. Introduction

We study the limiting, or asymptotic, distributions of order quantity and inventory level in a nonlinear inventory system, where the order quantity is limited by an upper- or lower-constraint, but the demand is not constrained by this limit. In other words, we explore the stochastic response of an inventory system where a (maximum or minimum) constant constraint on the order quantity is in the interior of the support of the demand quantity. This model has many real-world applications, which we now elaborate to motivate this research.

Forbidden returns (FR). There are supply chain scenarios where the customer is allowed to return the products but the seller is not allowed to do so. In this sense, the demand can be either positive or negative, where negative demand implies products returned from customers exceed those demanded by them in a period, but orders to suppliers are constrained to nonnegative values. The existence of negative demand has been reported in the literature. For example, in the book publishing and electronic industries, unwanted items can be returned to the manufacturer. In general, the negative demand assumption holds if we accept the net flow of goods can be reversed temporarily. In the upstream, the inversion of the material flow is less common in business-to-business environments. Typically, businesses are not allowed to return their surplus products to suppliers; this can be represented mathematically by a nonnegative assumption on the order quantity. This assumption is common in the classic inventory control literature; however, as highlighted by Chatfield & Pritchard (2013), this assumption is rare in the bullwhip effect and supply chain dynamics literature.

Minimum order quantity (MOQ). The lower limit of the order quantity may also be a positive number, representing a MOQ. The business can either order above (or at) the MOQ, or not order at all. This case is frequently observed in the upstream supply chain. For example, minimum order quantities are commonly imposed by off-shore suppliers, due to low profit margins, high set-up and shipment costs, and batching requirements in the manufacturing process. In the online grocery industry, it is common practice to set an MOQ as a prerequisite to receive discounted, or free, shipping. Zhou, Zhao, & Katehakis (2007) give an interesting discussion on the role of minimum order quantities in supply chains. Moreover, the famous (s, S) ordering policy, optimal in many cost scenarios

* Corresponding author.
E-mail addresses: wangx46@cardiff.ac.uk (X. Wang), s.m.disney@exeter.ac.uk (S.M. Disney), ponteborja@uniovi.es (B. Ponte).

https://doi.org/10.1016/j.ejor.2022.04.020

(Scarf, 1960; Zheng, 1991), is an ordering rule such that if the inventory position is below a reorder point s, an order is placed to bring the inventory position back to the order-up-to level S. The (s, S) policy can be seen as an MOQ constraint, where $S - s$ is the MOQ, whereas the demand can be lower than $S - s$.

Capacity constrained (CC). In production systems, maximum production quantities are often set due to limited internal capacity; however, customer demand is free to go beyond the capacity limit. The maximum order quantity is also used by some material requirements planning (MRP) software to prevent the automatic generation of unreasonably large orders. Besides, upper bound constraints may also emerge from external sources. For example, vendors may impose maximum order quantity constraints to ration customers when the capacity or raw material availability is limited to prevent some customers over ordering and other customers under-fulfilled. Also, maximum order quantities may be strategically imposed by single-source vendors that sell to multiple buyers to avoid opportunistic behaviours such as the hold-up problem (Dahel, 2003).

Fig. 1 shows a time series of real demand and its distribution. The product is from the consumer electronics industry, whose distribution is approximately normal, with two negative demands in the three-year period. Fig. 1 also shows the time series and distribution plot of simulated orders generated by the three constrained order policies that we study in this paper: FR, MOQ (with $s = 50$ and $S = 150$), and CC (capacity constraint set to 300). Unit lead-times were assumed. In all these cases, when the desired order quantity exceeds the lower, or upper, limit allowed, the actual order quantity has to be switched either to this limit or another designated value, greatly increasing the analytical complexity of the inventory model. Indeed, traditional analytical techniques for linear time invariant (LTI) models can no longer be applied in these nonlinear scenarios. This perspective explains why these complex nonlinear behaviours are understudied in the literature despite their obvious practical relevance and acts as the major driver of this research.

In this paper, we make the following contributions:

- We show the equivalence of three order-constrained inventory systems; the FR, MOQ, and CC systems can be represented in a general form for analysis.
- We derive the limiting density functions of the order quantity and inventory level in the constrained inventory system, under transportation delay and demand autocorrelation. The derivation is based on distribution truncation, convolution, and mixture; no assumptions regarding the demand distribution need to be made.
- For Gaussian demand, we propose a state space algorithm to approximate the first- and second-order moments of order quantity and inventory level, based on the properties of truncated normal random variables. This algorithm can be used in scenarios with a transportation delay and demand correlation. We discuss the applicability, efficiency, and effectiveness of this algorithm.
- We investigate how the order constraint influences the variance of the order quantity and the inventory level. We reveal how the order constraint creates a mixture effect and a truncation effect that affects the trade-off between the order and inventory variances. We also show how the constraint may reduce the sum of the order and inventory variances.

The rest of this paper is organized as follows. In Section 2, we review related literature. Section 3 gives the notation and the main results characterizing the distributions of orders and inventory levels in the order-constrained inventory model with unit delay and independently and identically distributed (i.i.d.) demand. Extensions to arbitrary transportation delay and correlated demand are made in Section 4. In Section 5, we propose an approximation algorithm for calculating the first- and second-order moments of the stochastic response generated by a Gaussian demand. Section 6 is devoted to operational aspects of our model, where a numerical analysis examines the variance amplification and the service level performance in this model. Finally, Section 7 concludes and highlights avenues for future research.

2. Literature review

This research contributes to the literature from three perspectives: (i) to the *bullwhip effect* research, which usually explores the inventory control problem through a linear quadratic Gaussian (LQG) model; (ii) to nonlinear *supply chain dynamics* research, which includes constraints on system variables but mainly investigates deterministic demand series; and (iii) to the *classical inventory control* literature, especially those on the (s, S) policy and the capacitated system, which predominantly models demand as a nonnegative, i.i.d. and integer time series.

LQG models have been extensively applied in bullwhip research because of their analytical tractability. In the seminal work of Lee, Padmanabhan, & Whang (1997), the bullwhip effect triggered by demand signal processing is simulated by a Gaussian input with first-order autocorrelation. The probability of the demand and order quantity being negative is assumed negligible by setting the coefficient of variation of demand sufficiently small. In addition, to preserve linearity of the model, unmet demand is backlogged. These assumptions were frequently adopted in later research on bullwhip effect and inventory amplification (Chen & Lee, 2009; Disney, Gaalman, Hedenstierna, & Hosoda, 2015). The connection between the second-order moment and the cost incurred is established by the assumption that the inventory (order) cost function is convex, under which the cost increases with the inventory (order) variance. In the special case of normally distributed demand, piecewise linear and convex cost function and optimal safety stock (production capacity), the cost is a multiple of the standard deviation of the inventory (orders) (Boute, Disney, Gijsbrechts, & Van Mieghem, 2021).

On the other hand, it has long been known that real-life supply chains are constrained. For instance, in the Beer Game experiment, participants are allowed to backlog unsatisfied demand to the next period, but excess inventory cannot be returned to their supplier (Sterman, 1989). This can be seen more clearly in their model of participants' behaviour; the order quantity is limited by a $\max\{0, \cdot\}$ constraint, i.e., the order quantity is not allowed to be negative. There are various attempts to study the dynamic behaviour of such system using nonlinear dynamical systems theory (Larsen, Morecroft, & Thomsen, 1999; Laugesen & Mosekilde, 2006). The research interest focuses on system stability; the primary finding is that complex behaviour of the inventory system can be induced by nonlinear constraints and irrational decisions. The nonlinear effects may even dominate the dynamics of the system, resulting in very complex behaviours, such as chaos and super-chaos (phenomena where the system behaviour is sensitive to perturbation in the system state and parameter values, which is difficult or impossible to predict, Wang, Disney, & Wang, 2012). However, this research stream often focuses on system stability and a deterministic demand signal (a step function is used in most cases). Attempts to extend the nonlinear inventory system model to stochastic demand have been predominantly simulation based, e.g., Chatfield & Pritchard (2013) for the lower constraint and Ponte, Wang, de la Fuente, & Disney (2017) for the upper constraint to the order quantity. More recently, Disney, Ponte, & Wang (2021) study the dynamics of a lost-sales inventory system where the inventory level is nonnegative. They show that both order and inventory distribu-

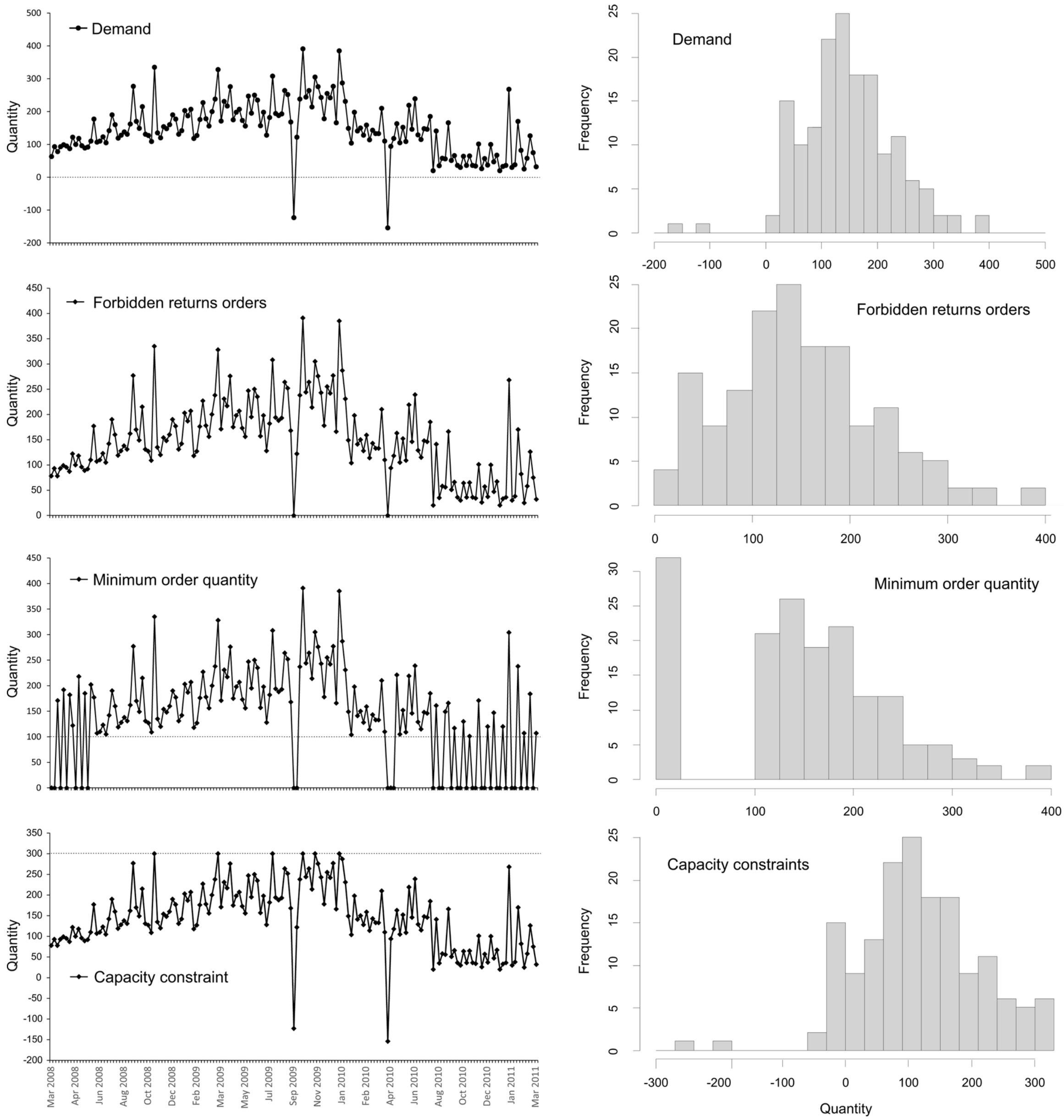

Fig. 1. A real demand pattern and the simulated response of the base stock policy with three different order constraints (forbidden returns, minimum order quantity, and a capacity constraint).

tions can be derived from the censored demand distribution. We show this is not the case in an order-constrained system.

