LARGE DEVIATIONS-BASED ASYMPTOTICS FOR INVENTORY CONTROL IN SUPPLY CHAINS

IOANNIS CH. PASCHALIDIS and YONG LIU

Center for Information and Systems Engineering (CISE) and Department of Manufacturing Engineering, Boston University, Boston, Massachusetts 02215
yannisp@bu.edu • liuyong@bu.edu

We consider a model of a capacitated single-class supply chain consisting of production facilities (stages) in tandem. External demand is met from the available finished goods inventory maintained in front of the most downstream facility; unsatisfied demand is backlogged. Every stage orders from its upstream facility, thus production is constrained by the local production capacity and the availability of upstream inventory. We propose production policies in two separate cases: (1) when each facility has information about its local inventory only, and (2) when each facility has knowledge of the total downstream inventory. In case (1) the proposed policy guarantees that stockout probabilities at each stage stay bounded below given constants (service level constraints). In case (2) the proposed policy minimizes total expected inventory cost subject to desirable service-level constraints. In both cases the parameters of the proposed policies are obtained analytically based on large deviations asymptotics, which leads to drastic computational savings compared to simulation. An important feature of our model is that it accommodates autocorrelated demand and service processes, both critical features of modern failure-prone manufacturing systems. We demonstrate that detailed distributional information on demand and service processes, which is incorporated into large deviations asymptotics, is critical in inventory control decisions. We discuss extensions to a multiclass setting and to a model where unsatisfied demand is lost instead of backordered.

Received April 2000; revisions received May 2001, February 2002; accepted May 2002.
Subject classifications: Inventory/production: service-level approximations/heuristics. Inventory/production, uncertainty: correlations in demand and capacity processes. Probability, applications: large deviations.
Area of review: Stochastic Models.

1. INTRODUCTION

Manufacturing has recently gone through significant restructuring. A recent survey (*The Economist* 1998) emphasized that "no factory is an island." Companies are becoming more global. They consist of factories, suppliers, distributors, and customer service centers scattered around the globe. As a result, modern manufacturing enterprises have recognized that production cannot be viewed separately from the physical distribution of goods. Instead, both activities should be perceived as indispensable parts of a *supply chain.*

Manufacturing is also becoming more *customer oriented.* In an era of increased competition, customers are more demanding and require products delivered in a timely manner wherever they happen to be located. In addition to product functionality, companies are recognizing the significance of *quality of service* (QoS) in acquiring and maintaining market share. The increasing reliance on information technology and the emergence of e-commerce have increased the importance of effectively managing the supply chain and at the same time are providing more tools to that end.

Our primary objective is to develop effective policies for inventory control in supply chains that address the difficulties present in the new manufacturing environment. The fundamental trade-off in inventory control is between *producing,* which accumulates inventory and incurs inventory costs, and *idling,* which leads to stockouts and unsatisfied demand. A *production policy* resolves this trade-off and determines at each point in time whether the production facilities at all stages of the supply chain should be producing or idling.

There is a large literature on production inventory systems (see Kapuściński and Tayur 1999 for a survey). The single-stage, single-class version of the problem is significantly simpler. It has been shown in a variety of settings (all without set-up costs) that a so-called *base-stock* policy (i.e., produce when inventory falls below a prescribed level and idle otherwise) is optimal (see Evans 1967, Gavish and Graves 1980, Sobel 1982, Federgruen and Zipkin 1986, Akella and Kumar 1986, and Kapuściński and Tayur 1998). In multiclass single-stage systems the optimal policy is not, in general, known. In these systems a production policy involves both idling and scheduling decisions (deciding on which classes to work on, if any). There have been results only for special cases (Zheng and Zipkin 1990, Ha 1997, Véricourt et al. 2000) or approximations and heuristics for the general case (Wein 1992, Peña-Perez and Zipkin 1997, Veatch and Wein 1996, Glasserman 1996, and Bertsimas and Paschalidis 2001). In a multiple-stage, single-class system, and *without capacity limits*, Clark and Scarf in their seminal paper (1960) have shown the optimality of a production policy where each facility follows a base-stock policy based on the total inventory available locally and in the downstream facilities (we will refer to this as *echelon* inventory). Their result has been generalized in several directions (Federgruen and Zipkin 1984, Chen and Song 2001). In the more general case, where capacity limits exist and demand and service processes are autocorrelated, such a policy is not necessarily optimal. However, the simplicity

0030-364X/03/5103-0437 $05.00
1526-5463 electronic ISSN

Operations Research © 2003 INFORMS
Vol. 51, No. 3, May–June 2003, pp. 437–460

of its structure makes it attractive. Under a similar echelon policy Glasserman and Tayur (1995) proposed a perturbation analysis approach to compute the hedging points in a capacitated single-class multistage system, and Glasserman (1997) has developed asymptotics to approximate stockout probabilities under renewal demand and constant production capacities.

In this paper we propose and analyze two base-stock production policies. Our first policy uses only local inventory information at each stage of the supply chain. Our second policy has similar structure with the policy proposed by Clark and Scarf (1960), that is, each stage makes decisions based on the local and total downstream inventory. In both cases, we introduce constraints that ensure that stockout probabilities stay bounded below given desirable levels. Such *service-level* constraints provide a more natural representation of customer satisfaction and are closely watched by manufacturing managers. This is in contrast to most of the work in the literature that considers policies minimizing expected inventory and backorder costs. Our analysis is general enough to accommodate temporal dependencies in demand and production processes. In practice, demand for various products might have strong correlations with a variety of phenomena, such as sales events, weather patterns, state of the economy, etc. Moreover, manufacturing facilities are *stochastic* and *failure-prone*, which creates dependencies in the production process. Under such assumptions, analyzing stockout probabilities exactly is intractable. We will instead rely upon *large deviations* techniques that lead to asymptotically tight approximations of stockout probabilities as they approach zero. As a result, we will be able to analytically obtain the appropriate base-stock levels for both policies we consider. Related techniques have been recently used by Bertsimas and Paschalidis (2001) to devise production policies in a multiclass, single-stage setting. Approximation techniques of this type, but in the simpler case of renewal demand and production processes, have been introduced by Glasserman (1996, 1997).

Among the main contributions of our work we consider:

Dependencies in the demand and service processes. As outlined above, this allows us to model more realistic demand situations and failure-prone production facilities. On the technical front, we have been able to use the full power of *large deviations* techniques to extend the asymptotics of Glasserman (1997) beyond the case of renewal demand and constant production capacities. Our results are "network" large deviations results (tandem queues in particular). Such results have only been obtained in limited network cases, as in Bertsimas et al. (1998b), which we use in the case of local inventory information. Our echelon inventory results take into account the strong coupling between different stages of the supply chain and, to the best of our knowledge, are the first of such "network" results to do so in the presence of stochastic and autocorrelated demand and production processes. Our echelon inventory main result has an interesting interpretation: It identifies a bottleneck stage whose production capacity is "responsible" for stockouts of finished goods. But this bottleneck stage *is not* necessarily the one with the smallest mean capacity; it is determined by more detailed *distributional* information on all stochastic processes involved. In this sense, such distributional information is critical in making inventory control decisions.

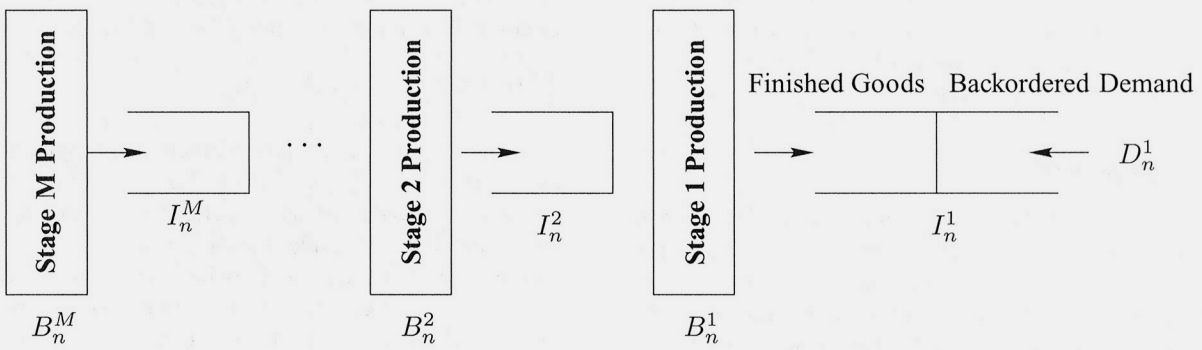
Relevance of the large deviations asymptotics. Our numerical results demonstrate that the large deviations asymptotics are fairly *accurate* in a wide range of desired stockout probabilities, including relatively large ones. Key to this are some heuristics we propose to compute a prefactor in front of the large deviations exponential. This prefactor can be more tightly characterized in the simpler case of renewal demand and constant production capacities (see Glasserman 1997).

The remainder of this paper is organized as follows. In §2, we provide the detailed model of the multistage, single-class production-inventory system, introduce the production policies we will consider, and outline our approach. In §3, we review some basic large deviations results we will be using in this paper. In §4, we analyze our first production policy, which is based on local inventory information. In §5, we treat our second policy, which uses echelon inventory information. We discuss extensions to the multiclass case and to a model of lost sales (instead of backorders) in §6. Numerical results that assess the accuracy of the proposed analytical approach are in §7 and conclusions in §8.

2. THE MODEL

Figure 1 depicts the supply-chain model we consider in this paper. This system produces a single product class and consists of M production facilities in tandem. We will say that production consists of M stages. External demand is met from the finished goods inventory maintained in front of the stage 1 production facility and is backordered if inventory is not available. Every production facility is fed by its upstream facility; in particular, to produce one unit facility i, $i = 1, \ldots, M-1$, requires one unit of the product of facility $i-1$. We assume that facility M is fed with an infinite supply of raw material. In front of every facility i, $i = 2, \ldots, M$, there is an inventory buffer that holds the final product of that facility and from which facility $i-1$ draws material for its production. We assume a periodic review policy in which time is divided into time slots of equal duration. For all $i = 1, \ldots, M$ and n we let B_n^i denote the amount that the facility at stage i can produce during time slot n (production capacity). We also let D_n^1 denote the amount of external orders arriving at stage 1 during time slot n. Finally, we let I_n^i, $i = 1, \ldots, M$, denote the inventory in front of stage i at the beginning of time slot n. In intermediate stages $i = 2, \ldots, M$ the inventory I_n^i is constrained to be nonnegative. In contrast, we allow the inventory at stage 1, I_n^1, to take negative values to denote backordering; when I_n^1 is negative $-I_n^1$ is equal to the amount of backordered demand.

Figure 1. The model of the supply chain.



The system evolves as follows. At the beginning of time slot $n+1$, the inventory at stage 1 is given by

$$I_{n+1}^1 = I_n^1 - D_n^1 + P_n^1,$$

where P_n^1 denotes the amount produced during time slot n by the facility at stage 1, which is determined by the production policy we select and confined by the production capacity B_n^1 and the available upstream inventory I_n^2. The quantity P_n^1 can be also viewed as the demand for stage 2, which operates in a similar manner and generates demand for stage 3. Thus, the whole supply chain is driven by the external demand.

The demand process $\{D_n^1;\ n \in \mathbb{Z}\}$ and the production processes $\{B_n^i;\ n \in \mathbb{Z}\}$, $i = 1, \ldots, M$, are mutually independent, possibly autocorrelated, arbitrary stationary stochastic processes that satisfy certain mild technical conditions (some form of a sample path large deviations principle). These conditions are satisfied by renewal processes, Markov-modulated processes, and in general stationary processes with mild mixing conditions (for details see Bertsimas et al. 1998a, b, and 1999). For stability purposes[1] we assume that

$$\mathbf{E}[D_n^1] < \min_{i=1,\ldots,M} \mathbf{E}[B_n^i], \tag{1}$$

which by stationarity carries over to all time slots n. Stability can be shown under both base-stock policies we will consider in this paper by using techniques from Baccelli and Liu (1992). For the case of an echelon base-stock policy a stability proof is given in Glasserman and Tayur (1994).

Our objective is to find a policy within a selected class of production policies that minimizes expected inventory costs and guarantees that the steady-state stockout probability $\mathbf{P}[I_n^1 \leqslant 0]$, at some arbitrary time slot n, does not exceed a desirable small value ϵ. We will be referring to this as a *service-level* constraint. In this paper, we will propose policies in two separate cases: (1) when each stage i has knowledge of its local inventory I_n^i only, and (2) when each facility i has knowledge of the total downstream inventory $I_n^i + I_n^{i-1} + \cdots + I_n^1$. In both cases, we will implement a base-stock policy.

In particular, in case (1) every stage i sets a *hedging point* or *safety stock* w_i for its local inventory I_n^i and implements the production policy: produce if I_n^i falls below w_i, and idle otherwise. In the single-stage ($M = 1$) case, this policy has been analyzed in Bertsimas and Paschalidis (2001). In a multistage system, however, there is strong coupling between stages because upstream inventory can constrain downstream production, which makes exact analysis particularly hard. To bypass this problem, we will use a *decomposition approach*. More specifically, we will focus in a regime where coupling between stages becomes weaker. For every stage this is the case if the safety stock in the upstream buffer is large enough, so that downstream production is rarely constrained by upstream inventory availability. In effect, each stage can be viewed as an independent single-stage system, and the results in Bertsimas and Paschalidis (2001) can be applied. To that end, though, we need to characterize the demand for every stage i by "propagating" the external demand through the downstream stages $1, 2, \ldots, i-1$.

The policy obtained via the decomposition approach, although it maintains the service-level constraint at stage 1, might not necessarily be efficient in terms of expected inventory cost. Information of inventory availability in other stages might lead to lower such cost by giving the opportunity to trade off inventory between different stages, i.e., lower the required safety stock in stages where inventory costs are high and compensate by increasing the safety stock in stages where costs are lower. Case (2) considers such a situation. Let $X_n^i = I_n^i + I_n^{i-1} + \cdots + I_n^1$, $i = 1, \ldots, M$, denote the total downstream inventory from stage i; we will be referring to this quantity as *echelon* inventory at stage i. Consider the policy according to which every stage i sets a hedging point or safety stock w_i for X_n^i and produces if X_n^i falls below w_i and idles otherwise. We will analyze this policy and devise a production policy that minimizes expected inventory costs subject to given service-level constraints.

In both cases (1) and (2) we need to obtain the stockout probability to be able to maintain the service-level constraints. An exact expression is intractable, especially in view of the rather complicated (autocorrelated) models for the demand and production processes. To that end, we will employ *large deviations* theory. In the regime of small

stockout probabilities (or equivalently, large safety stocks) stockouts are rare events and are amenable to large deviations analysis. We will provide numerical results to demonstrate that the large deviations asymptotics are *accurate* when compared to simulations.

3. PRELIMINARIES

Before we proceed with our agenda and in the form of background on large deviations we first review some basic results, which will also help in establishing some of our notation. Consider a sequence of i.i.d. random variables X_i, $i \geqslant 1$, with mean $\mathbf{E}[X_1] = \bar{X}$. The strong law of large numbers asserts that $\sum_{i=1}^{n} X_i/n$ converges to $\bar{X}$, as $n \to \infty$, with probability 1 (w.p.1). Thus, for large n the event $\sum_{i=1}^{n} X_i > na$, where $a > \bar{X}$, (or $\sum_{i=1}^{n} X_i < na$, for $a < \bar{X}$) is a rare event. More specifically, its probability behaves as $e^{-nr(a)}$, as $n \to \infty$, where the function $r(\cdot)$ determines the rate at which the probability of this event is diminishing. Cramér's (1938) theorem determines $r(\cdot)$, and is considered the first large deviations statement. In particular,

$$r(a) = \sup_{\theta} \left(\theta a - \log \mathbf{E}[e^{\theta X_1}]\right).$$

Next, consider a sequence $\{S_1, S_2, \dots\}$ of random variables, with values in $\mathbb{R}$ and define

$$\Lambda_n(\theta) \triangleq \frac{1}{n} \log \mathbf{E}[e^{\theta S_n}]. \tag{2}$$

For the applications we have in mind, S_n is a partial sum process. Namely, $S_n = \sum_{i=1}^{n} X_i$, where X_i, $i \geqslant 1$, are identically distributed, possible dependent, random variables. Let

$$\Lambda(\theta) \triangleq \lim_{n\to\infty} \Lambda_n(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathbf{E}[e^{\theta S_n}]. \tag{3}$$

(We assume that the limit exists for all θ, where $\pm\infty$ are allowed both as elements of the sequence $\Lambda_n(\theta)$ and as limit points.) We will refer to $\Lambda(\cdot)$ as the *limiting* log-*moment generating function*. Let us also define

$$\Lambda^*(a) \triangleq \sup_{\theta}(\theta a - \Lambda(\theta)), \tag{4}$$

which will be referred to as the *large deviation rate function*. Under a technical assumption (see Dembo and Zeitouni 1998) it has been shown (Gärtner-Ellis Theorem) that for large enough n and for small $\epsilon > 0$,

$$\mathbf{P}[S_n \in (na - n\epsilon, na + n\epsilon)] \sim e^{-n\Lambda^*(a)}. \tag{5}$$

This can be viewed as an extension of Cramér's theorem to autocorrelated stochastic processes. When (5) holds, we say that $\{S_n\}$ satisfies a *large deviations principle* (LDP) with *rate function* $\Lambda^*(\cdot)$. The notation "$\sim$" should be interpreted as "asymptotically behaves"; more rigorously, the logarithm of the probability divided by n converges to $-\Lambda^*(a)$, as $n \to \infty$.

It is important to note that $\Lambda(\cdot)$ and $\Lambda^*(\cdot)$ are convex duals (Legendre transforms of each other) (Dembo and Zeitouni 1998). Namely, along with (4), it also holds that

$$\Lambda(\theta) = \sup_{a}(\theta a - \Lambda^*(a)).$$

In this paper, we are also estimating the tail probabilities of the form $\mathbf{P}[S_n \leqslant na]$ or $\mathbf{P}[S_n \geqslant na]$. We therefore define large deviations rate functions associated with such tail probabilities. Consider the case where $S_n = \sum_{i=1}^{n} X_i$, the random variables X_i, $i \geqslant 1$, being identically distributed, and let $m = \mathbf{E}[X_1]$. It can be shown (see Dembo and Zeitouni 1998) that $\Lambda^*(m) = 0$. Let us now define

$$\Lambda^{*+}(a) \triangleq \begin{cases} \Lambda^*(a) & \text{if } a > m, \\ 0 & \text{if } a \leqslant m, \end{cases}$$

and

$$\Lambda^{*-}(a) \triangleq \begin{cases} \Lambda^*(a) & \text{if } a < m, \\ 0 & \text{if } a \geqslant m. \end{cases} \tag{6}$$

Notice that $\Lambda^{*+}(a)$ is nondecreasing and $\Lambda^{*-}(a)$ is nonincreasing function of a, respectively. The convex duals of these functions are

$$\Lambda^{+}(\theta) \triangleq \begin{cases} \Lambda(\theta) & \text{if } \theta \geqslant 0, \\ +\infty & \text{if } \theta < 0, \end{cases}$$

and

$$\Lambda^{-}(\theta) \triangleq \begin{cases} \Lambda(\theta) & \text{if } \theta \leqslant 0, \\ +\infty & \text{if } \theta > 0, \end{cases} \tag{7}$$

respectively. In particular, $\Lambda^{*-}(a) = \sup_{\theta}(\theta a - \Lambda^{-}(\theta))$ and $\Lambda^{*+}(a) = \sup_{\theta}(\theta a - \Lambda^{+}(\theta))$.