Lastly, in the inventory theory literature, a lot of attention was given to the inventory and the order distribution in either the (s, S) or the capacitated systems. Using renewal theory, Karlin (1958) has derived an expression for the inventory distribution (also see Scarf, 1963); for the same model Schultz (1983) provides the order distribution and the order variance. Approximation algorithms for computing the order and inventory variance under the (s, S) policy have been proposed by Schneider, Rinks, & Kelle (1995) and later utilized by Kelle & Milne (1999) to study the bullwhip effect. More recently, Noblesse, Boute, Lambrecht, & Houdt (2014) study the bullwhip effect in a continuous-review (s, S) system with compound Poisson demand. They derive the distribution of orders as well as the time between orders.

Another, seemingly detached, research stream considers the capacitated inventory system, mostly focusing on the optimal ordering policy when the order quantity is upper-constrained. Federgruen & Zipkin (1986a,b) show that a base stock policy is optimal under the assumption of stationary demand. Tayur (1993) notes the similarities between the capacitated inventory system and a D/G/1 queuing system, and used the analogy of an infinite-stage uncapacitated supply chain system to derive the stationary distribution of the inventory level. Using the *shortfall* concept they show that the inventory distribution can be represented by an infinite convolution between the demand distribution and the shortfall distribution. The stability of the capacitated system is discussed in Glasserman & Tayur (1994), where stationarity conditions are derived. Approximation methods of the optimal base stock level are proposed in Glasserman & Tayur (1995) and Glasserman (1997). Kapuściński & Tayur (1998) extend Federgruen & Zipkin (1986a,b)'s capacitated model to the case of cyclic demand, and Parker & Kapuściński (2004) extend the model to a two-echelon system. Both studies found a (modified)

Table 1
Commonly used notation.

Sets	
$\mathcal{A}$	The set of admissible order quantities
$\mathcal{A}'$	The complement of $\mathcal{A}$
$\mathcal{N}$	The set of positive integers
$\mathcal{R}$	The set of real numbers
Variables (time-dependent random processes)	
ε	Gaussian i.i.d. variable with zero mean and unit variance
d	Demand
o	Order quantity
$\tilde{o}$	Desired order quantity
i	Net inventory level after fulfillment and consumption
w	Work-in-process; $w_t = 0$ when $L = 1$
IP	Inventory position, $IP_t = i_t + w_t$; $IP_t = i_t$ when $L = 1$
τ	Order cycle length
Inventory system parameters	
α_T	Target service level
L	Transportation lead time
S	Order-up-to level
Functions and distributions	
$x_{\mathcal{A}}$	Truncation of x with $\mathcal{A}$: $x_{\mathcal{A}} = x$ if $x \in \mathcal{A}$; $x_{\mathcal{A}} = 0$ otherwise
$[f(x)]_{\mathcal{A}}$	Truncation of f with $\mathcal{A}$: $[f(x)]_{\mathcal{A}} = f(x)$ if $x \in \mathcal{A}$; $[f(x)]_{\mathcal{A}} = 0$ otherwise
$f(x) * g(x)$	Convolution between $f(x)$ and $g(x)$
$\phi(x)$, $\Phi(x)$	pdf and cdf of standard normal distribution
$\mathbb{E}(x)$, μ_x	The expectation vector of x; $\dim \mathbb{E}(x) = \dim x$
$\mathbb{V}(x)$, σ_x^2	The covariance matrix of x; $\dim \mathbb{V}(x) = \dim x \times \dim x$

base stock policy to be optimal. Levi, Roundy, Shmoys, & Truong (2008) develop an approximation algorithm for the optimal policy. Gavirneni, Kapuściński, & Tayur (1999) study the case of information sharing in a two-echelon capacitated supply chain and proved the optimality of the order-up-to policy under this scenario.

In this work, unlike previous research, we do not focus on the optimal policy under the constraint, but on the stochastic response and variability of a constrained base stock inventory system. From this perspective, this study complements the above research in several ways. First, previous research did not reveal the connection between the systems with different order constraints, whereas we show they are equivalent. This can also be seen from the similar structure of order and inventory distributions in these systems. Second, previous research, based on renewal theory or the D/G/1 queuing analogy, requires that the random variables in demand are nonnegative. We relax this assumption by allowing negative demand using stopping time and mixture distributions. Third, most previous research assumes unit lead-time and i.i.d. demand, whereas our approach allows us to incorporate arbitrary lead-times and auto-correlated demand in a straightforward manner. Fourth, we propose a novel approximation to calculate the order and inventory variances which outperforms previous approximation methods.

3. Distributions of order and inventory

In this section, we present an analysis under the assumptions of unit transportation delay and temporally independent demand series. These assumptions will be relaxed later in Sections 4.1 and 4.2, respectively. We use the notation outlined in Table 1 throughout this paper. Other notation will be introduced as needed.

3.1. The stopping time problem

We first present a general form of our problem that can represent a variety of inventory systems with order-constraints, including the FR, MOQ, and CC sytems. We then show this general form is actually a stopping time problem. We consider a periodic review system, with $t \in \mathcal{N}$. In this section we assume the demand is an i.i.d. random process following a distribution whose probability density function (pdf) $f_d(x)$ is defined on the real line $\mathcal{R}$. Here we do not need to specify the type of distribution, but merely assume that f_d is integrable. The transportation delay of the inventory system is one period; that is, the order placed at the end of period t will be received and available to satisfy demand during period $t+1$. At the end of the period, after demand has been satisfied, the observed inventory levels and open orders are subtracted from the order-up-to level S to determine the replenishment orders. A description of the sequence of events for unit lead times is given in Fig. 2.

The inventory balance equation is

$$i_t = i_{t-1} + o_{t-1} - d_t. \tag{1}$$

The variable i is the inventory level and o the order quantity. A base stock policy is used to make replenishment decisions. That is, the *desired order quantity* is the difference between the order-up-to level S and the inventory position IP_t (when the lead-time is one period, $IP_t = i_t$). Under FR, the actual order quantity equals desired order quantity if it is positive, and zero otherwise:

$$o_t = \begin{cases} S - i_t & \text{if } S - i_t > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Under the MOQ policy, the actual order quantity equals the desired order quantity if inventory i_t is below the re-order point s, and zero otherwise:

$$o_t = \begin{cases} S - i_t & \text{if } i_t < s \\ 0 & \text{otherwise.} \end{cases}$$

In the CC setting, the actual production quantity equals the desired order quantity if it is less than the production capacity C_p, otherwise the actual production quantity equals C_p:

$$o_t = \begin{cases} S - i_t & \text{if } S - i_t < C_p \\ C_p & \text{otherwise.} \end{cases}$$

We will now show the above three systems can be represented by a single general form. Let the admissible region $\mathcal{A}$ be the half real line divided by the real number C_1, i.e., $(-\infty, C_1)$ or $(C_1, +\infty)$. Note that $C_1 \notin \mathcal{A}$. Also let C_2 be a real number not in $\mathcal{A}$, $C_2 \in \mathcal{A}' = \mathcal{R} \setminus \mathcal{A}$, $\mathcal{A}'$ is the complement set of $\mathcal{A}$ with respect to $\mathcal{R}$. Thus $o \in \mathcal{A} \cup \{C_2\}$, and it is possible that $C_1 = C_2$. We can write the order-

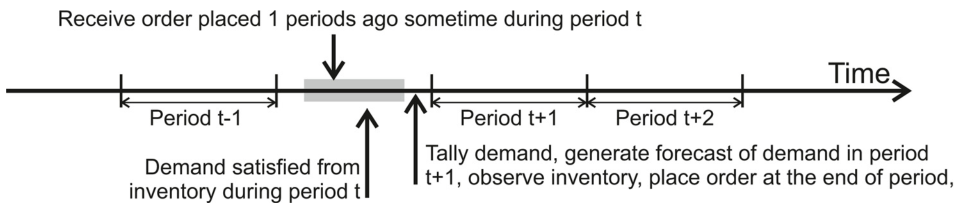

Fig. 2. Sequence of events in the base stock replenishment policy with unit lead-time.

Table 2
Illustration of the order process in a forbidden returns system.

t	Demand	Desired order quantity	Actual order quantity	Inventory level	Number of periods from last admissible order
1	6.11	6.11	6.11	3.89	1
2	3.51	3.51	3.51	6.49	1
3	−3.49	−3.49	0	13.49	1
4	2.01	−1.48	0	11.48	2
5	2.38	0.90	0.90	9.10	3
6	−3.58	−3.58	0	13.58	1
7	9.43	5.85	5.85	4.15	2
8	−5.83	−5.83	0	15.83	1
9	7.11	1.28	1.28	8.72	2
10	3.53	3.53	3.53	6.47	1

constrained system as

$$o_t = \begin{cases} S - i_t & \text{if } S - i_t \in \mathcal{A} \\ C_2 & \text{otherwise.} \end{cases} \tag{2}$$

Intuitively, C_1 is the bound for the actual order quantity, and C_2 is the actual order quantity if the desired order quantity falls outside of this bound.

- Under FR, $\mathcal{A} = (C_1 = 0, +\infty)$ and $C_2 = C_1 = 0$. That is, when the desired order quantity is less than zero, the actual order quantity is zero.
- With MOQ, $\mathcal{A} = (C_1 = S - s, +\infty)$ and $C_2 = 0$; when the desired quantity is less than $S - s$, the actual order quantity is zero.
- In the CC setting, $\mathcal{A} = (-\infty, C_p)$ and $C_1 = C_2 = C_p$; when the desired production quantity is greater than the capacity C_p, the actual production quantity is C_p.

The actual order quantity can either be *free* or *constrained*, corresponding to the two cases in (2). An order is *free* when the actual order quantity equals the desired quantity, and it is *constrained* if the actual order quantity equals the fixed value C_2. The level of the impact of the order constraint is determined by the probability that demand falls in the admissible region, $\int_{\mathcal{A}} f_d(x)\mathrm{d}x$. When this integral equals one, the constraint is never binding and the order-constrained system becomes identical to a linear system. We use the notation of the order constraint being *loose* or *tight* to indicate the value of this integral being high or low.

Before the formal analysis, we give an intuitive example of the stopping time problem. Table 2 is adopted from a simulation output in a forbidden returns system where $\mu_d = 5$, $\sigma_d = 5$ and $S = 10$. It shows how the orders are generated from the demand. The desired order quantity is always the sum of demand since the last free order. However, the actual order quantity must be either in the admissible region or be equal to C_2. Specifically, if the sum of demand falls into the admissible region, then the order quantity is the sum of demand since the previous free order. Otherwise, the order quantity is C_2. For instance, in period 4, even if $d_4 > 0$, we still have $o_4 = 0$ because $d_4 + d_3 < 0$; in period 5, $o_5 = 0.9$ as $d_3 + d_4 + d_5 = 0.9$, despite $d_5 = 2.38$.

We will now show the general form (2) can be translated into a stopping time problem. To see this we first define $\{\hat{d}, \hat{o}, \hat{\mathcal{A}}\}$ as $\{d, o, \mathcal{A}\}$ subtracted by C_2, such that $\hat{d} = d - C_2$, $\hat{o} = o - C_2$, and $\hat{\mathcal{A}} = \mathcal{A} - C_2 = \{x - C_2 | x \in \mathcal{A}\}$.

Lemma 1. *Let* $\tau = \min\{k > 0 \mid \hat{o}_{t-k} \in \hat{\mathcal{A}}\}$, *then*

$$\hat{o}_t = \left(\sum_{k=0}^{\tau-1} \hat{d}_{t-k}\right)_{\hat{\mathcal{A}}}. \tag{3}$$

Proof. Note $i_{t-\tau} + o_{t-\tau} = S$ and $\sum_{k=1}^{\tau-1} o_{t-k} = (\tau - 1)C_2$. Thus

$$\begin{aligned} i_t &= i_{t-1} + o_{t-1} - d_t && \text{From (1)} \\ &= i_{t-\tau} + \sum_{k=1}^{\tau} o_{t-\tau} - \sum_{k=0}^{\tau-1} d_{t-k} && \text{Recursively expanding } i_{t-1} \\ &= S + (\tau - 1)C_2 - \sum_{k=0}^{\tau-1} (\hat{d}_{t-k} + C_2) && i_{t-\tau} + o_{t-\tau} = S, o_{t-k} = C_2 \text{ for } k < \tau \\ &= S - \sum_{k=0}^{\tau-1} \hat{d}_{t-k} - C_2. && \text{Collecting together terms} \end{aligned}$$

Substituting the last expression into (2) yields (3). □

To ensure the existence of τ, there needs to be at least one order that belongs to $\mathcal{A}$ in the order history. This can be guaranteed by assuming the recurrence of the random walk $\sum_t d_t$ on $\mathcal{A}$, given by the Chung–Fuchs theorem (Sato, 1999). When $C_2 \neq 0$ (for instance, in the CC system), we only need to subtract C_2 from d, o and $\mathcal{A}$ to transform the problem into the form of (3). Therefore, for notational convenience, we assume $C_2 = 0$ in the subsequent analysis. That is, the constrained order equals zero. However, in the numerical examples, we will present the results from the CC system with $C_2 \neq 0$.