Using the Gärtner-Ellis Theorem it can be shown (Bertsimas et al. 1998b) that for all $\epsilon_1, \epsilon_2 > 0$ there exists n_0 such that for all $n \geqslant n_0$

$$e^{-n(\Lambda^{*-}(a)+\epsilon_2)} \leqslant \mathbf{P}[S_n \leqslant na] \leqslant e^{-n(\Lambda^{*-}(a)-\epsilon_1)}, \tag{8}$$

$$e^{-n(\Lambda^{*+}(a)+\epsilon_2)} \leqslant \mathbf{P}[S_n \geqslant na] \leqslant e^{-n(\Lambda^{*+}(a)-\epsilon_1)}. \tag{9}$$

On a notational remark, in the sequel we will be denoting by

$$S_{i,j}^{X} \triangleq \sum_{k=i}^{j} X_k, \quad i \leqslant j, \tag{10}$$

the partial sums of the random sequence $\{X_i;\ i \in \mathbb{Z}\}$ and make the convention $S_{i,j}^{X} = 0$, if $i > j$. We will be also denoting by $\Lambda_X(\cdot)$ and $\Lambda_X^*(\cdot)$ the limiting log-moment generating function and the large deviations rate function, respectively, of the process X.

4. THE DECOMPOSITION APPROACH—A LOCAL INFORMATION CASE

In this section we consider the case where each stage has knowledge of its local inventory only. We start by reviewing the single-stage problem; our analysis of the multistage problem will build on this.

4.1. Single-Stage Problem

In agreement with the notation introduced in §2 we will be using I_n, D_n, and B_n to denote the inventory, demand, and production capacity, respectively, during time slot n. The system we are dealing with is a make-to-stock system and has been analyzed (as a special case) in Bertsimas and Paschalidis (2001).

We implement a base-stock policy that maintains a safety stock of w in the inventory. The inventory evolves as follows:

$$I_{n+1} = \min\{I_n - D_n + B_n, w\}.$$

The objective is to keep the stockout probability less than some given threshold $\epsilon > 0$, i.e., $\mathbf{P}[I_n \leqslant 0] \leqslant \epsilon$. Define the shortfall L_n as the gap between the current inventory and the hedging point, i.e., $L_n \triangleq w - I_n$. In terms of L_n, the dynamics of the system can be written as

$$L_{n+1} = [L_n + D_n - B_n]^+ \triangleq \max\{L_n + D_n - B_n, 0\}, \qquad (11)$$

and we have the following equality:

$$\mathbf{P}[I_n \leqslant 0] = \mathbf{P}[L_n \geqslant w].$$

We can view L_n as the queue length of a discrete-time $G/G/1$ queue with arrival process $\{D_n; n \in \mathbb{Z}\}$ and service process $\{B_n; n \in \mathbb{Z}\}$. Hence, the problem is to characterize the steady-state overflow probability $\mathbf{P}[L_n \geqslant w]$. This has been done in Bertsimas and Paschalidis (2001) using large deviations techniques. On a notational remark, in the sequel we will be dropping the reference to the time slot (subscript n) when referring to steady-state quantities. For example, we will be denoting by L the steady-state queue length at an arbitrary time slot.

THEOREM 1 (SINGLE STAGE, BERTSIMAS AND PASCHALIDIS 2001). *For the single-stage system (cf. (11)), the steady-state queue length L satisfies*

$$\lim_{w\to\infty} \frac{1}{w} \log \mathbf{P}[L \geqslant w] = -\theta_s^*,$$

where $\theta_s^* > 0$ is the largest root of the equation $\Lambda_D^+(\theta) + \Lambda_B(-\theta) = 0$.

More intuitively, for large enough w we have

$$\mathbf{P}[I \leqslant 0] = \mathbf{P}[L \geqslant w] \sim e^{-w\theta_s^*},$$

thus the minimum w that satisfies $\mathbf{P}[I \leqslant 0] \leqslant \epsilon$ is

$$w = -\frac{\log \epsilon}{\theta_s^*}.$$

Note that at the origin $\Lambda_D^+(\theta) + \Lambda_B(-\theta)$ equals zero and its slope is $\mathbf{E}[D_1] - \mathbf{E}[B_1]$, which is negative by the stability condition (1). It is possible that $\Lambda_D^+(\theta) + \Lambda_B(-\theta)$ never crosses the horizontal axis for $\theta > 0$, in which case we will say $\theta_s^* = +\infty$ and a hedging point of zero should be used (just-in-time (JIT) policy).

4.2. Multiple Stages

We now return to our original problem with M stages. We propose a base-stock policy that maintains a safety stock equal to w_i for the (local) inventory of every stage i, $i = 1, \ldots, M$. In particular, stage i produces until the local inventory I_n^i reaches the hedging point w_i and idles if $I_n^i \geqslant w_i$. The amount produced by stage i constitutes demand for the upstream stage $i+1$, for $i = 1, \ldots, M-1$; we will denote it by D_n^{i+1}. Note that the demand for stage i, D_n^i, $i = 2, \ldots, M$, is constrained by the downstream capacity B_n^{i-1} and the available inventory I_n^i.

The dynamics for the supply chain are

$$I_{n+1}^i = \min\{I_n^i - D_n^i + B_n^i, I_n^i - D_n^i + I_n^{i+1}, w_i\}, \quad i = 1, \ldots, M-1,$$

$$I_{n+1}^M = \min\{I_n^M - D_n^M + B_n^M, w_M\}.$$

The demand for stage i (or, equivalently, production of stage $i-1$) is given by

$$D_n^i = I_{n+1}^{i-1} - I_n^{i-1} + D_n^{i-1}, \quad i = 2, \ldots, M.$$

As in the single-stage case, we define the inventory shortfall for stage i as $L_n^i \triangleq w_i - I_n^i$, $i = 1, \ldots, M$, and the dynamics of the supply chain can be written as

$$L_{n+1}^i = \max\{L_n^i + D_n^i - B_n^i, L_n^i + D_n^i + L_n^{i+1} - w_{i+1}, 0\}, \quad i = 1, \ldots, M-1,$$

$$L_{n+1}^M = \max\{L_n^M + D_n^M - B_n^M, 0\}.$$

The demand for stage i can now be expressed as

$$D_n^i = L_n^{i-1} - L_{n+1}^{i-1} + D_n^{i-1}, \quad i = 2, \ldots, M. \qquad (12)$$

The major difficulty for analyzing this model and characterizing the stockout probabilities is that the production is constrained not only by its own capacity, but also by the upstream inventory. To bypass this difficulty we will *decouple* the various stages by ignoring the upstream inventory constraint on the downstream production. More specifically, the proposed decomposition amounts to assuming that the system operates according to a policy which satisfies $I_n^{i+1} \geqslant B_n^i$, $i = 1, \ldots, M-1$, almost surely for all time slots n. We can intuitively argue that this decomposition is in fact accurate when the inventory level of the upstream stage is high enough; then the influence of the upstream inventory constraint will be insignificant when compared to the capacity constraint. We later argue (see also §7) that we can approximate this behavior by enforcing a small enough stockout probability for the upstream stage. In the decoupled system, the dynamics of the supply chain can be simplified as follows:

$$I_{n+1}^i = \min\{I_n^i - D_n^i + B_n^i, w_i\}, \quad i = 1, \ldots, M,$$
$$L_{n+1}^i = \max\{L_n^i + D_n^i - B_n^i, 0\}, \quad i = 1, \ldots, M. \qquad (13)$$

Next note that the dynamics in (13) are exactly the dynamics of M decoupled make-to-order G/G/1 queues. In particular, as in the single-stage problem discussed above, L_n^i can be interpreted as the queue length in a discrete-time G/G/1 queue with arrival process $\{D_n^i; n \in \mathbb{Z}\}$ and service process $\{B_n^i; n \in \mathbb{Z}\}$ (see Figure 2). Hence, Theorem 1 holds. To apply it, however, we need the large deviations rate functions of the processes $\{D_n^i; n \in \mathbb{Z}\}$. For $i = 1$, $\{D_n^1; n \in \mathbb{Z}\}$ is the external demand process, whose large deviations rate function is assumed known. For the remaining stages $i = 2, \ldots, M$, recall that D_n^i is the demand for stage i generated by stage $i-1$. In the equivalent make-to-order version of the system D_n^i can be interpreted as the number of departures from the stage $i-1$ queue during time slot n. To see that, consider the queue corresponding to stage $i-1$, which has queue length equal to L_n^{i-1} at time slot n. (12) simply states that the queue length at slot n (L_n^{i-1}) plus the number of arrivals at slot n (D_n^{i-1}) is equal to the queue length at slot $n+1$ (L_{n+1}^{i-1}) plus the number of departures during slot n (D_n^i).

The following theorem characterizes the large deviations behaviour of the departure process $\{D_n^i; n \in \mathbb{Z}\}$, for all $i = 2, \ldots, M$. It is a corollary of a result in Bertsimas et al. (1998b) that characterizes the departure process of a G/GI/1 queue using a continuous-time model. For a proof it suffices to establish a correspondence of the random variables in the discrete-time model with the ones in a continuous-time G/G/1 queue and invoke the result in Bertsimas et al. (1998b); we omit the details.

THEOREM 2 (DEPARTURE PROCESS). *The partial sum of the departure process of the $G/G/1$ queue of stage $i-1$ satisfies*

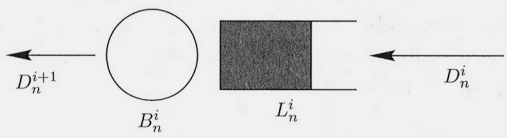
$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left[\sum_{j=1}^{n} D_j^i \geqslant na\right] = -\Lambda_{D^i}^{*+}(a), \quad i = 2, \ldots, M,$$

where $\Lambda_{D^i}^{*+}(a) = \Lambda_{B^{i-1}}^{*+}(a) + \Lambda_{\Gamma^{i-1}}^{*+}(a)$, *and*

$$\Lambda_{\Gamma^{i-1}}^{*+}(a) = \sup_{\{\theta \mid \Lambda_{D^{i-1}}^{+}(\theta) + \Lambda_{B^{i-1}}(-\theta) < 0\}} \left[\theta a - \Lambda_{D^{i-1}}^{+}(\theta)\right].$$

In Bertsimas et al. (1998b), it is also shown that the departure process satisfies the exact same technical properties that the arrival process does (some form of a sample path large deviations principle). This is key because to apply Theorem 1 to every stage $i = 2, \ldots, M$ in isolation we need the demand process D^i to satisfy a large deviations principle. Moreover, to derive the demand for the upstream stage $i+1$ we need to apply Theorem 2 which requires some form of a sample path large deviations principle. We now have all the ingredients to analyze L_n^i for every stage i in isolation. The result is summarized in the following theorem.

Figure 2. The equivalent G/G/1 queue of stage i, $i = 1, \ldots, M$, in a decoupled multistage supply chain.



THEOREM 3. *For every stage $i = 1, \ldots, M$ of the decoupled system, the steady-state queue length L^i satisfies*

$$\lim_{w_i\to\infty} \frac{1}{w_i} \log \mathbf{P}[L^i \geqslant w_i] = -\theta_{L,i}^*,$$

where $\theta_{L,i}^*$ is the largest root of the equation $\Lambda_{D^i}^{+}(\theta) + \Lambda_{B^i}(-\theta) = 0$, $\Lambda_{D^i}^{+}(\theta)$, for $i = 2, \ldots, M$, is the convex dual of $\Lambda_{D^i}^{*+}(a)$, and $\Lambda_{D^i}^{*+}(a)$ is as specified in Theorem 2.

Assume now that the stockout probability for stage 1 needs to be bounded by some ϵ_1. To operate in a regime where the decomposition is fairly accurate (see also §7) we can select the service level of stage i, ϵ_i, to be the same as, or an order of magnitude less than, the corresponding requirement, ϵ_{i-1}, for its downstream stage $i-1$. Using the results of this section, we obtain the hedging points:

$$w_i = -\frac{\log \epsilon_i}{\theta_{L,i}^*}, \quad i = 1, \ldots, M.$$

To improve the accuracy of the asymptotics, especially for fairly large ϵs, we can introduce a prefactor $f_i(w_i)$ and consider the approximation

$$\mathbf{P}[I^i \leqslant 0] = \mathbf{P}[L^i \geqslant w_i] \approx f_i(w_i)\, e^{-w_i\theta_{L,i}^*}, \quad i = 1, \ldots, M,$$

where $f_i(w_i)$ is in general any function that satisfies $\log(f_i(w_i))/w_i \to 0$ as $w_i \to \infty$ (cf. Theorem 3). Notice that this is true for any polynomial function of w_i. For renewal demand and production processes $f_i(w_i)$ is a constant (Asmussen 1987), and it is equal to 1 under M/M/1 assumptions. We will use a constant for the more general case as well. In particular, we will set $f_i(w_i) = c_i$ which yields the following approximation:

$$\mathbf{P}[I^i \leqslant 0] = \mathbf{P}[L^i \geqslant w_i] \approx c_i\, e^{-w_i\theta_{L,i}^*}, \quad i = 1, \ldots, M. \quad (14)$$

The coefficient c_i can be estimated by assuming that the above is the exact distribution of the queue length process and matching expectations to obtain (see Bertsimas and Paschalidis 2001 for details)

$$c_i = \theta_{L,i}^* \mathbf{E}[L^i]. \quad (15)$$

Note that in the decoupled system $\mathbf{E}[L^i]$ is independent of w_i, and can be obtained either by approximations of the expected queue length in a G/G/1 queue (as in Bertsimas and Paschalidis 2001) or by concurrent simulation (where one sample path of the stochastic processes involved is used to obtain $\mathbf{E}[L^i]$ for all i). Hence, the hedging point satisfies

$$w_i = -\frac{\log(\epsilon_i/c_i)}{\theta_{L,i}^*}, \quad i = 1, \ldots, M. \quad (16)$$

Numerical results that help assess the accuracy of the proposed approximation are given in §7.

5. THE MULTIECHELON APPROACH—A GLOBAL INFORMATION CASE

In this section we consider the case where echelon inventory information is available at every stage $i = 1, \dots, M$. This will allow us to trade off inventory between various stages to reduce expected inventory costs while maintaining the service level constraints. We will be using the model and notation introduced in §2. In particular, X_n^i denotes the echelon inventory at time slot n, and stage i and is defined as $X_n^i = I_n^i + \cdots + I_n^1$, $i = 1, \dots, M$. We implement an echelon base-stock production policy that maintains a hedging point or safety stock of w_i for X_n^i. More specifically, the facility at stage i produces until X_n^i reaches w_i and idles otherwise. Clearly, $w_1 \leqslant w_2 \leqslant \cdots \leqslant w_M$.

As in §4, we define the shortfall of echelon i inventory as $Y_n^i \triangleq w_i - X_n^i$, which implies $\mathbf{P}[X_n^i \leqslant 0] = \mathbf{P}[Y_n^i \geqslant w_i]$. The dynamics of the echelon inventory are

$$X_{n+1}^i = \min\{X_n^i - D_n^1 + B_n^i, w_i, X_n^i - D_n^1 + I_n^{i+1}\}, \quad i = 1, \dots, M-1, \tag{17}$$

$$X_{n+1}^M = \min\{X_n^M - D_n^1 + B_n^M, w_M\}. \tag{18}$$

In terms of the shortfalls the dynamics can be written as

$$Y_{n+1}^i = \max\{Y_n^i + D_n^1 - B_n^i, 0, Y_n^{i+1} + D_n^1 - (w_{i+1} - w_i)\}, \quad i = 1, \dots, M-1, \tag{19}$$

$$Y_{n+1}^M = \max\{Y_n^M + D_n^1 - B_n^M, 0\}. \tag{20}$$

5.1. Large Deviations Analysis of the Stockout Probability

Our first result, which is the main result of this section, is a large deviations result for the steady-state probability $\mathbf{P}[Y^1 \geqslant w_1]$, which is equal to the steady-state stockout probability $\mathbf{P}[X^1 \leqslant 0]$. We will first prove the result and then interpret it to gain insight. Recall that as in the previous section, we drop the subscript n when referring to steady-state quantities. On a notational remark, in the sequel we will be using $\mathcal{O}_i$ to denote the ith-dimensional simplex, i.e.,

$$\mathcal{O}_i = \left\{ (\xi_1, \dots, \xi_i) \mid \xi_j \in [0,1], j = 1, \dots, i, \ \sum_{j=1}^{i} \xi_j = 1 \right\}.$$

THEOREM 4. *Assume the hedging points $w_1, w_2, \dots, w_M$ in the multiechelon system (cf. (19), (20)) satisfy*

$$w_i = \beta_{i-1} w_1, \quad i = 2, \dots, M,$$

where β_i are constants and $1 \leqslant \beta_1 \leqslant \cdots \leqslant \beta_{M-1}$. The steady-state shortfall Y^1 of echelon 1 satisfies

$$\lim_{w_1 \to \infty} \frac{1}{w_1} \log \mathbf{P}[Y^1 \geqslant w_1] = -\theta_{G,1}^*, \tag{21}$$

where $\theta_{G,1}^*$ is determined by

$$\begin{aligned}
\theta_{G,1}^* = \min\Bigg[& \inf_{a>0} \frac{1}{a} \inf_{x_0 - x_1 = a} \left(\Lambda_{D^1}^{*+}(x_0) + \Lambda_{B^1}^{*-}(x_1)\right), \\
& \inf_{a>0} \frac{1}{a} \inf_{\substack{x_0 - \xi_1 x_1 - \xi_2 x_2 = a\beta_1 \\ (\xi_1, \xi_2) \in \mathcal{O}_2}} \left(\Lambda_{D^1}^{*+}(x_0) + \xi_1 \Lambda_{B^1}^{*-}(x_1) + \xi_2 \Lambda_{B^2}^{*-}(x_2)\right), \dots, \\
& \inf_{a>0} \frac{1}{a} \inf_{\substack{x_0 - \xi_1 x_1 - \cdots - \xi_M x_M = a\beta_{M-1} \\ (\xi_1, \dots, \xi_M) \in \mathcal{O}_M}} \left(\Lambda_{D^1}^{*+}(x_0) + \xi_1 \Lambda_{B^1}^{*-}(x_1) + \cdots + \xi_M \Lambda_{B^M}^{*-}(x_M)\right) \Bigg].
\end{aligned} \tag{22}$$

To establish this result we will obtain (i) a sample path characterization of Y_n^1, (ii) a lower and an upper bound on $\mathbf{P}[Y^1 \geqslant w_1]$, and (iii) show that the upper and lower bounds match up to the first degree in the exponent.