Lemma 1 shows that τ is a stopping time, the time at which $\sum_{k=1}^{\tau-1} d_{t-k}$ falls into $\mathcal{A}$. We define τ as the degree of the order o_t, indicating the most recent admissible order quantity was made τ periods ago (the last column in Table 2). We also define $o[\tau]$ as the order quantity with the degree τ. In the FR and MOQ settings, $\{\tau | o[\tau] \in \mathcal{A}\}$ can be intuitively understood as the length of the ordering cycles, i.e., the number of periods between the current and the previous positive order. Both τ and $o[\tau]$ are time dependent variables, but we suppress the subscript t to avoid notational clutter when no additional confusion is introduced, e.g. when discussing a distribution function. For the subsequent analysis, it is convenient to denote the desired order quantity as $\tilde{o}_t = S - i_t = \sum_{k=0}^{\tau-1} d_{t-k}$. Lemma 1 can then be reiterated as *the order quantity* *is the desired order quantity truncated by* $\mathcal{A}$, or $o_t = (\tilde{o}_t)_{\mathcal{A}}$. By definition, the free order quantity $\tilde{o}_t$ has the same degree as o_t.

3.2. The distribution of order quantity

We derive the distribution of the order quantity as follows. For any $o_t[\tau]$, we have $d_{t-\tau+1} \notin \mathcal{A}$, $d_{t-\tau+1} + d_{t-\tau+2} \notin \mathcal{A}$, …, $\sum_{k=1}^{\tau-1} d_{t-k} \notin \mathcal{A}$. Thus, the distribution of the desired order quantity $\tilde{o}[\tau]$ can be derived by recursive convolution. The distribution of the actual order quantity $o[\tau]$ is then available through the rectification of $\tilde{o}[\tau]$. Here we use rectification in a general sense; rectification refers to a modification to a distribution function when its inadmissible part is reset to zero, and the value of the rectified pdf at C_2 (assumed to be zero) is the integral of the original pdf over the inadmissible region. That is, $f(x \in \mathcal{A}') = 0$ and $f(0) = \int_{\mathcal{A}'} f(x)\mathrm{d}x$. In contrast, a truncated pdf disregards the negative part (see Table 1 for the formal definition of truncation). Since $o_t[\tau]$ are i.i.d., $f_{o[\tau]}$ become component distributions of a mixture, which can be simply summed to yield f_o. We begin by looking at an individual $f_{\tilde{o}[\tau]}$.

When $\tau = 1$, there was a free order one period ago; the desired order quantity in this period equals the demand observed, and the distribution of $\tilde{o}[1]$ is the demand distribution:

$$f_{\tilde{o}[1]}(x) = f_d(x). \tag{4}$$

The distribution of $\tilde{o}[2]$ can be derived as follows. $\tilde{o}_t[2]$ means that its most recent free order was made 2 periods ago. Therefore $d_{t-1} \notin \mathcal{A}$, and the distribution of $\tilde{o}_t[2]$ can be derived by the following convolution:

$$f_{\tilde{o}[2]}(x) = f_d(x) * [f_d(x)]_{\mathcal{A}'} = f_d(x) * [f_{\tilde{o}[1]}(x)]_{\mathcal{A}'},$$

where $[f_d(x)]_{\mathcal{A}'}$ and $[f_{\tilde{o}[1]}(x)]_{\mathcal{A}'}$ are the demand and order distributions truncated by $\mathcal{A}'$. The distribution of $\tilde{o}[3]$ can be derived similarly. Given $d_{t-2} \notin \mathcal{A}$ and $d_{t-2} + d_{t-1} \notin \mathcal{A}$, we have

$$f_{\tilde{o}[3]}(x) = f_d(x) * [f_{\tilde{o}[2]}(x)]_{\mathcal{A}'}.$$

Generally, the distribution of $\tilde{o}[\tau]$ ($\tau > 1$) can be derived by recursive convolution as

$$f_{\tilde{o}[\tau]}(x) = f_d(x) * [f_{\tilde{o}[\tau-1]}(x)]_{\mathcal{A}'}. \tag{5}$$

The distribution of $o[\tau]$ is then the truncation of the distribution of $\tilde{o}[\tau]$, $f_{o[\tau]}(x) = [f_{\tilde{o}[\tau]}(x)]_{\mathcal{A}}$. It is known that if the demand follows a normal distribution, then $\tilde{o}[2]$ follows the skew-normal distribution (Azzalini, 1985; Henze, 1986), but the type of distribution for $\tilde{o}[\tau]$ is generally undefined when the demand distribution is unknown or when $\tau > 2$.

For the probability of actual order quantity, we need to derive the probability of each component. We denote the probability that a free order has degree τ as p_τ where

$$p_\tau = \int_{\mathcal{R}} f_{o[\tau]}(x)\mathrm{d}x = \int_{\mathcal{A}} f_{\tilde{o}[\tau]}(x)\mathrm{d}x. \tag{6}$$

Note, the integration of $f_{\tilde{o}[\tau]}(x)$ over $\mathcal{R}$, although not truncated, is not unity. This is because $[f_{\tilde{o}[\tau-1]}(x)]_{\mathcal{A}'}$ is not normalized. In this sense, $[f_{\tilde{o}[\tau]}(x)]_{\mathcal{A}'}$ is not a truncated distribution as conventionally defined, which requires $F_{\tilde{o}[\tau]}(+\infty) = 1$. In fact, the integration of $f_{\tilde{o}[\tau]}$ gives the probability that *any* order has a degree of τ, in other words, the time difference between two consecutive free orders is at least τ. This can be seen as the unconstrained counterpart of p_τ, thus can be denoted $\tilde{p}_\tau$. Intuitively,

$$\begin{aligned}\tilde{p}_\tau &= \int_{\mathcal{R}} f_{\tilde{o}[\tau]}(x)\mathrm{d}x = \int_{\mathcal{R}} f_d(x) * [f_{\tilde{o}[\tau-1]}(x)]_{\mathcal{A}'}\mathrm{d}x \\ &= \int_{\mathcal{R}} f_d(x)\mathrm{d}x \int_{\mathcal{A}'} f_{\tilde{o}[\tau-1]}(x)\mathrm{d}x = \int_{\mathcal{A}'} f_{\tilde{o}[\tau-1]}(x)\mathrm{d}x = 1 - \sum_{k=1}^{\tau-1} p_k.\end{aligned}$$

The last equality can be easily verified by induction. Note, orders with degree one do not have any preceding constrained orders; orders with degree two have one preceding constrained order, and so on. Thus, the average time difference between two successive free orders is $\sum_k kp_k$.

Since the average length of the zero order series is $\sum_k (k-1)p_k$, instantly we have:

$$\Pr\{o = 0\} = \frac{\sum_{k=1}^{\infty}(k-1)p_k}{\sum_{k=1}^{\infty} kp_k}, \text{ and}$$

$$\Pr\{o \in \mathcal{A}\} = 1 - \frac{\sum_{k=1}^{\infty}(k-1)p_k}{\sum_{k=1}^{\infty} kp_k} = \frac{1}{\sum_{k=1}^{\infty} kp_k}.$$

We can then derive the expression for the distribution of orders:

Proposition 1. *In the order-constrained system, the distribution of order quantity is*

$$f_o(x) = \frac{\delta(x)\sum_{k=1}^{\infty}(k-1)p_k + \sum_{k=1}^{\infty} f_{o[k]}(x)}{\sum_{k=1}^{\infty} kp_k}. \tag{7}$$

Here $\delta(x)$ is the Dirac Delta function with $\delta(0) = 1$ and $\delta(x) = 0$ for $x \neq 0$.

The distribution of order quantity is a mixture distribution with a discrete $o_t = 0$ part and a continuous $o_t \in \mathcal{A}$ part. Schultz (1983) provides the order quantity distribution under an (s, S) policy and integer, nonnegative demand. Schultz's density function has a similar structure to (7), where $\sum_k (k-1)p_k$ is analogous to the renewal function representing the average number of periods the accumulative demand is above s. The difference is that when demand is allowed to be negative, the renewal model can no longer be applied.

Fig. 3 shows the distribution of the order quantity in the three order-constrained systems. In the FR system, $\mathcal{A} = (0, +\infty)$. In the MOQ system, $\mathcal{A} = (2, +\infty)$. In the CC system, $\mathcal{A} = (-\infty, 10)$. The order distributions are also compared with the demand distribution, which is assumed to be Gaussian, with $\mu_d = 5$ and $\sigma_d = 5$. In the FR system, the mode of o_t no longer equals μ_d. It is slightly smaller. This can be intuitively explained as follows. The mean of $\tilde{o}[1]$ equals μ_d. For $\tilde{o}[\tau]$ where $\tau > 1$, as $\tilde{o}_t[\tau] = \sum_{k=0}^{\tau-1} d_{t-k}$ and $\sum_{k=1}^{\tau-1} d_{t-k} < 0$, we have $\mathbb{E}(\tilde{o}[\tau]) < \mu_d$ where $\mathbb{E}(\cdot)$ is the expectation operator. This gives $f_o(x)$ a negative skew. This effect is more prominent when the order constraint becomes tighter. In the CC system, the admissible region is symmetric to the admissible region in the FR system about μ_d, therefore the order distribution under CC is a horizontal reflection of the order distribution under FR about μ_d. In the MOQ system, the inadmissible region $\mathcal{A}'$ contains both negative and positive parts and it is generally indefinite whether $\mathbb{E}(\tilde{o}[\tau])$ is monotone in τ or not. Although $\mathcal{A}$ is tighter in the MOQ setting than in FR setting, the skewness does not change significantly. To illustrate the above analysis, Fig. 4 shows the first ten component distributions, $\tilde{o}[\tau \leqslant 10]$, under the same settings as in Fig. 3. Each point represents one component distribution, the horizontal and vertical axes show their respective mean and standard deviation, and the size of the points are proportional to their probabilities.

3.3. The distribution of inventory

With unit lead-time, the distribution of inventory is straightforward to derive from the order distribution. From (1), (2), the inventory iteration can be written as

$$i_t = \begin{cases} S - d_t & \text{if } o_{t-1} \in \mathcal{A} \\ i_{t-1} - d_t & \text{otherwise.} \end{cases} \tag{8}$$

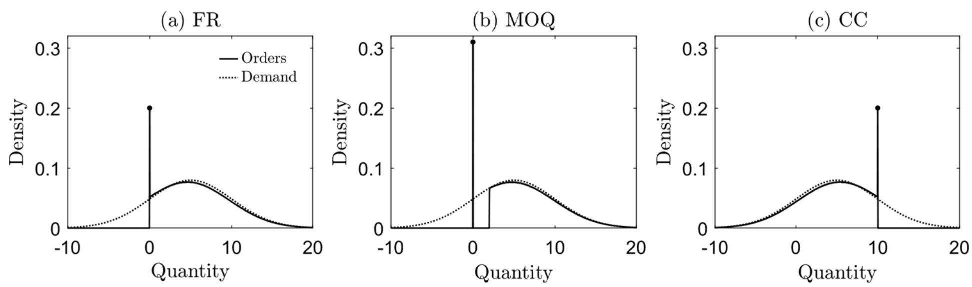

Fig. 3. Order distribution in order-constrained systems, $d \sim N(5,5)$, $S = 10$. (a) The FR system; (b) the MOQ system with $s = 8$. (c) The CC system with $C_p = 10$.