We start by obtaining a sample path characterization of Y_n^1. Suppose that at time 0, the echelon inventories are all equal to the corresponding safety stocks, i.e., $Y_0^i = 0$, for all i. At time 1, the shortfall of the echelon 1 inventory is

$$\begin{aligned}
Y_1^1 &= \max\{Y_0^1 + D_0^1 - B_0^1, 0, Y_0^2 + D_0^1 - (w_2 - w_1)\} \\
&= \max\{0, D_0^1 - \min\{B_0^1, (w_2 - w_1)\}\} \\
&= \max\{0, D_0^1 - r_{1,1}^1\},
\end{aligned}$$

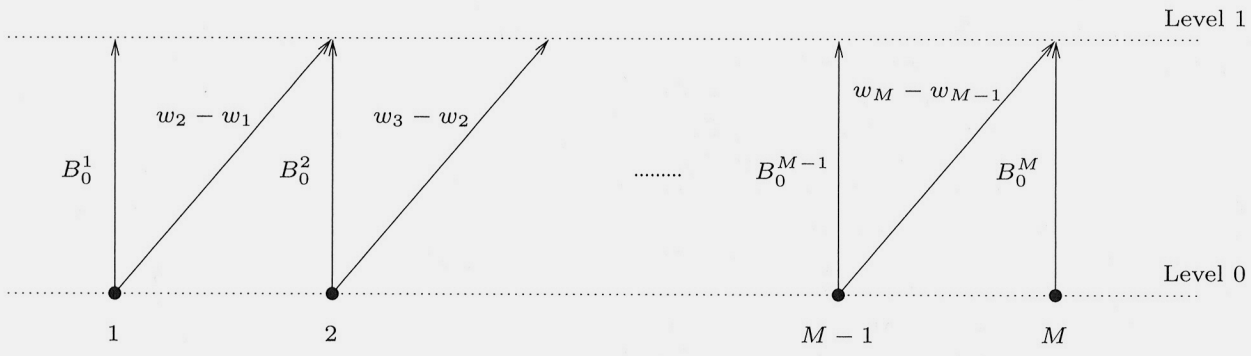
where $r_{1,1}^1 \triangleq \min\{B_0^1, (w_2 - w_1)\}$. Figure 3 depicts a graph in which $r_{1,1}^1$ can be interpreted as the length of the shortest path from point 1 (corresponding to stage 1) at level 0 to level 1. In general, we will use $r_{n,m}^i$ to denote the length of the shortest path among the paths with m hops for stage i, where n is the number of levels on the graph. For the remaining stages $i = 2, \dots, M$, we have a similar characterization, i.e.,

$$Y_1^i = \max\{0, D_0^1 - r_{1,1}^i\},$$

where $r_{1,1}^i = \min\{B_0^i, (w_{i+1} - w_i)\}$, $i = 2, \dots, M-1$, and $r_{1,1}^M = B_0^M$. In accordance with the notation we just introduced, note that in the graph of Figure 3, $r_{1,1}^i$, $i = 1, \dots, M$, denotes the length of the shortest path from point i at level 0 to level 1. At time $n = 2$,

$$\begin{aligned}
Y_2^1 &= \max\{0, Y_1^1 + D_1^1 - B_1^1, Y_1^2 + D_1^1 - (w_2 - w_1)\} \\
&= \max\{0, D_1^1 - B_1^1, D_1^1 - (w_2 - w_1), D_0^1 + D_1^1 \\
&\qquad - B_1^1 - r_{1,1}^1, D_0^1 + D_1^1 - r_{1,1}^2 - (w_2 - w_1)\} \\
&= \max\{0, D_1^1 - \min\{B_1^1, (w_2 - w_1)\}, \\
&\qquad D_0^1 + D_1^1 - \min\{B_1^1 + B_0^1, B_1^1 + (w_2 - w_1), \\
&\qquad (w_2 - w_1) + B_0^2, (w_2 - w_1) + (w_3 - w_2)\}\} \\
&= \max\{0, D_1^1 - r_{2,1}^1, D_0^1 + D_1^1 - r_{2,2}^1\},
\end{aligned}$$

Figure 3. The paths for each stage at time slot 1 (one-level graph).



where $r_{2,1}^1 \triangleq \min\{B_1^1, (w_2 - w_1)\}$ and $r_{2,2}^1 \triangleq \min\{B_1^1 + B_0^1, B_1^1 + (w_2 - w_1), (w_2 - w_1) + B_0^2, (w_2 - w_1) + (w_3 - w_2)\}$. Figure 4 depicts a two-level graph in which $r_{2,1}^1$ denotes the length of the shortest path from point 1 at level 0 to level 1, and $r_{2,2}^1$ denotes the length of the shortest path from point 1 at level 0 to level 2. Similar results can be obtained for other stages.

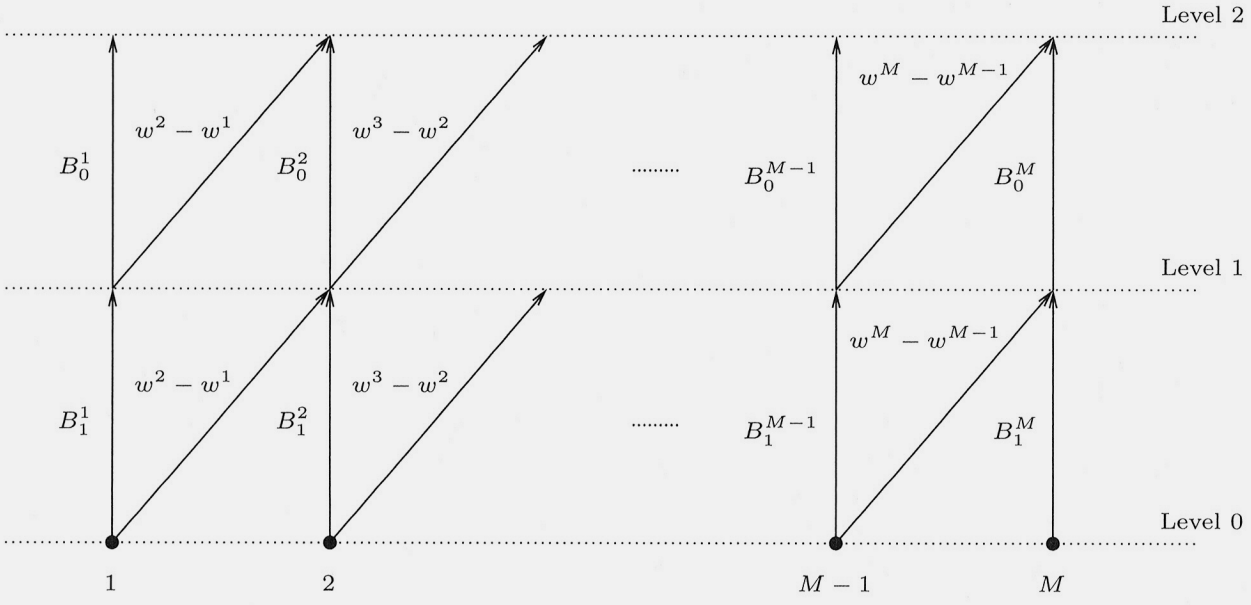
In general, the shortfall of stage 1 at time slot n is given by

$$Y_n^1 = \max\left\{0, \max_{1 \leqslant m \leqslant n}\left[\sum_{j=1}^{m} D_{n-j}^1 - r_{n,m}^1\right]\right\}, \tag{23}$$

where $r_{n,m}^1$ is equal to the length of the shortest path from point 1 at level 0 to level m in an n-level graph. A similar characterization of Y_n^1 in terms of shortest paths in a graph is given by Glasserman (1997), but for renewal demand and deterministic production processes. As we will see in the sequel, stochasticity in the production processes and dependencies in all processes involved substantially complicate the picture and require a different and more involved large deviations analysis than the one in Glasserman (1997).

Let us now denote by $\{\hat{D}_n^1;\ n \in \mathbb{Z}\}$ the time-reversed stochastic process obtained from the demand process $\{D_n^1;\ n \in \mathbb{Z}\}$. In particular, for any $k \in \mathbb{Z}$, $(\hat{D}_1^1, \hat{D}_2^1, \ldots, \hat{D}_k^1)$ has the same distribution as $(D_k^1, D_{k-1}^1, \ldots, D_1^1)$. Similarly, let $\{\hat{B}_n^i;\ n \in \mathbb{Z}\}$ denote the time-reversed production process $\{B_n^i;\ n \in \mathbb{Z}\}$ of stage i, $i = 1, \ldots, M$. Notice that because of stationarity, $\sum_{j=1}^{m} D_{n-j}^1$ has the same distribution as $\sum_{j=1}^{m} \hat{D}_j^1$. More generally, $\sum_{j=k}^{l} D_j^1$ (or $\sum_{j=k}^{l} \hat{D}_j^1$) has the same distribution as $\sum_{j=1}^{l-k+1} D_j^1$ (or $\sum_{j=1}^{l-k+1} \hat{D}_j^1$), that is, the distribution of the partial sum of demands (or time-reversed demands) during a time period depends only on the length of the period and not on the starting time. The same is true for the production processes and their reverse processes as well. Moreover, demand and production processes are independent of each other. Using these observations, Y_n^1 has the same distribution as the right-hand side of the

Figure 4. The paths of each stage at time slot 2 (two-level graph).



following equation:

$$Y_n^1 \stackrel{D}{=} \max\left\{0, \max_{1\leqslant m\leqslant n}\left[\sum_{j=1}^{m}\hat{D}_j^1 - \min_{\substack{m_1+l_1+m_2+l_2+\cdots+m_M=m\\ 0\leqslant m_i\leqslant m,\ l_i\in\{0,1\}\\ l_i=0\Rightarrow m_{i+1},l_{i+1},\cdots,m_M=0}}\left(\sum_{i=1}^{m_1}\hat{B}_i^1 + l_1(w_2-w_1) + \sum_{i=k_1+1}^{k_1+m_2}\hat{B}_i^2 + l_2(w_3-w_2) + \cdots + \sum_{i=k_{M-1}+1}^{k_{M-1}+m_M}\hat{B}_i^M\right)\right]\right\}, \quad (24)$$

where "$\stackrel{D}{=}$" denotes equality in distribution, and $k_i = i + \sum_{j=1}^{i} m_j$, for $i = 1, \ldots, M-1$. Due to the stability condition (1), a steady-state distribution exists for Y_n^1. In particular, Y_n^1 converges to Y^1 as $n \to \infty$. Therefore, using (24) we obtain

$$Y^1 \stackrel{D}{=} \max_{m\geqslant 0}\left[S_{1,m}^{\hat{D}^1} - \min_{\substack{m_1+l_1+m_2+l_2+\cdots+m_M=m\\ 0\leqslant m_i\leqslant m,\ l_i\in\{0,1\}\\ l_i=0\Rightarrow m_{i+1},l_{i+1},\cdots,m_M=0}}\left(S_{1,m_1}^{\hat{B}^1} + l_1(w_2-w_1) + S_{k_1+1,k_1+m_2}^{\hat{B}^2} + l_2(w_3-w_2) + \cdots + S_{k_{M-1}+1,k_{M-1}+m_M}^{\hat{B}^M}\right)\right], \quad (25)$$

where we use the notation introduced in (10) with the convention $\sum_{i=k+1}^{k} X_i = 0$ for any process $\{X_i;\ i \in \mathbb{Z}\}$. To facilitate handling the above expression, let us denote by G_m the argument of the maximum, i.e.,

$$Y^1 \stackrel{D}{=} \max_{m\geqslant 0} G_m.$$

We will proceed with establishing the large deviations result in (21). To that end, and in the standard large deviations methodology, we will develop a lower bound and an upper bound and show that the corresponding exponents match. We start from the lower bound. We will use the fact that for a process X and its time-reversed version $\hat{X}$, $\Lambda_X(\theta) = \Lambda_{\hat{X}}(\theta)$, which can be seen from (3). Consequently, $\Lambda_X^*(a) = \Lambda_{\hat{X}}^*(a)$.

Lower Bound. The lower bound result is summarized in the following proposition.

PROPOSITION 1. *Assume the hedging points $w_1, w_2, \ldots, w_M$ in the multiechelon system (cf. (19), (20)) satisfy $w_i = \beta_{i-1} w_1$, $i = 2, \ldots, M$, where β_i are constants and $1 \leqslant \beta_1 \leqslant \cdots \leqslant \beta_{M-1}$. The steady-state shortfall Y^1 of echelon 1 satisfies*

$$\liminf_{w_1\to\infty} \frac{1}{w_1}\log \mathbf{P}[Y^1 \geqslant w_1] \geqslant -\theta_{G,1}^*,$$

where $\theta_{G,1}^*$ is given in (22).

PROOF. For any $m \geqslant 0$ we have

$$\frac{1}{w_1}\log \mathbf{P}[Y^1 \geqslant w_1] = \frac{1}{w_1}\log \mathbf{P}\Big[\max_{m\geqslant 0} G_m \geqslant w_1\Big] \geqslant \frac{1}{w_1}\log \mathbf{P}[G_m \geqslant w_1]. \quad (26)$$

Choose $a > 0$ and write $w_1 = ma$. Then $w_i - w_{i-1} = m(\beta_{i-1} - \beta_{i-2})a$ for $i = 2, \ldots, M$, where $\beta_0 \triangleq 1$. Using (26) we obtain

$$\frac{1}{w_1}\log \mathbf{P}[Y^1 \geqslant w_1] \geqslant \frac{1}{ma}\log \mathbf{P}[G_m \geqslant ma], \quad (27)$$

and because we are interested in the regime $w_1 \to \infty$ it suffices to analyze the behaviour of the right-hand side of (27) for large values of m. To that end, select $x_i \geqslant 0$, $i = 0, \ldots, M$, $l_i \in \{0, 1\}$, $i = 1, \ldots, M-1$, and $0 \leqslant m_i \leqslant m$, $i = 1, \ldots, M$, such that $m_1 + l_1 + m_2 + l_2 + \cdots + l_{M-1} + m_M = m$,

$$mx_0 - m_1x_1 - l_1m(\beta_1 - 1)a - m_2x_2 - l_2m(\beta_2 - \beta_1)a - \cdots - m_Mx_m = ma,$$

and $l_i = 0$ implies $m_{i+1}, l_{i+1}, \ldots, m_M = 0$ for $i = 1, \ldots, M-1$. To obtain a lower bound on $\mathbf{P}[G_m \geqslant ma]$ we will construct particular sample path scenarios characterized by x_i, l_i, and m_i that lead to $G_m \geqslant ma$. More specifically, we have

$$\begin{aligned}
&\mathbf{P}[G_m \geqslant ma]\\
&\geqslant \mathbf{P}\left[\max_{\substack{m_1+l_1+\cdots+m_M=m\\ 0\leqslant m_i\leqslant m,\ l_i\in\{0,1\}\\ l_i=0\Rightarrow m_{i+1},l_{i+1},\cdots,m_M=0}}\left(S_{1,m}^{\hat{D}^1} - S_{1,m_1}^{\hat{B}^1} - l_1(w_2-w_1) - \cdots - S_{k_{M-1}+1,k_{M-1}+m_M}^{\hat{B}^M}\right) \geqslant ma\right]\\
&\geqslant \mathbf{P}\left[S_{1,m}^{\hat{D}^1} - S_{1,m_1}^{\hat{B}^1} - l_1ma(\beta_1-1) - \cdots - S_{k_{M-1}+1,k_{M-1}+m_M}^{\hat{B}^M} \geqslant ma\right]\\
&= \mathbf{P}\Big[S_{1,m}^{\hat{D}^1} - S_{1,m_1}^{\hat{B}^1} - \cdots - S_{k_{M-1}+1,k_{M-1}+m_M}^{\hat{B}^M}\\
&\qquad \geqslant ma(1 + l_1(\beta_1-1) + \cdots + l_{M-1}(\beta_{M-1}-\beta_{M-2}))\Big].
\end{aligned} \quad (28)$$

We can distinguish M cases, depending on the values we select for x_i, l_i, and m_i. In particular:

Case 1. Select $l_1 = \cdots = l_{M-1} = 0$ which implies $m_1 = m$ and $x_0 - x_1 = a$. Then from (28) we obtain

$$\begin{aligned}
\mathbf{P}[G_m \geqslant ma] &\geqslant \mathbf{P}\left[S_{1,m}^{\hat{D}^1} - S_{1,m}^{\hat{B}^1} \geqslant ma\right]\\
&\geqslant \mathbf{P}[S_{1,m}^{\hat{D}^1} \geqslant mx_0]\mathbf{P}[S_{1,m}^{\hat{B}^1} \leqslant mx_1]\\
&\geqslant e^{-m[\Lambda_{\hat{D}^1}^{*+}(x_0) + \Lambda_{\hat{B}^1}^{*-}(x_1) + \epsilon]},
\end{aligned}$$

where the last inequality above is due to the LDP principle for the processes $\hat{D}^1$ and $\hat{B}^1$ (cf. (8) and (9)) and holds for large enough m and all $\epsilon > 0$. Using (27), taking the limit as $w_1 \to \infty$, optimizing over x_0, x_1, and a to obtain a tighter bound, and recalling that demand and production processes have identical large deviations rate functions with their time-reversed versions, we conclude

$$\liminf_{w_1\to\infty} \frac{1}{w_1} \log \mathbf{P}[Y^1 \geqslant w_1] \geqslant -\inf_{a>0} \frac{1}{a} \inf_{x_0-x_1=a} [\Lambda^{*+}_{D^1}(x_0) + \Lambda^{*-}_{B^1}(x_1)]. \tag{29}$$

Case i, $i = 2, \ldots, M$. Select $l_1, \ldots, l_{i-1} = 1$, $l_i, \ldots, l_{M-1} = 0$. This implies $x_0 - \xi_1 x_1 - \cdots - \xi_i x_i = a\beta_{i-1}$, where $\xi_i = m_i/m$, $i = 1, \ldots, M$. Then from (28) we obtain

$$\begin{aligned}
&\mathbf{P}[G_m \geqslant ma] \\
&\quad \geqslant \mathbf{P}\left[S^{\hat{D}^1}_{1,m} - S^{\hat{B}^1}_{1,m_1} - \cdots - S^{\hat{B}^i}_{k_{i-1}+1,k_{i-1}+m_i} \geqslant ma\beta_{i-1}\right] \\
&\quad \geqslant \mathbf{P}\left[S^{\hat{D}^1}_{1,m} \geqslant mx_0\right]\mathbf{P}\left[S^{\hat{B}^1}_{1,m_1} \leqslant m_1 x_1\right] \\
&\qquad \cdots \mathbf{P}\left[S^{\hat{B}^i}_{k_{i-1}+1,k_{i-1}+m_i} \leqslant m_i x_i\right] \\
&\quad \geqslant e^{-m[\Lambda^{*+}_{\hat{D}^1}(x_0) + \xi_1\Lambda^{*-}_{\hat{B}^1}(x_1) + \cdots + \xi_i\Lambda^{*-}_{\hat{B}^i}(x_i) + \epsilon]},
\end{aligned}$$

where the last inequality above is due to the LDP principle for the processes $\hat{D}^1, \hat{B}^1, \ldots, \hat{B}^i$ (cf. (8) and (9)) and holds for large enough m and all $\epsilon > 0$. Note that because $m_1 + l_1 + m_2 + l_2 + \cdots + l_{M-1} + m_M = m$, our selection of l_i's and m_i's implies $m_1 + m_2 + \cdots + m_i = m - (i-1)$, which by its turn implies $\xi_1 + \cdots + \xi_i = 1$ as $m \to \infty$. As in case 1, we use (27), take the limit as $w_1 \to \infty$, and optimize over $x_0, x_1, \ldots, x_i$ and a to obtain a tighter bound, that is,

$$\begin{aligned}
\liminf_{w_1\to\infty} \frac{1}{w_1} \log \mathbf{P}[Y^1 \geqslant w_1] &\geqslant -\inf_{a>0} \frac{1}{a} \inf_{\substack{x_0-\xi_1x_1-\cdots-\xi_ix_i=a\beta_{i-1} \\ (\xi_1,\ldots,\xi_i)\in\mathcal{O}_i}} \left[\Lambda^{*+}_{D^1}(x_0)\right. \\
&\qquad \left. + \xi_1\Lambda^{*-}_{B^1}(x_1) + \cdots + \xi_i\Lambda^{*-}_{B^i}(x_i)\right].
\end{aligned} \tag{30}$$

The tightest lower bound is obtained by summarizing (29) and (30) for all $i = 2, \ldots, M$, i.e.,

$$\liminf_{w_1\to\infty} \frac{1}{w_1} \log \mathbf{P}[Y^1 \geqslant w_1] \geqslant -\theta^*_{G,1},$$

where $\theta^*_{G,1}$ is given by (22). □

Upper Bound. Next we will establish an upper bound on the probability of interest.