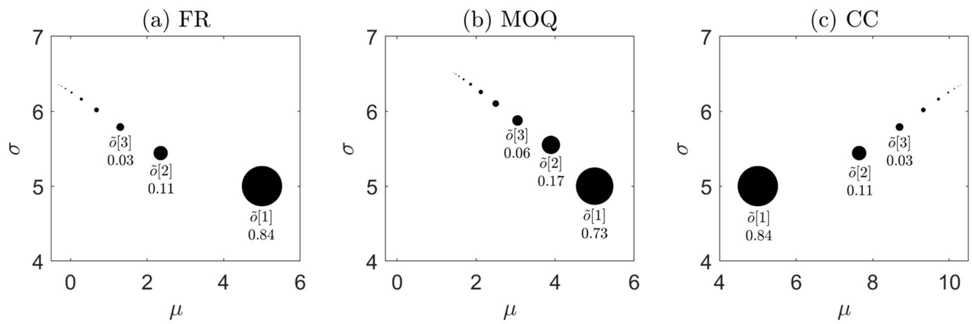

Fig. 4. Point process representation of the component distributions, $d \sim N(5,5)$, $S = 10$. (a) The FR system; (b) the MOQ system with $s = 8$. (c) The CC system with $C_p = 10$. The size of the circle is proportional to the probability of the component distribution. The probabilities are given under the circles.

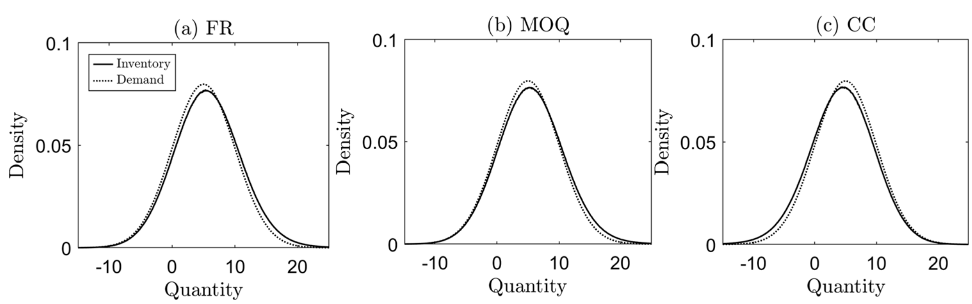

Fig. 5. Distribution of the inventory levels in order-constrained systems, $d \sim N(5,5)$, $S = 10$. (a) The FR system; (b) the MOQ system with $s = 8$. (c) The CC system with $C_p = 10$.

This can be formulated with the stopping time $\tau = \min\{k > 0 | o_{t-k} \in \mathcal{A}\}$ as

$$i_t[\tau] = S - \sum_{k=0}^{\tau-1} d_{t-k} = S - \tilde{o}_t[\tau]. \tag{9}$$

Eq. (9) shows the inventory level in any given period equals S minus the desired order quantity in that period. This can also be seen in the numerical illustration in Table 2. Further, Eq. (9) shows the distribution of inventory is also a mixture distribution. The main difference is that the inventory level is not constrained. That is, the component distributions are not truncated. Generally, $f_{i[\tau]}(S-x) = f_{\tilde{o}[\tau]}(x)$ holds for all $\tau > 1$, that is, $f_{i[\tau]}$ is $f_{\tilde{o}[\tau]}$ reflected about $S/2$. Therefore we have,

Proposition 2. *In the forbidden returns system, the distribution of the inventory is*

$$f_i(S-x) = \frac{\sum_k f_{\tilde{o}[k]}(x)}{\sum_k k p_k}, \tag{10}$$

where $k = 1, \ldots, \infty$.

Fig. 5 shows the inventory distributions in the same setting as in Fig. 3. Here $S - \mu_d = 5$ so the inventory distribution in the linear system is the same as the demand distribution. Since the inventory distributions are continuous and smooth, their nature can be illustrated more precisely with the descriptive statistics shown in Table 3. We see the FR inventory distribution exhibits both a higher mean and mode and a wider spread (higher variance) than the equivalent linear inventory distribution; which causes the peak density to decrease. It is also positively skewed and leptokurtic (the kurtosis larger than 3, meaning the distribution has fatter tails than the Gaussian). The CC inventory distribution is symmetrical to the FR inventory distribution. In the MOQ system, the monotonicity of $\mathbb{E}(\tilde{o}[\tau])$ cannot be guaranteed and we cannot establish how the inventory distribution will change from the Gaussian. In this example, the MOQ constraint has a milder impact on the mean and mode of the inventory distribution compared with the FR constraint.

Table 3
Characteristics of the inventory distribution in order-constrained systems.

	Gaussian	FR	MOQ	CC
Mean	5	5.63	5.48	4.37
Standard deviation	5	5.31	5.33	5.31
Mode	5	5.48	5.32	4.52
Density at mode	0.080	0.077	0.077	0.077
Skewness	0	0.15	0.15	−0.15
Kurtosis	3	3.25	3.25	3.25

4. Extensions: Transportation lead time and correlated demand

Having explored the base case for unit lead time and i.i.d. customer demand in Section 3, we now extend our analysis to examine the cases of arbitrary lead-time and auto-correlated demand.

4.1. i.i.d. demand and a transportation delay

Under the base stock policy, the transportation delay does not affect the distribution of the order quantity when the demand is i.i.d. In cases where the lead time is longer than one period, we only need to replace the inventory level i_t by the inventory position IP_t in (1) and (2), and the results in Section 3.2 hold as the order distribution does not change. However, the delay does affect the inventory distribution, as the inventory balance equation now becomes

$$i_t = i_{t-1} + o_{t-L} - d_t. \tag{11}$$

As in Section 3.3, now we need to express the inventory level as the sum of past demand, from which we have the following result:

Lemma 2. *In the order-constrained system where* $L > 1$, $i_t = S - \sum_{k=0}^{L+\upsilon-2} d_{t-k}$, *where the stopping time* $\upsilon = \min\{k > 0 \mid o_{t-L-k+1} \in \mathcal{A}\}$.

Proof. Define as before $\tau = \min\{k > 0 \mid o_{t-k} \in \mathcal{A}\}$. From (11), we have

$$i_t = IP_t + w_t = IP_{t-\tau} + \sum_{k=1}^{\tau} o_{t-k} - \sum_{k=0}^{\tau-1} d_{t-k} - \sum_{k=1}^{L-1} o_{t-k}. \tag{12}$$

The derivation of the last step utilises the inventory position balance equation, $IP_t = IP_{t-\tau} + \sum_{k=1}^{\tau} o_{t-k} - \sum_{k=0}^{\tau-1} d_{t-k}$, and that $w_t = \sum_{k=1}^{L-1} o_{t-k}$. Since $o_{t-\tau} \in \mathcal{A}$, we have $IP_{t-\tau} + o_{t-\tau} = S$. So (12) becomes

$$i_t = S - \sum_{k=\tau}^{L-1} o_{t-k} - \sum_{k=0}^{\tau-1} d_{t-k}.$$

Again, since $o_{t-\tau} \in \mathcal{A}$, we have $\sum_{k=\tau}^{L-1} o_{t-k} = \sum_{k=\tau}^{L+\upsilon-2} d_{t-k}$, where $\upsilon = \min\{k > 0 | o_{t-L-k+1} \in \mathcal{A}\}$. The result then directly follows. □

Lemma 2 allows us to obtain the distribution of inventory when a lead time is present in a manner similar to the one introduced in Section 3. In Lemma 2, υ is defined as the degree of inventory. $\upsilon = 1$ means that the most recent free order before period $t-L$ is o_{t-L}, i.e., $o_{t-L} \in \mathcal{A}$. The distribution of $i[1]$ equals the distribution of lead-time demand because all the demand that occurred from period $t-L+1$ to period t contributes to i_t irrespective of the value of demand:

$$f_{i[1]}(S-x) = f_d^L(x), \tag{13}$$

where $f_d^L(x)$ denotes the lead-time demand distribution; that is, $f_d^L(x)$ is $f_d(x)$ convolved L times. The inventory level with the second degree $i[2]$ means that $o_{t-L-1} \in \mathcal{A}$, $o_{t-L} = 0$, therefore $d_{t-L} \notin \mathcal{A}$. The distribution of $i[2]$ is thus the demand distribution truncated by $\mathcal{A}'$, $[f_d(x)]_{\mathcal{A}'}$, convolved with the lead-time demand distribution, $f_d^L(x)$:

$$f_{i[2]}(S-x) = f_d^L(x) * [f_d(x)]_{\mathcal{A}'} = f_d^L(x) * [f_{\tilde{o}[1]}(x)]_{\mathcal{A}'}.$$

Observe, if i_t is of degree $\tau > 1$, then we have $i_t = \sum_{k=0}^{L-1} d_{t-k} + \tilde{o}_{t-L}$. Also, o_{t-L} is constrained and $\tilde{o}_{t-L}$ is of degree $\tau-1$ because the most recent free order before $t-L$ was $\tau-1$ periods ago. The above result can be generalized to

$$f_{i[\tau]}(S-x) = f_d^L(x) * [f_{\tilde{o}[\tau-1]}(x)]_{\mathcal{A}'}. \tag{14}$$

The overall inventory distribution is a mixture of the above component distributions which has the same form as (10):

$$f_i(x) = \frac{\sum_k f_{i[k]}(x)}{\sum_k k p_k}.$$

We can now provide a general representation for the order and inventory distributions under constrained order and arbitrary lead-time. Let $g_1 = f_d$ be the distribution of demand and $h_1 = f_d^L$ be the distribution of lead-time demand. In the linear inventory system, they give the order and inventory distribution respectively. In the order-constrained system, for $k = 2, \dots, \infty$, construct the function series by recursive convolution:

$$g_k = (g_{k-1})_{\mathcal{A}'} * g_1 \text{ and } h_k = (g_{k-1})_{\mathcal{A}'} * h_1.$$

Then

$$f_{\tilde{o}}(x) = \frac{\sum_{k=1}^{\infty} g_k(x)}{\sum_{k=1}^{\infty} k p_k} \text{ and } f_i(S-x) = \frac{\sum_{k=1}^{\infty} h_k(x)}{\sum_{k=1}^{\infty} k p_k}.$$

4.2. Correlated demand

An important issue in inventory theory is the temporal autocorrelation between consecutive demands. It is relatively easy to incorporate demand autocorrelation into linear inventory systems, as it does not affect the linearity of the system. In our model, we achieve this by extending the previous analysis to the multivariate case. Under the assumption that demand is ARMA(p,q), it follows a multivariate distribution $f_d(\mathbf{x}|\mathrm{M}, \Sigma)$, where $\mathbf{x}$ is the demand vector, M is the expectation and Σ is the autocovariance matrix of $\mathbf{x}$. In fact, the i.i.d. demand case described in Section 3 is simply $\mathrm{M} = \mu_d \mathbf{1}$ and $\Sigma = \sigma_d I$, where $\mathbf{1}$ is an all-one vector and I is the identity matrix, both with proper dimensions.

In the case of correlated demand, $\mathrm{M} = \mu_d \mathbf{1}$ remains unchanged as the stationarity assumption is still in place. We only need to specify the covariance matrix Σ. For an ARMA(p, q) demand model, the demand is generated by the following equations:

$$\begin{aligned} y_t &= \Phi y_{t-1} + \Theta \varepsilon_t, \\ d_t &= Z y_t + \mu_d, \end{aligned} \tag{15}$$

where

$$\Phi = \begin{pmatrix} \phi_1 & 1 & & \\ \phi_2 & & \ddots & \\ \vdots & & & 1 \\ \phi_r & & & \end{pmatrix}$$

is a square matrix representing the auto-regression,

$$\Theta = \begin{pmatrix} 1 & \theta_1 & \dots & \theta_r \end{pmatrix}^T$$

is the input matrix representing the moving average, Z is a row vector in the form of $(1\ 0\ \dots\ 0)$, and $r = \max(p, q)$. ε_t is a scalar i.i.d. random variable (Harvey, 1990). Under this model, the elements of Σ are $\sigma_{jk}^2 = \mathbb{E}(d_j d_k) - \mu_d^2 = Z\mathbb{E}(y_j y_k^T)Z^T$. In the case of $k \leqslant j$, we can write y_j as

$$\begin{aligned} y_j &= \Phi y_{j-1} + \Theta \varepsilon_j \\ &= \Phi^2 y_{j-2} + \Phi \Theta \varepsilon_{j-1} + \Theta \varepsilon_j \\ &= \dots \\ &= \Phi^{j-k} y_k + \sum_{l=1}^{j-k-1} \Phi^{j-k-l} \Theta \varepsilon_{k+l} + \Theta \varepsilon_j. \end{aligned}$$

Therefore $\mathbb{E}(y_j y_k^T) = \Phi^{j-k} \Sigma_{yy}$, where Σ_{yy} is the covariance of y. Σ_{yy} can be directly derived from Φ and Θ using the Kronecker product, see Wang & Disney (2017).