PROPOSITION 2. *Assume the hedging points* $w_1, w_2, \ldots, w_M$ *in the multiechelon system (cf.* (19), (20)*) satisfy* $w_i = \beta_{i-1}w_1$, $i = 2, \ldots, M$, *where* β_i *are constants and* $1 \leqslant \beta_1 \leqslant \cdots \leqslant \beta_{M-1}$. *The steady-state shortfall* Y^1 *of echelon* 1 *satisfies*

$$\limsup_{w_1\to\infty} \frac{1}{w_1} \log \mathbf{P}[Y^1 \geqslant w_1] \leqslant -\bar{\theta}^*_{G,1}, \tag{31}$$

where

$$\bar{\theta}^*_{G,1} \triangleq \min(\theta^*_1, \beta_1\theta^*_2, \ldots, \beta_{M-1}\theta^*_M), \tag{32}$$

and where

$$\theta^*_i \triangleq \sup_{\left\{\theta\geqslant 0:\ \sup_{(\xi_1,\ldots,\xi_i)\in\mathcal{O}_i} (\Lambda_{D^1}(\theta) + \xi_1\Lambda_{B^1}(-\theta) + \cdots + \xi_i\Lambda_{B^i}(-\theta)) < 0\right\}} \theta, \quad i = 1, \ldots, M. \tag{33}$$

PROOF. We have

$$\begin{aligned}
&\mathbf{P}[Y^1 \geqslant w_1] \\
&\quad = \mathbf{P}[\max_{m\geqslant 0} G_m \geqslant w_1] \\
&\quad = \mathbf{P}\left[\max_{m\geqslant 0}\left[S^{\hat{D}^1}_{1,m} - \min_{\substack{m_1+l_1+m_2+l_2+\cdots+m_M=m \\ 0\leqslant m_i\leqslant m,\ l_i\in\{0,1\} \\ l_i=0\Rightarrow m_{i+1},l_{i+1},\ldots,m_M=0}} \left(S^{\hat{B}^1}_{1,m_1}\right.\right.\right. \\
&\qquad\qquad \left.\left.\left. + l_1(w_2 - w_1) + S^{\hat{B}^2}_{k_1+1,k_1+m_2} + l_2(w_3 - w_2) \right.\right.\right. \\
&\qquad\qquad \left.\left.\left. + \cdots + S^{\hat{B}^M}_{k_{M-1}+1,k_{M-1}+m_M}\right)\right] \geqslant w_1\right] \\
&\quad = \mathbf{P}\left[\max\left\{\max_{m\geqslant 0}\left(S^{\hat{D}^1}_{1,m} - S^{\hat{B}^1}_{1,m}\right),\right.\right. \\
&\qquad \max_{\substack{m\geqslant 0 \\ m_1+m_2=m-1}} \left(S^{\hat{D}^1}_{1,m} - S^{\hat{B}^1}_{1,m_1} - (w_2 - w_1) - S^{\hat{B}^2}_{k_1+1,k_1+m_2}\right), \ldots, \\
&\qquad \max_{\substack{m\geqslant 0 \\ m_1+m_2+\cdots+m_M=m-(M-1)}} \left(S^{\hat{D}^1}_{1,m} - S^{\hat{B}^1}_{1,m_1} - (w_2 - w_1)\right. \\
&\qquad\qquad \left.\left. - \cdots - S^{\hat{B}^M}_{k_{M-1}+1,k_{M-1}+m_M}\right)\right\} \geqslant w_1\Bigg] \\
&\quad \leqslant \mathbf{P}\left[\max_{m\geqslant 0}\left(S^{\hat{D}^1}_{1,m} - S^{\hat{B}^1}_{1,m}\right) \geqslant w_1\right] \\
&\qquad + \mathbf{P}\left[\max_{\substack{m\geqslant 0 \\ m_1+m_2=m-1}} \left(S^{\hat{D}^1}_{1,m} - S^{\hat{B}^1}_{1,m_1} - S^{\hat{B}^2}_{k_1+1,k_1+m_2}\right) \geqslant \beta_1 w_1\right] \\
&\qquad + \cdots + \mathbf{P}\left[\max_{\substack{m\geqslant 0 \\ m_1+m_2+\cdots+m_M=m-(M-1)}} \left(S^{\hat{D}^1}_{1,m} - S^{\hat{B}^1}_{1,m_1} - \cdots\right.\right. \\
&\qquad\qquad \left.\left. - S^{\hat{B}^M}_{k_{M-1}+1,k_{M-1}+m_M}\right) \geqslant \beta_{M-1} w_1\right].
\end{aligned} \tag{34}$$

In the second equality above we consider all sample paths that can lead to a value larger than w_1. In particular, the first such sample path corresponds to $l_1 = 0$, the ith sample path corresponds to $l_1 = \cdots = l_{i-1} = 1$ and $l_i = 0$, for $i = 2, \ldots, M-1$, and the Mth sample path corresponds to

$l_1 = \cdots = l_{M-1} = 1$. The first inequality above bounds the probability of the maximum by the sum of the individual probabilities. Hence, it suffices to bound each term in the right-hand side of (34). We distinguish M cases:

Case 1. For the first probability in the right-hand side of (34) and for $\theta \geqslant 0$ we have

$$\begin{aligned}
&\mathbf{P}\left[\max_{m\geqslant 0}\left(S^{\hat{D}^1}_{1,m}-S^{\hat{B}^1}_{1,m}\right)\geqslant w_1\right]\\
&\quad\leqslant \mathbf{E}\left[e^{\theta\max_{m\geqslant 0}\left(S^{\hat{D}^1}_{1,m}-S^{\hat{B}^1}_{1,m}\right)}\right]e^{-\theta w_1}\\
&\quad\leqslant \sum_{m\geqslant 0}\mathbf{E}\left[e^{\theta\left(S^{\hat{D}^1}_{1,m}-S^{\hat{B}^1}_{1,m}\right)}\right]e^{-\theta w_1}\\
&\quad\leqslant \left[K_1'(\theta)+\sum_{m\geqslant m_0}e^{m(\Lambda_{\hat{D}^1}(\theta)+\Lambda_{\hat{B}^1}(-\theta)+\epsilon_1)}\right]e^{-\theta w_1}\\
&\quad\leqslant K_1(\theta,\epsilon_1)e^{-\theta w_1}\quad \text{if } \Lambda_{\hat{D}^1}(\theta)+\Lambda_{\hat{B}^1}(-\theta)<0,
\end{aligned}$$

where m_0 is sufficiently large and $\epsilon_1 > 0$. In the first inequality above we used the Markov inequality. In the third inequality above we have split the summation in two parts. Specifically, terms corresponding to $m = 0, \ldots, m_0$ are summarized in $K_1'(\theta)$. For the remaining terms we use the existence of the limiting log-moment generating function (cf. (3)). Finally, in the last inequality above, because the exponent is negative (for sufficiently small ϵ_1) the infinite series converges to some $K_1''(\theta, \epsilon_1)$, which when combined with $K_1'(\theta)$ yields $K_1(\theta, \epsilon_1)$. Optimizing over θ to obtain the tightest bound yields

$$\begin{aligned}
&\mathbf{P}\left[\max_{m\geqslant 0}\left(S^{\hat{D}^1}_{1,m}-S^{\hat{B}^1}_{1,m}\right)\geqslant w_1\right]\\
&\quad\leqslant \inf_{\{\theta\geqslant 0:\Lambda_{D^1}(\theta)+\Lambda_{B^1}(-\theta)<0\}}K_1(\theta,\epsilon_1)e^{-\theta w_1}\\
&\quad\leqslant K_1(\theta_1^*,\epsilon_1)e^{-\theta_1^* w_1},
\end{aligned}$$

where θ_1^* is as defined in (33).

Case i, $i = 2, \ldots, M$. For the ith probability in the right-hand side of (34) and for $\theta \geqslant 0$ we have

$$\begin{aligned}
&\mathbf{P}\left[\max_{\substack{m\geqslant 0\\ m_1+m_2+\cdots+m_i=m-(i-1)}}\left(S^{\hat{D}^1}_{1,m}-S^{\hat{B}^1}_{1,m_1}-\cdots-S^{\hat{B}^i}_{k_{i-1}+1,k_{i-1}+m_i}\right)\geqslant\beta_{i-1}w_1\right]\\
&\quad\leqslant \mathbf{E}\left[\exp\left\{\theta\max_{\substack{m\geqslant 0\\ m_1+\cdots+m_i=m-(i-1)}}\left(S^{\hat{D}^1}_{1,m}-S^{\hat{B}^1}_{1,m_1}-\cdots-S^{\hat{B}^i}_{k_{i-1}+1,k_{i-1}+m_i}\right)\right\}\right]e^{-\theta\beta_{i-1}w_1}\\
&\quad\leqslant \sum_{\substack{m\geqslant 0\\ m_1+\cdots+m_i=m-(i-1)}}\mathbf{E}\left[e^{\theta(S^{\hat{D}^1}_{1,m}-S^{\hat{B}^1}_{1,m_1}-\cdots-S^{\hat{B}^i}_{k_{i-1}+1,k_{i-1}+m_i})}\right]e^{-\theta\beta_{i-1}w_1}\\
&\quad\leqslant \sum_{m\geqslant 0}m^i\sup_{m_1+\cdots+m_i=m-(i-1)}\mathbf{E}\left[e^{\theta(S^{\hat{D}^1}_{1,m}-S^{\hat{B}^1}_{1,m_1}-\cdots-S^{\hat{B}^i}_{k_{i-1}+1,k_{i-1}+m_i})}\right]e^{-\theta\beta_{i-1}w_1}\\
&\quad\leqslant \left[\sum_{m\geqslant m_0}m^iK_i'(\theta,\epsilon_i)\cdot\sup_{(\xi_1,\ldots,\xi_i)\in\mathscr{O}_i}e^{m(\Lambda_{\hat{D}^1}(\theta)+\xi_1\Lambda_{\hat{B}^1}(-\theta)+\cdots+\xi_i\Lambda_{\hat{B}^i}(-\theta)+\epsilon_i)}\right]e^{-\theta\beta_{i-1}w_1}\\
&\quad\leqslant K_i(\theta,\epsilon_i)e^{-\theta\beta_{i-1}w_1}\\
&\quad\text{if }\sup_{(\xi_1,\ldots,\xi_i)\in\mathscr{O}_i}(\Lambda_{\hat{D}^1}(\theta)+\xi_1\Lambda_{\hat{B}^1}(-\theta)+\cdots+\xi_i\Lambda_{\hat{B}^i}(-\theta))<0,
\end{aligned}$$