In the case of $j < k$, write y_k as

$$y_k = \Phi^{k-j} y_j + \sum_{l=1}^{k-j-1} \Phi^{k-j-l} \Theta \varepsilon_{j+l} + \Theta \varepsilon_k,$$

from which we have $\mathbb{E}(y_j y_k^T) = \Sigma_{yy} (\Phi^{k-j})^T$. So the elements of Σ can be written as

$$\sigma_{jk}^2 = \begin{cases} Z\Phi^{j-k}\Sigma_{yy}Z^T & \text{if } k \leqslant j \\ Z\Sigma_{yy}(\Phi^{k-j})^T Z^T & \text{if } k > j. \end{cases} \tag{16}$$

Next, we need to rewrite the component distributions in the multivariate form. Define the following partitions:

$$\mathcal{O}[\tau] = \left\{ \mathbf{x} \middle| \sum_{j=1}^{k} x_j \in \mathcal{A}', \ 1 < k < \tau - 1 \right\} \subset \mathcal{R}^{\tau-1},$$

$$\mathcal{P}[\tau] = \left\{ \mathbf{x} \middle| \sum_{j=1}^{k} x_j \in \mathcal{A}', \ 1 < k < \tau - 1 \right\} \subset \mathcal{R}^{\tau},$$

and

$$\mathcal{I}[\tau] = \left\{ \mathbf{x} \middle| \sum_{j=1}^{k} x_j \in \mathcal{A}', \ 1 < k < \tau - 1 \right\} \subset \mathcal{R}^{L+\tau-1}.$$

We have

$$f_{\tilde{o}[\tau]}(x) = \int_{\mathcal{O}[\tau]} f_d(\mathbf{x}) \mathrm{d}\mathbf{x},$$

$$\tilde{p}_\tau = \int_{\mathcal{P}[\tau]} f_d(\mathbf{x}) \mathrm{d}\mathbf{x},$$

and

$$f_{i[\tau]}(S - x) = \int_{\mathcal{I}[\tau]} f_d(\mathbf{x}) \mathrm{d}\mathbf{x},$$

where $\mathbf{x}$ is the demand vector with dimensions of τ, τ and $L + \tau$ respectively. The distribution of order quantity and inventory can then be derived using the approached described in Sections 3 and Section 4.1.

In a linear OUT policy, the minimum mean square error (MMSE) forecast of the ARMA demand minimises the variance of inventory level and the inventory cost. However, in the order constrained policy, this result is yet to be established. Meanwhile, for general ARMA demand, adopting MMSE forecasts causes the order-up-to level S to be time-varying, complicating analysis (as it will create a time-varying admissible region). We leave this for future research, and proceed herein with a constant order-up-to level.

5. An approximation algorithm

It can be seen that the computation of the order and inventory distribution and their moments is difficult due to the fact that it involves the calculation of a countably infinite number of component distributions and truncated distributions. Numerically, convolution of density functions requires discretization, the granularity of which greatly impacts the accuracy of the computation. In this section, we introduce an approximation algorithm which allows us to compute the mean and variance of order quantity and inventory much faster than conducting numerical convolutions. This algorithm works under the assumptions that the demand, and all component distributions are Gaussian. The reason for choosing this assumption is that the variance of a truncated *Gaussian* distribution can be written in closed form facilitating the recursive computation. However, it is important to note the component distributions are not Gaussian. Therefore, the accuracy of the algorithm depends on the divergence between the component distributions and the Gaussian distribution.

This algorithm builds upon the results from Nurminen, Rui, Ardeshiri, Bazanella, & Gustafsson (2016), who derived the first- and second-order moments of a multivariate normal distribution where the first variable is truncated. We first introduce several variables that will be used in the algorithm. Let M and Σ be the mean vector and the co-variance matrix of the multivariate distribution. The subscript $\mathcal{A}$ denotes the truncation operation by $\mathcal{A}$. Let $m_{\mathcal{A}}$ be the mean truncation coefficient and $s_{\mathcal{A}}$ the variance truncation coefficient. They are calculated as

$$m_{\mathcal{A}} = \frac{\phi(\underline{z}) - \phi(\bar{z})}{\Phi(\bar{z}) - \Phi(\underline{z})} \tag{17}$$

and

$$s_{\mathcal{A}} = 1 + \frac{\underline{z}\phi(\underline{z}) - \bar{z}\phi(\bar{z})}{\Phi(\bar{z}) - \Phi(\underline{z})} - (m_{\mathcal{A}})^2, \tag{18}$$

where $\underline{z} = (\inf \mathcal{A} - \mu_1)/\lambda_{1,1}$ and $\bar{z} = (\sup \mathcal{A} - \mu_1)/\lambda_{1,1}$ can be understood as the standardized infimum and supremum of $\mathcal{A}$. μ_1 and $\lambda_{1,1}$ are the elements of M and Λ at position 1 and (1,1) respectively. The lower triangular matrix Λ is the Cholesky decomposition of Σ such that $\Sigma = \Lambda\Lambda^T$. Therefore μ_1 and $\lambda_{1,1}$ are the mean and standard deviation of the first element.

Lemma 3. *Suppose* $\begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix}$ *is a random vector following a multivariate normal distribution with mean* M *and covariance* Σ*, then*

$$\mathbb{E}\begin{pmatrix} (x_1)_{\mathcal{A}} & x_2 & \dots & x_n \end{pmatrix} = \Lambda \begin{pmatrix} m_{\mathcal{A}} \\ \mathbf{0} \end{pmatrix} + \mathrm{M},$$

$$\mathbb{V}\begin{pmatrix} (x_1)_{\mathcal{A}} & x_2 & \dots & x_n \end{pmatrix} = \Lambda \begin{pmatrix} s_{\mathcal{A}} & \mathbf{0} \\ \mathbf{0} & I \end{pmatrix} \Lambda^T.$$

The logic of this algorithm is to derive the conditional mean and variance of the demand vector $\begin{pmatrix} d_t \in \mathcal{A} & d_{t+1} & \dots & d_{t+n} \end{pmatrix}$ and $\begin{pmatrix} d_t \in \mathcal{A}' & d_{t+1} & \dots & d_{t+n} \end{pmatrix}$ from the unconditional mean and variance of $\begin{pmatrix} d_t & d_{t+1} & \dots & d_{t+n} \end{pmatrix}$, using Lemma 3. The conditional mean and variance are then used to approximate the mean, variance, and probability of the component distributions $\tilde{o}[\tau]$ and $i[\tau]$. To initialize the algorithm, we take an $n+1$ dimensional demand vector $\begin{pmatrix} d_t & d_{t+1} & \dots & d_{t+n} \end{pmatrix}$. Its mean M and the variance Σ can be derived with the method specified in Section 4.2. We denote them as M_d and Σ_d for consistency. Since the covariance matrix of a general ARMA demand is not diagonal, the arbitrarily chosen n affects the accuracy and efficiency of the approximation. We also need to introduce several binary matrices for the calculation. Let $u_k = \begin{pmatrix} 1 & \mathbf{0}_{1\times(n-k-1)} \end{pmatrix}$, $v_k = \begin{pmatrix} \mathbf{1}_{1\times(L+1)} & \mathbf{0}_{1\times(n-L-k-1)} \end{pmatrix}$, $\mathrm{Q}_1 = \begin{pmatrix} \mathbf{0}_{(n-1)\times 1} & I \end{pmatrix}$, $\mathrm{Q}_{k>1} = \begin{pmatrix} u_k^T & I \end{pmatrix}$.

In the first step, M_1 and Σ_1 give the mean and covariance of $\begin{pmatrix} d_{t+1} & \dots & d_{t+n} \end{pmatrix}$ conditional on $d_t \in \mathcal{A}$:

$$\mathrm{M}_1 = \mathrm{Q}_1 \left[\Lambda_d \begin{pmatrix} (m_d)_{\mathcal{A}} \\ \mathbf{0} \end{pmatrix} + \mathrm{M}_d \right], \tag{19}$$

$$\Sigma_1 = \mathrm{Q}_1 \Lambda_d \begin{pmatrix} (s_d)_{\mathcal{A}} & \mathbf{0} \\ \mathbf{0} & I \end{pmatrix} \Lambda_d^T \mathrm{Q}_1^T. \tag{20}$$

Q_1 is an $(n-1) \times n$ matrix composed of an all-zero first column and an identity matrix. It reduces the dimension of the vector (or the matrix) by one by deleting the first element (or the first row and column). Since $\tilde{o}_t[1] = d_t \mid d_{t-1} \in \mathcal{A}$, the mean of $\tilde{o}[1]$ is given by the first element of M_1, and its variance is given by the first diagonal element of Σ_1:

$$\mu_1 = \mathbb{E}(\tilde{o}[1]) = u_1 \mathrm{M}_1,$$

$$\sigma_1^2 = \mathbb{V}(\tilde{o}[1]) = u_1 \Sigma_1 u_1^T.$$

The first element of u_j is one and zero otherwise. It takes the first (diagonal) element of the respective mean vector (covariance matrix). Note, under autocorrelated demand, we don't have $\mathbb{E}(\tilde{o}[1]) = \mu_d$ and $\mathbb{V}(\tilde{o}[1]) = \sigma_d^2$ anymore as $o_{t+1}[1] = d_{t+1}$ is conditional on $d_t \in \mathcal{A}$. This can also be seen in (19) and (20) as Λ is no longer diagonal.

In the next step, M_1 and Σ_1 are used to update $(m_1)_{\mathcal{A}'}$, $(s_1)_{\mathcal{A}'}$ and Λ_2. We can then derive the conditional mean and variance of the vector $\begin{pmatrix} d_{t+1}+d_{t+2} & \dots & d_{t+n}\end{pmatrix}$ conditional on $d_t \in \mathcal{A}$ and $d_{t+1} \in \mathcal{A}'$:

$$\mathrm{M}_2 = Q_2\left[\Lambda_1\begin{pmatrix}(m_1)_{\mathcal{A}'}\\ \mathbf{0}\end{pmatrix} + \mathrm{M}_1\right], \tag{21}$$

$$\Sigma_2 = Q_2\Lambda_1\begin{pmatrix}(s_1)_{\mathcal{A}'} & \mathbf{0}\\ \mathbf{0} & I\end{pmatrix}\Lambda_1^T Q_2^T. \tag{22}$$

The matrix Q_2 is $(n-2)\times(n-1)$ which reduces the dimensions of M_1 and Σ_1 by one by adding up the first two elements of M_1 and the first two rows and columns of Σ_1. Since $\tilde{o}_t[2] = d_{t-1} + d_t$ where $d_{t-1} \in \mathcal{A}'$ and $d_{t-2} \in \mathcal{A}$, the first element of M_2 and Σ_2 are the mean and variance of $\tilde{o}[2]$:

$$\mu_2 = \mathbb{E}(\tilde{o}[2]) = u_2\mathrm{M}_2,$$

$$\sigma_2^2 = \mathbb{V}(\tilde{o}[2]) = u_2\Sigma_2 u_2^T.$$

We assume here that $\tilde{o}[2]$ is normally distributed, but it is in fact skew-normal. Hence μ_2 and σ_2^2 are only approximations to the real mean and variance of $\tilde{o}_2$.

Generally, the iteration at step k is performed as follows:

$$\mathrm{M}_k = Q_k\left[\Lambda_{k-1}\begin{pmatrix}(m_{k-1})_{\mathcal{A}'}\\ \mathbf{0}\end{pmatrix} + \mathrm{M}_{k-1}\right], \tag{23}$$

$$\Sigma_k = Q_k\Lambda_{k-1}\begin{pmatrix}(s_{k-1})_{\mathcal{A}'} & \mathbf{0}\\ \mathbf{0} & I\end{pmatrix}\Lambda_{k-1}^T Q_k^T. \tag{24}$$

Eqs. (23) and (24) approximate the mean and covariance of $\begin{pmatrix}\sum_{j=1}^{k} d_{t+j} & d_{t+k+1} & \dots & d_{t+n}\end{pmatrix}$ conditional on $d_t \in \mathcal{A}$, $d_{t+1} \in \mathcal{A}'$, $d_{t+1}+d_{t+2} \in \mathcal{A}'$ up until $\sum_{j=1}^{k-1} d_{t+j} \in \mathcal{A}'$. The mean and variance of $\tilde{o}[k]$ is approximated by the first element of M_k and the first diagonal element of Σ_k respectively:

$$\mu_k = \mathbb{E}(\tilde{o}[k]) = u_k\mathrm{M}_k,$$

$$\sigma_k^2 = \mathbb{V}(\tilde{o}[k]) = u_k\Sigma_k u_k^T.$$

The above procedure is easy to understand, recalling that each untruncated component distribution is derived by convolving the demand distribution and the previous component distribution truncated by $\mathcal{A}'$. μ_k and σ_k will be used to calculate the truncation coefficients $(m_k)_{\mathcal{A}}$ and $(s_k)_{\mathcal{A}}$ for the next iteration.

The distribution of actual order quantity with degree k is the truncation of the kth untruncated component distribution by $\mathcal{A}$, $o_t[k] = (\tilde{o}[k])_{\mathcal{A}}$. Its mean and variance can be approximated as follows:

$$\mathbb{E}(o[k]) = u_k\left[\Lambda_k\begin{pmatrix}(m_k)_{\mathcal{A}}\\ \mathbf{0}\end{pmatrix} + \mathrm{M}_k\right],$$

$$\mathbb{V}(o[k]) = u_k\Lambda_k\begin{pmatrix}(s_k)_{\mathcal{A}} & \mathbf{0}\\ \mathbf{0} & I\end{pmatrix}\Lambda_k^T u_k^T.$$

The inventory level with degree $k > 1$ equals the order-up-to level S minus the lead-time demand,

$$\mathbb{E}(S - i[k]) = v_k\mathrm{M}_k,$$

$$\mathbb{V}(i[k]) = v_k\Sigma_k v_k^T.$$

The vector v_k is used to sum up the first $(L+1)$ elements of M_k (or the first $(L+1)\times(L+1)$ sub-matrices of Σ_k) to account for lead-time demand. When $k=1$, the mean and variance of $i[1]$ equals the mean (reflected at $S/2$) and variance of lead-time demand.

Lastly, the probability of components with degree k is approximated by

$$p_k = [\Phi(\bar{z}_k) - \Phi(\underline{z}_k)]\prod_{j=1}^{k-1}\left[1 - \Phi(\bar{z}_j) + \Phi(\underline{z}_j)\right].$$

where $\underline{z}_k = (\inf\mathcal{A} - \mu_k)/\sigma_k$ and $\bar{z}_k = (\sup\mathcal{A} - \mu_k)/\sigma_k$. We use $\Phi(\bar{z}_k) - \Phi(\underline{z}_k)$ to approximate the probability that $\tilde{o}_k \in \mathcal{A}$. The probability of constrained orders $o_t = 0$ can then be calculated via the equations given in Section 3.2. The mean and variance of orders are thus available via those of a mixture distribution, following Wald's equation and the Blackwell–Girshick equation:

$$\mathbb{E}(x) = \mathbb{E}(\mathbb{E}(x[k])),$$

$$\mathbb{V}(x) = \mathbb{E}(\mathbb{V}(x[k])) + \mathbb{V}(\mathbb{E}(x[k])),$$

$x \in \{o, i\}$. The constant order quantities should also be included. Fig. 6 shows the iterative relationship between variables involved in this algorithm. We name this algorithm Truncated Gaussian Convolution (TGC), as it assumes that all components are Gaussian and is based on truncated Gaussian convolution. The only time consuming operation involved in this algorithm is the Cholsky decomposition, with a complexity of $\mathcal{O}(n^3)$. We will evaluate the effectiveness and the efficiency of the TGC together with the economic analysis in Section 6.

6. Numerical analysis and economic implications

The purpose of the numerical analysis is to demonstrate the performance of the TGC algorithm and to reveal the economic impact of the order constraint. For the former, we test the TGC algorithm in terms of approximation accuracy, computational efficiency, and robustness under other demand distributions. For the latter, we conduct a simulation-based analysis via three performance measures prominent in inventory control: the order variance amplification phenomenon, a.k.a. the bullwhip effect, the trade-off between the order-up-to level and service level, and the trade-off between order and inventory variance amplification. The analysis encompasses scenarios of order constraints (FR, MOQ, and CC), auto-correlated demand and arbitrary lead-time. Prior to presenting the analysis, we reemphasize the equivalence between the FR and CC systems under symmetric demand distributions. Therefore, adjusting the mean demand in the FR system will have the same effect as adjusting the capacity in the CC system.

We compare TGC with the computing methods of: numerical convolution, simulation, and two intuitive approximations, namely, the truncated demand (TD) to approximate order distribution (as the order distribution is the same as the demand distribution in the linear system), and lead-time demand (LTD) to approximate the inventory distribution (as the inventory distribution is a reflected and translated lead-time demand distribution in the linear system). Schneider et al. (1995) proposed the following approximation for the order variance in the (s,S) system based on the mean and variance of demand (we label it as SRK after the authors' names)

$$\sigma_o^2 \approx \sigma_d^2 + \frac{2\mu_d^2(S-s)^2}{\mu_d^2 + 2\mu_d(S-s) + \sigma_d^2}.$$

See also Kelle & Milne (1999) for an application of the SRK approximation in a supply chain context. We include the SRK approximations in the bullwhip analysis as well. This approximation

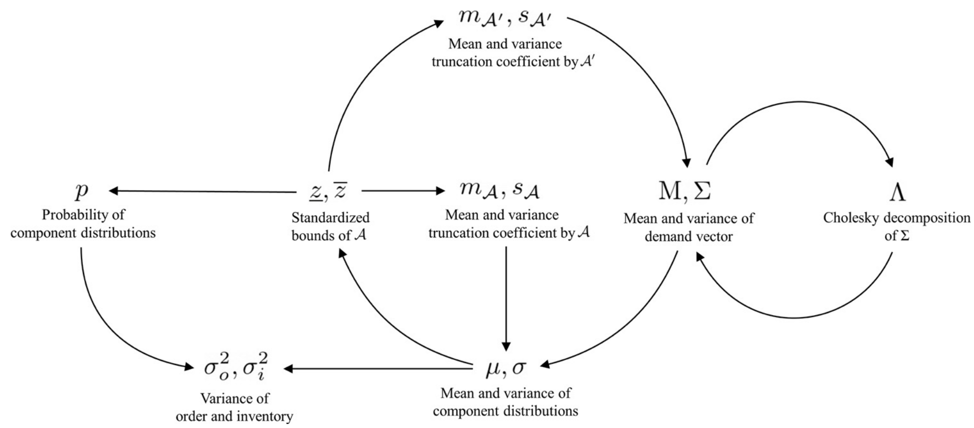

Fig. 6. The iterative relationship of variables in the TGC algorithm.

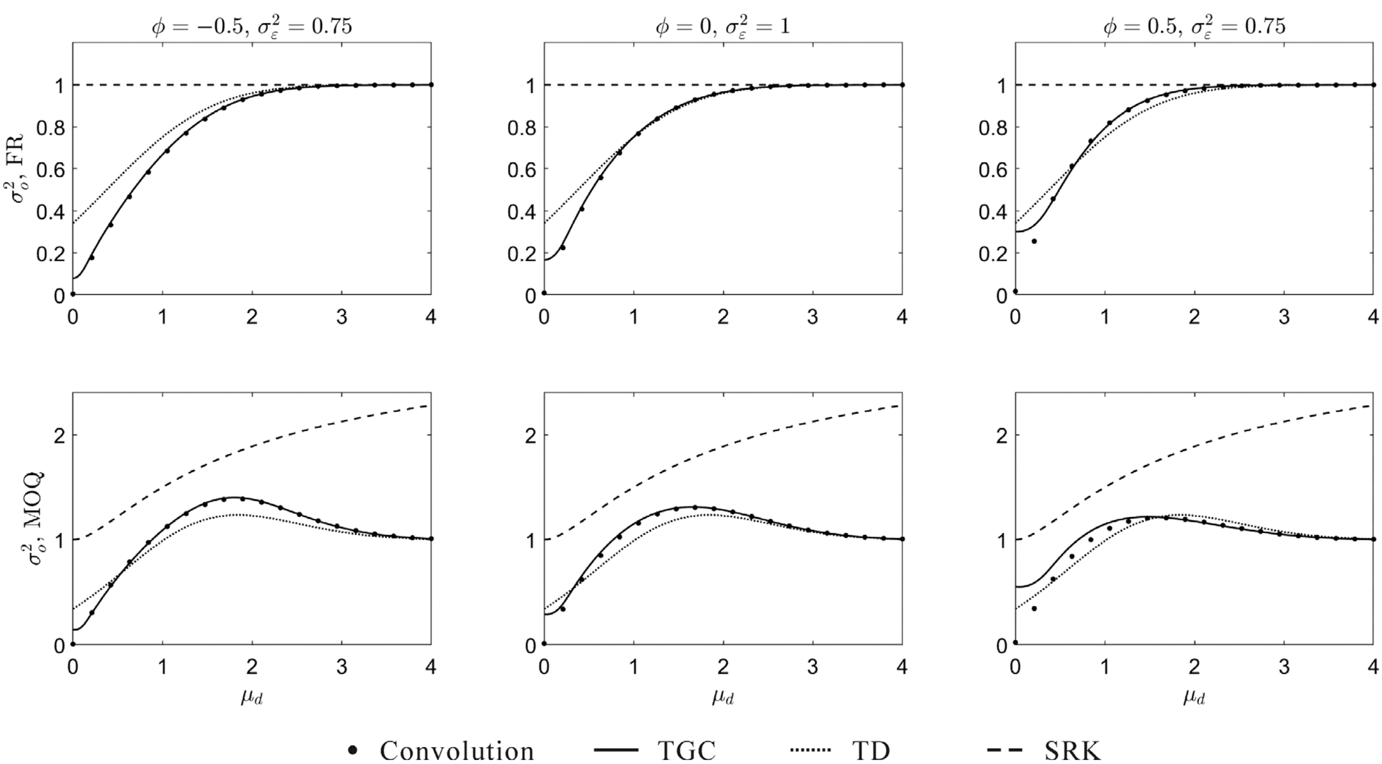

Fig. 7. The order variance with varying mean demand, demand correlation and order constraint.

is intended for the MOQ system only, while the FR system can be derived from the MOQ system by letting $s = S$. Therefore in the FR system, SRK approximates the order variance as the demand variance.

6.1. Bullwhip analysis

The bullwhip ratio is a good indicator of the impact of the order-constraint, defined as the ratio between order variance and demand variance. In the linear base-stock policy under i.i.d. demand with MMSE forecasts, $o_t = d_t$, the order variance equals the demand variance and the bullwhip ratio is one (Disney, Maltz, Wang, & Warburton, 2016). However, the bullwhip ratio in this nonlinear environment is no longer constant. Fig. 7 shows the evolution of this metric when the mean demand changes from 0 to 4 under auto-correlated demand and the FR and MOQ order constraints respectively. The auto-correlation model is chosen to be the first order auto-regressive, AR(1), demand process where Φ and Θ in (15) are both scalars such that $\Phi = \phi$ and $\Theta = 1$. The demand process can be written as $y_t = \phi y_{t-1} + \varepsilon_t$ and $d_t = y_t + \mu_d$. We illustrate scenarios of $\phi = -0.5$ (negative correlation); $\phi = 0$ (no correlation) and $\phi = 0.5$ (positive correlation). We also adjust the value of σ_ε^2 in three scenarios to make the demand variance constant at $\sigma_d^2 = \sigma_\varepsilon^2/(1-\phi^2) = 1$. In the MOQ system, we use $s = 0$ and $S = 1$.

First, in a nonlinear system with constrained orders, the mean demand (or more generally, the tightness of the constraint) has a significant impact on the order variance. This clearly contrasts with our knowledge in the linear system, where the mean demand does not affect the order variance. We see the real order variance in the FR system is increasing concave in the mean demand, and asymptotes to the demand variance when μ_d is sufficiently large. Similarly, the order variance is increasing concave in the capacity under the CC system. In the MOQ system, the order variance will first increase, and then decrease, in μ_d. This is because of the distance between the discrete and the continuous components (at $C_2 = 0$

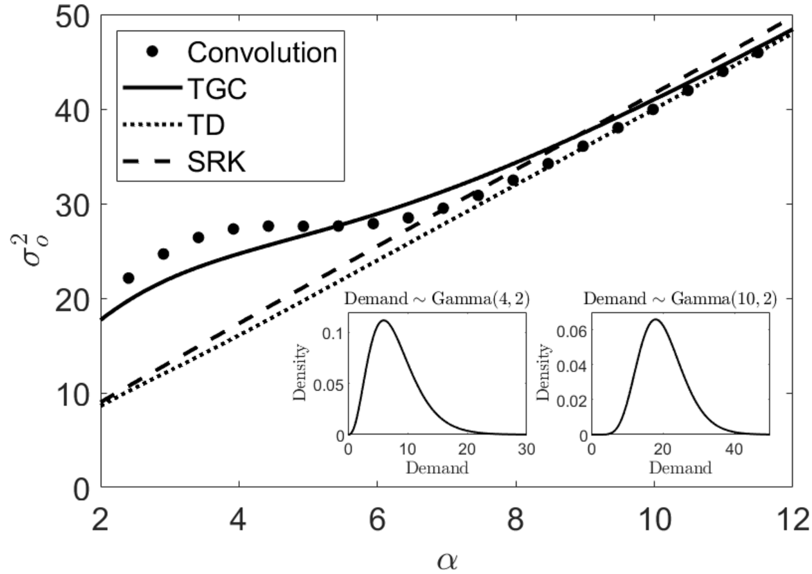

Fig. 8. The order variance in an (s, S) system with Gamma demand.

and $\mathcal{A}$ respectively) in the order distribution, such that a reduction in the probability of the discrete component does not always lead to an increase in order variance.

Second, we observe that TGC is a better approximation than TD. The advantages are two fold. From an accuracy perspective, even for i.i.d. demand, TGC is more accurate than TD when the mean demand is low. This is because TD does not account for the component mixture in the order distribution. Furthermore, as TD does not contain any information about the demand correlation, it cannot reflect the demand correlation properly. As shown in Fig. 7, in scenarios where the demand variance is equal but correlation is different, TD always gives the same approximation. However, the TGC approximation can account for the demand correlation. When compared with SRK in the MOQ system, we see that TGC significantly outperforms SRK at approximating the real order variance, and SRK does not take into account the demand auto-correlation. We need to note that SRK performs better when the MOQ becomes larger. This is because a key step in the derivation of the SRK approximation requires large MOQ values (see Appendix 1 of Schneider et al., 1995). On the other hand, the proposed TGC approximation performs less satisfactorily under large MOQ values as the component distributions deviate from Gaussian significantly.

The robustness of TGC is examined with Gamma demand in Fig. 8. The Gamma(α, β) distribution is given by $f_d(x) = \beta^{-\alpha} x^{\alpha-1} e^{-x/\beta} / \Gamma(\alpha)$ where α is the shape parameter, β is the scale parameter and $\Gamma(\cdot)$ is the Gamma function. This gives $\mu_d = \alpha\beta$ and $\sigma_d^2 = \alpha\beta^2$. We vary α between 2 and 12 with $\beta = 2$. As the support of the Gamma distribution is nonnegative, the FR constraint does not affect the order distribution. Therefore we use an MOQ system with $s = 0$ and $S = 1$. The result shows that TGC approximates the order variance better than TD and SRK when α is small, in terms of the distance from the real values. When α becomes large, the Gamma distribution becomes more like a Gaussian distribution (see the subplots in Fig. 8), and the accuracy of TGC improves. Meanwhile, TD converges faster than TGC in this example. This is because μ_d increases with α, and the order constraint (in this case $o_t > 1$) gradually becomes loose. SRK performs similar to TD but shows an increasing divergence when α increases.

The efficiency of TGC compared with the numerical convolution method and Monte-Carlo simulation is shown in Table 4. The task is to calculate the order variance in the FR system when demand is i.i.d. with $\mu_d = 1.5$ and $\sigma_d = 1$. We do not vary the parameters as they do not affect the computation time. The platform is a personal computer with INTEL® i7-8650 CPU at 1.9 GHz and 16 GB RAM. For the benchmark variance value, we use numerical convolution as the real value requires infinite convolutions which is practically impossible. To reduce the discretization error, a fine grid (10^5) is used to discretize the density functions. The number of iteration (30) is chosen such that the absolute change in the converging value is less than the maximum rounding error under double precision ($2^{-52} \approx 2.22 \times 10^{-16}$). Next, the approximation of TGC is given with an discrepancy of 10^{-3} from the benckmark value. For the methods where iteration is needed (simulation and numerical convolution), we show the computation time needed in order to achieve the same level of accuracy. The results by TD and SRK are also included. It is quite obvious that TGC possesses great superiority in terms of computational efficiency with the same level of accuracy between numerical convolution and simulation methods. On the other hand, TD and SRK are quick since no iteration is involved, but their approximations have large deviations from the benchmark. Finally, it is important to note the reliability of TGC decreases as the constraint becomes tighter, since the inadmissible part of the demand distribution is not negligible, and the component distributions cannot be safely approximated by Gaussian distributions.

6.2. Order-up-to level and service level

The achieved service level is another important performance measure of the inventory system, which can be adjusted by altering the safety stock and the order-up-to level. Since the order-up-to level horizontally shifts the inventory distribution and does not change its shape, it should be set such that the achieved service level (the probability that the inventory level is positive) equals the target service level. If the achieved service level does not equal the target, then the business will either be over- or under-stocked and the inventory cost will not be minimized. In this sense, the distance between achieved and target service level can be used to measure inventory control performance. Since the exact inventory distribution is asymptotically available, we can always achieve the target service level, at least numerically. In this section, we are interested in the service level performance using only the inventory variance estimate.

Table 4
Computational efficiency of the numerical methods.

	Benchmark*	TGC†	Simulation‡	Convolution§	TD	SRK
Order Variance, σ^2	0.8959	0.8969	(0.8950, 0.8968)	0.8950	0.8884	1.0000
Time (seconds)	43.66	0.17	5.78	5.52	0.01	0.003

* 10^5 data points are used to discretize the pdfs up until $o[30]$.
† $n = 100$.
‡ 95% confidence interval of 6400 samples over 1000 periods.
§ 10^5 data points are used to discretize the pdfs up until $o[5]$.

Table 5
Actual service level when safety stock is calculated by LTD and TGC. (Note: Bold font indicates cases where the TGC approximation does not perform worse than the LTD approximation).

		$\mu_d = 1$			$\mu_d = 2$		
		$\phi = -0.5$	$\phi = 0$	$\phi = 0.5$	$\phi = -0.5$	$\phi = 0$	$\phi = 0.5$
LTD	$L = 1$	0.900	0.898	0.895	0.894	0.865	0.820
	$L = 5$	0.900	0.899	0.898	0.891	0.887	0.873
	$L = 10$	0.900	0.900	0.900	0.892	0.892	0.884
TGC	$L = 1$	0.901	**0.898**	**0.905**	0.891	**0.879**	**0.895**
	$L = 5$	**0.900**	**0.899**	**0.901**	**0.891**	**0.890**	**0.889**
	$L = 10$	**0.900**	**0.900**	0.901	**0.893**	**0.892**	**0.892**

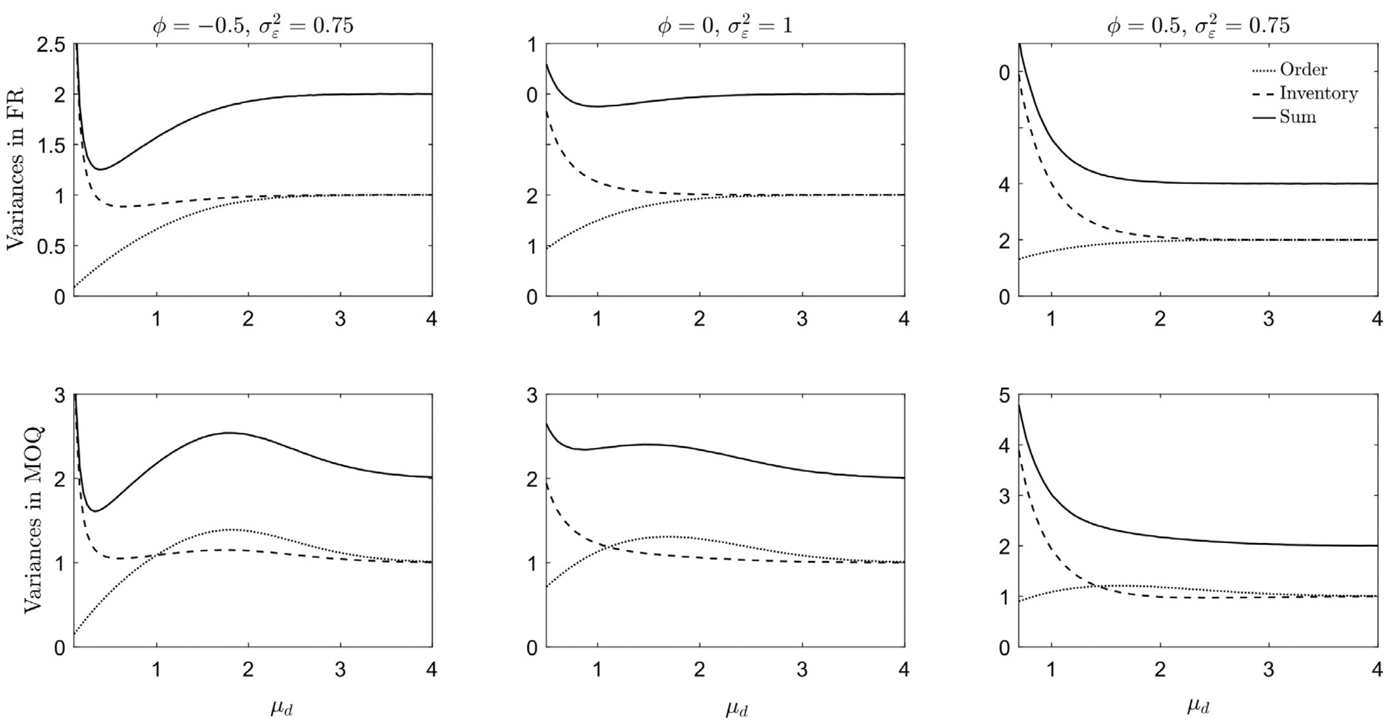

Fig. 9. Sum of order and inventory variance with varying mean demand, demand correlation, and order constraint.

Table 5 examines the performance of TGC in terms of service level in a capacitated system. The same three auto-correlated demand scenarios as in Section 6.1 are considered. The capacity $C_p = 3$, the lead-time values are $L = \{1, 5, 10\}$, and the mean demand $\mu_d = \{1, 2\}$. The order-up-to level is determined in the following way: $S = L\mu_d + z(\alpha_T)\sigma_i$, where first term is the mean of lead-time-demand, $z(\alpha_T)$ is the α_T-quantile of the standard Gaussian distribution. The second term $z(\alpha_T)\sigma_i$ is the safety stock. Under TGC, σ_i is derived by taking into account the truncations and mixtures. As a comparison, under the LTD approximation, the mean and variance of inventory are simply taken as those of the lead-time demand in the linear unconstrained case (the demand auto-correlation is still considered). A target service level of $\alpha_T = 90\%$ is assumed. The achieved service level is derived via simulation in the CC system.

By varying the lead-time and the demand autocorrelation, we see both approximations produce less satisfactory results when the constraint becomes tighter (that is, when μ_d becomes greater). This is true for both the LTD and the TGC approximations. There is a positive relationship between μ_d and the achieved service level; that is, given a capacity constraint, the achieved service level decreases as μ_d increases. This contrasts with a linear unconstrained inventory system model, in which the actual service level is affected by the order-up-to level but not the mean demand. The reason is the inventory distribution becomes more negatively skewed as μ_d becomes greater, leading to a thicker negative tail and lower achieved service level. The discrepancy between the target and the achieved service level also decreases in lead-time and decreases in demand correlation. A possible explanation is that long lead-time and negative correlation both lead to more "normal" component distributions, due to the central limit theorem and the risk pooling effect. We also observe TGC generally outperforms the LTD approximation in terms of reaching the target service level; the benefit could be as high as 7.5% (when $\mu_d = 2$, $\phi = 0.5$ and $L = 1$). This indicates significant cost savings can be gained with a more accurate inventory distribution estimate in the order-constrained system.

6.3. Sum of order and inventory variances

It is known that the standard deviation of inventory has a strong impact on inventory-related costs. Under certain assumptions (such as piece-wise linear and convex costs and Gaussian inventory distribution), this relationship is even linear. Hence, the study of inventory variance provides insight for inventory cost management. Similarly, the standard deviation of the replenishment orders (or production targets in manufacturing settings) directly creates capacity costs and also contributes to the inventory cost in upstream suppliers. The relative magnitudes of the order and inventory standard deviations alter the economic consequences (Boute et al., 2021). However, in order to preserve the analytical simplicity of quadratic problems, here we consider the simple sum of the variances.

Ponte et al. (2017) has observed that a capacity constraint on the order quantity has a smoothing effect on orders, reducing the order variance while increasing inventory variance. We can now explain this observation. When there is a constant constraint on the order quantity, there are two effects taking place. The first one is the *mixture effect*, where the order distribution becomes mixed, with component distributions that have different means. This increases the order variance. There is also the *truncation effect*, where all component distributions are truncated by the constraint, which tends to decrease the order variance. For the order quantity, the truncation effect is dominant, and the order variance decreases as the order constraint becomes tighter. However, the component distributions of inventory are not truncated as there is no constraint on the inventory level. This means the inventory variance increases with a tighter constraint (although not monotonically) as only the mixture effect is present. We also observe that as the constraint tightens, the order variance reduces slower than the inventory variance increases. However, the inventory variance grows significantly when the constraint becomes very tight. We can take advantage of this phenomenon and reduce order variance by adjusting the constraint, while keeping the inventory variance to an acceptable level, to balance the trade-off between inventory and order costs.

Fig. 9 shows the order variance, the inventory variance, and their sum. The mean demand, demand correlation and order constraints are the same as in Fig. 7. We let the lead-time $L = 1$ so that the order and inventory variances are comparable. The immediate observation is that for the FR system, when the demand is independent or negatively correlated, decreasing μ_d leads to a reduction in the variance of the sum, due to the reasons elaborated above. Moreover, by decreasing μ_d, a larger benefit can be achieved in the negative correlation case. As the auto-regression coefficient increases, the benefit of reducing μ_d becomes insignificant. The same effect can be observed in the CC system when the capacity is tuned. However, when $C_1 \neq C_2$ as in the MOQ system, the order variance does not increase with the mean demand, but increases then decreases. Consequently, the sum of variances may not have a global minimum, but instead possess multiple local maximums and minimums, as in the negative correlation and no correlation cases.

We can easily infer what would happen when the objective is a weighted sum of order and inventory variances (that is, $\gamma\sigma_o^2 + (1-\gamma)\sigma_i^2$ where $0 \leqslant \gamma \leqslant 1$). The weight γ does not change the tightness of the order constraint, therefore it does not affect the accuracy of the TGC algorithm. However, since the order variance decreases, and the inventory variance increases when the constraint tightens, the weight does affect how much the weighted sum can be reduced by the constraint. For instance, by increasing the weight for the order variance, there would be a greater reduction in the weighted sum of variances.

7. Conclusions

We have investigated the order and inventory distributions in a periodic review, continuous state, nonlinear inventory system, where the order quantity is (either upper- or lower-) constrained by a constant. This model is shown to have a wide application in practice, covering cases of forbidden returns, minimum order quantity, and production capacity. Our analytical framework is compatible with long transportation delays and demand autocorrelation.

Both the order and inventory distributions are a mixture of distributions, where the components can be derived by iterative convolutions of full and truncated distributions. For the Gaussian demand, we have proposed an algorithm that is able to approximate the density function and the moments, taking advantage of the analytical tractability of the truncated Gaussian distribution. The performance of the algorithm decreases with the tightness of the constraint, but it generally provides more accurate estimates than the intuitive approximations and is more efficient than numerical convolution. When applying this algorithm to set the safety stock, the accuracy leads to improved service level.

We have revealed the order constraint has a smoothing effect on orders but a variance amplifying effect on inventory. This is due to both distribution truncation and mixture effects. Combined, the order constraint is able to reduce the sum of order and inventory variance. This finding exposes a mechanism by which the order-inventory trade-off can be influenced by the order constraint. Strategically constraining orders has the potential to significantly enhance the dynamic behaviour of production and distribution systems.

Managerially, our analysis provides additional insights for managers to understand the effect of the order constraint on the fluctuations of order quantity and inventory level. The approximation algorithm allows for accurate and efficient estimations of the magnitude of this effect. Moreover, it enlarges the toolbox for controlling the trade-off between order and inventory variances. We can now choose to influence the mean demand, introduce a minimum order quantity, or adjust the capacity. Such measures may be easier to implement than the proportional control method (Disney & Towill, 2003) usually advocated to control the bullwhip effect, as the core algorithm does not need to be changed. The ordering decision can be made by simply comparing the originally recommended quantity and the constraint value. In future research, it would be interesting to use our modelling approach to extract demand information from observed constrained order information. Furthermore, the effect of demand auto-correlation and lead-time on the inventory level performance in the order-constrained systems is worth further exploration, especially when the order-up-to level is dynamic over time due to the demand forecast.

Acknowledgements

We would like to thank the anonymous referees for their detailed and constructive reports that led to many improvements of this paper. Also, Borja Ponte's work in this article was financially supported by the Spanish State Research Agency (Agencia Estatal de Investigación) of the Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033), via the project 'Speeding Up the transition towards Resilient circular economy networks: forecasting, inventory and production control, reverse logistics and supply chain dynamics" (SPUR, grant ref. PID2020-117021GB-I00).

References

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics, 12*(2), 171–178.

Boute, R., Disney, S. M., Gijsbrechts, J., & Van Mieghem, J. A. (2021). Dual sourcing and smoothing under non-stationary demand time series: Re-shoring with SpeedFactories. *Management Science, 68*(2), 1039–1057. https://doi.org/10.1287/mnsc.2020.3951.

Chatfield, D. C., & Pritchard, A. M. (2013). Returns and the bullwhip effect. *Transportation Research Part E: Logistics and Transportation Review, 49*(1), 159–175.

Chen, L., & Lee, H. L. (2009). Information sharing and order variability control under a generalized demand model. *Management Science, 55*(5), 781–797.

Dahel, N. (2003). Vendor selection and order quantity allocation in volume discount environments. *Supply Chain Management: An International Journal, 8*(4), 335–342.

Disney, S., & Towill, D. (2003). On the bullwhip and inventory variance produced by an ordering policy. *Omega, 31*(3), 157–167.

Disney, S. M., Gaalman, G. J. C., Hedenstierna, C. P. T., & Hosoda, T. (2015). Fill rate in a periodic review order-up-to policy under auto-correlated normally distributed, possibly negative, demand. *International Journal of Production Economics, 170*(B), 501–512.

Disney, S. M., Maltz, A., Wang, X., & Warburton, R. D. H. (2016). Inventory management for stochastic lead times with order crossovers. *European Journal of Operational Research, 248*(2), 473–486.

Disney, S. M., Ponte, B., & Wang, X. (2021). Exploring the nonlinear dynamics of the lost-sales order-up-to policy. *International Journal of Production Research, 59*, 5809–5830.

Federgruen, A., & Zipkin, P. (1986a). An inventory model with limited production capacity and uncertain demands I: The average cost criterion. *Mathematics of Operations Research, 11*, 193–207.

Federgruen, A., & Zipkin, P. (1986b). An inventory model with limited production capacity and uncertain demands II: The discounted cost criterion. *Mathematics of Operations Research, 11*, 208–215.

Gavirneni, S., Kapuściński, R., & Tayur, S. (1999). Value of information in capacitated supply chains. *Management Science, 45*(1), 16–24.

Glasserman, P. (1997). Bounds and asymptotics for planning critical safety stocks. *Operations Research, 45*, 244–257.

Glasserman, P., & Tayur, S. (1994). The stability of a capacitated, multiechelon production-inventory system under a base-stock policy. *Operations Research, 42*, 913–925.

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

Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.

Henze, N. (1986). A probabilistic representation of the 'skew-normal' distribution. *Scandinavian Journal of Statistics, 13*(4), 271–275.

Kapuściński, R., & Tayur, S. (1998). A capacitated production-inventory model with periodic demand. *Operations Research, 46*(6), 899–911.

Karlin, S. (1958). The application of renewal theory to the study of inventory policies. In K. Arrow, S. Karlin, & H. Scarf (Eds.), *Studies in the mathematical theory of inventory and production*. Stanford, California: Stanford University Press.

Kelle, P., & Milne, A. (1999). The effect of (s, S) ordering policy on the supply chain. *International Journal of Production Economics, 59*(1), 113–122.

Larsen, E. R., Morecroft, J. D. W., & Thomsen, J. S. (1999). Complex behaviour in a production-distribution model. *European Journal of Operational Research, 119*(1), 61–74.

Laugesen, J., & Mosekilde, E. (2006). Border-collision bifurcations in a dynamic management game. *Computers and Operations Research, 33*(2), 464–478.

Lee, H., Padmanabhan, V., & Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science, 43*, 546–558.

Levi, R., Roundy, R. O., Shmoys, D. B., & Truong, V. A. (2008). Approximation algorithms for capacitated stochastic inventory control models. *Operations Research, 56*(5), 1184–1199.

Noblesse, A. M., Boute, R. N., Lambrecht, M. R., & Houdt, B. V. (2014). Characterizing order processes of continuous review (s, S) and (r, nQ) policies. *European Journal of Operational Research, 236*(2), 534–547.

Nurminen, H., Rui, R., Ardeshiri, T., Bazanella, A., & Gustafsson, F. (2016). Mean and covariance matrix of a multivariate normal distribution with one doubly-truncated component. *Technical Report*. Linköpings Universitet.

Parker, R. P., & Kapuściński, R. (2004). Optimal policies for a capacitated two-echelon inventory system. *Operations Research, 52*(5), 739–755.

Ponte, B., Wang, X., de la Fuente, D., & Disney, S. M. (2017). Exploring nonlinear supply chains: The dynamics of capacity constraints. *International Journal of Production Research, 55*(14), 4053–4067.

Sato, K.-I. (1999). *Lévy processes and infinitely divisible distributions*. Cambridge University Press.

Scarf, H. (1960). The optimality of (S, s) policies in the dynamic inventory problem. In K. J. Arrow, S. Karlin, & P. Suppes (Eds.), *Mathematical methods in the social sciences*. Stanford University Press.

Scarf, H. (1963). A survey of analytic techniques in inventory theory. In H. E. Scarf, D. D. Gilford, & M. W. Shelley (Eds.), *Multistage inventory models and techniques*. Stanford University Press.

Schneider, H., Rinks, D. B., & Kelle, P. (1995). Power approximations for a two-echelon inventory system using service levels. *Production and Operations Management, 4*(4), 381–400.

Schultz, C. R. (1983). Computing demand properties at the wholesale warehouse level. *Naval Research Logistics Quarterly, 30*(1), 37–48.

Sterman, J. D. (1989). Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science, 35*(3), 321–339.

Tayur, S. (1993). Computing the optimal policy for capacitated inventory models. *Communications in Statistics: Stochastic Models, 9*, 585–598.

Wang, X., & Disney, S. M. (2017). Mitigating variance amplification under stochastic lead-time: The proportional control approach. *European Journal of Operational Research, 256*(1), 151–162.

Wang, X., Disney, S. M., & Wang, J. (2012). Stability analysis of constrained inventory systems with transportation delay. *European Journal of Operational Research, 223*(1), 86–95.

Zheng, Y. (1991). A simple proof for optimality of (s, S) policies in infinite-horizon inventory systems. *Journal of Applied Probability, 28*(4), 802–810.

Zhou, B., Zhao, Y., & Katehakis, M. N. (2007). Effective control policies for stochastic inventory systems with a minimum order quantity and linear costs. *International Journal of Production Economics, 106*(2), 523–531.