where $\xi_i = m_i/m$, $i = 1, \ldots, M$, m_0 is sufficiently large, and $\epsilon_i > 0$. As in Case 1, in the first inequality above we used the Markov inequality, in the fourth inequality above we used the existence of the limiting log-moment generating functions, and in the last inequality above we used the fact that the infinite series converges if the exponent is negative. Optimizing over θ to obtain the tightest bound

$$\begin{aligned}
&\mathbf{P}\left[\max_{\substack{m\geqslant 0\\ m_1+m_2+\cdots+m_i=m-(i-1)}}\left(S^{\hat{D}^1}_{1,m}-S^{\hat{B}^1}_{1,m_1}-\cdots-S^{\hat{B}^i}_{k_{i-1}+1,k_{i-1}+m_i}\right)\geqslant\beta_{i-1}w_1\right]\\
&\quad\leqslant K_i(\theta_i^*,\epsilon_i)e^{-\theta_i^*\beta_{i-1}w_1},
\end{aligned}$$

where θ_i^* is as defined in (33).

Summarizing Cases $1, \ldots, M$ and using (34) we obtain that for all small enough $\epsilon_1, \ldots, \epsilon_M > 0$ and for some $K_1(\theta_1^*, \epsilon_1), \ldots, K_M(\theta_M^*, \epsilon_M)$,

$$\mathbf{P}[Y^1\geqslant w_1]\leqslant K_1(\theta_1^*,\epsilon_1)e^{-\theta_1^*w_1}+\cdots+K_M(\theta_M^*,\epsilon_M)e^{-\theta_M^*\beta_{M-1}w_1}.$$

Letting $w_1 \to \infty$ we obtain (31). □

Upper and Lower Bounds Match. Finally, we will show that the upper bound has the same exponent as the lower bound.

PROPOSITION 3. *It holds that* $\theta^*_{G,1} = \bar{\theta}^*_{G,1}$, *where* $\theta^*_{G,1}$ *and* $\bar{\theta}^*_{G,1}$ *are defined in* (22) *and* (32), *respectively.*

PROOF. It suffices to show that

$$\begin{aligned}
&\inf_{a>0}\frac{1}{a}\inf_{\substack{x_0-\xi_1x_1-\cdots-\xi_ix_i=a\beta_{i-1}\\ (\xi_1,\ldots,\xi_i)\in\mathscr{O}_i}}\left(\Lambda^{*+}_{D^1}(x_0)+\xi_1\Lambda^{*-}_{B^1}(x_1)+\cdots+\xi_i\Lambda^{*-}_{B^i}(x_i)\right)\\
&=\beta_{i-1}\theta_i^*\\
&=\beta_{i-1}\left(\sup_{\left\{\theta\geqslant 0:\sup_{(\xi_1,\ldots,\xi_i)\in\mathscr{O}_i}(\Lambda_{D^1}(\theta)+\xi_1\Lambda_{B^1}(-\theta)+\cdots+\xi_i\Lambda_{B^i}(-\theta))<0\right\}}\theta\right),
\end{aligned}$$

for all $i = 1, \ldots, M$, where $\beta_0 \triangleq 1$. To that end, notice that

$$\beta_{i-1}\left(\sup_{\left\{\theta\geqslant 0:\ \sup_{(\xi_1,\ldots,\xi_i)\in\mathscr{O}_i}(\Lambda_{D^1}(\theta)+\xi_1\Lambda_{B^1}(-\theta)+\cdots+\xi_i\Lambda_{B^i}(-\theta))<0\right\}}\theta\right)$$
$$=\sup_{\left\{\theta\geqslant 0:\ \sup_{(\xi_1,\ldots,\xi_i)\in\mathscr{O}_i}(\Lambda_{D^1}(\theta/\beta_{i-1})+\xi_1\Lambda_{B^1}(-\theta/\beta_{i-1})+\cdots+\xi_i\Lambda_{B^i}(-\theta/\beta_{i-1}))<0\right\}}\theta.$$

We will use the following lemma, which was shown in Bertsimas et al. (1999, lemma 6.2).

LEMMA 1. *For $\Lambda^*(\cdot)$ and $\Lambda(\cdot)$ being convex duals and assuming that $\Lambda(\theta) < 0$ for sufficiently small $\theta > 0$, it holds that*

$$\inf_{a>0}\frac{1}{a}\Lambda^*(a) = \theta^*,$$

where θ^* is the largest root of the equation $\Lambda(\theta) = 0$.

Notice next that

$$\inf_{\substack{x_0-\xi_1x_1-\cdots-\xi_ix_i=a\beta_{i-1}\\(\xi_1,\ldots,\xi_i)\in\mathscr{O}_i}}\left(\Lambda^{*+}_{D^1}(x_0)+\xi_1\Lambda^{*-}_{B^1}(x_1)+\cdots+\xi_i\Lambda^{*-}_{B^i}(x_i)\right) \tag{35}$$

is a convex function of a as the value function of a convex optimization problem with a appearing only in the right-hand side of the constraints. Moreover, it can be shown that it is lower-semicontinuous (by Bertsimas et al. 1999, Lemma 6.3), and thus we can apply convex duality results (Rockafellar 1970) and use Lemma 1. Finally, for any $\xi_j \geqslant 0$, $j = 1, \ldots, i$, with $\xi_1 + \cdots + \xi_i = 1$, $\Lambda_{D^1}(\theta/\beta_{i-1}) + \xi_1\Lambda_{B^1}(-\theta/\beta_{i-1}) + \cdots + \xi_i\Lambda_{B^i}(-\theta/\beta_{i-1})$ is equal to zero at $\theta = 0$ and has negative derivative at the same point due to (1), which implies that

$$\sup_{(\xi_1,\ldots,\xi_i)\in\mathscr{O}_i}\left(\Lambda_{D^1}(\theta/\beta_{i-1})+\xi_1\Lambda_{B^1}(-\theta/\beta_{i-1})+\cdots+\xi_i\Lambda_{B^i}(-\theta/\beta_{i-1})\right) \tag{36}$$

takes negative values for sufficiently small $\theta > 0$. As a final step we show that the expression in (35) is the convex dual of the expression in (36). Indeed we have

$$\sup_a\left\{\theta a-\inf_{\substack{x_0-\xi_1x_1-\cdots-\xi_ix_i=a\beta_{i-1}\\(\xi_1,\ldots,\xi_i)\in\mathscr{O}_i}}\left(\Lambda^{*+}_{D^1}(x_0)+\xi_1\Lambda^{*-}_{B^1}(x_1)+\cdots+\xi_i\Lambda^{*-}_{B^i}(x_i)\right)\right\}$$
$$=\sup_a\sup_{\substack{x_0-\xi_1x_1-\cdots-\xi_ix_i=a\beta_{i-1}\\(\xi_1,\ldots,\xi_i)\in\mathscr{O}_i}}\left\{\theta a-\Lambda^{*+}_{D^1}(x_0)-\xi_1\Lambda^{*-}_{B^1}(x_1)-\cdots-\xi_i\Lambda^{*-}_{B^i}(x_i)\right\}$$
$$=\sup_{(\xi_1,\ldots,\xi_i)\in\mathscr{O}_i}\left\{\theta\,\frac{x_0-\xi_1x_1-\cdots-\xi_ix_i}{\beta_{i-1}}-\Lambda^{*+}_{D^1}(x_0)-\xi_1\Lambda^{*-}_{B^1}(x_1)-\cdots-\xi_i\Lambda^{*-}_{B^i}(x_i)\right\}$$
$$=\sup_{(\xi_1,\ldots,\xi_i)\in\mathscr{O}_i}\left\{\Lambda_{D^1}(\theta/\beta_{i-1})+\xi_1\Lambda_{B^1}(-\theta/\beta_{i-1})+\cdots+\xi_i\Lambda_{B^i}(-\theta/\beta_{i-1})\right\}. \quad \square$$

Combining Propositions 1, 2, and 3, we obtain Theorem 4. Some remarks are in order:

(1) Theorem 4 provides us with the asymptotic decay rate for the overflow probability of the shortfall at stage 1, or equivalently, with the asymptotic decay rate of the stockout probability for the echelon inventory at stage 1. More intuitively, Theorem 4 asserts that

$$\mathbf{P}[X^1 \leqslant 0] = \mathbf{P}[Y^1 \geqslant w_1] \sim e^{-\theta^*_{G,1}w_1}. \tag{37}$$

(2) The proof of Theorem 4 characterizes the most likely path that leads to stockouts and provides intuition on how they occur. Recall from the proof that we have shown (cf. (32))

$$\theta^*_{G,1} = \min(\theta^*_1, \beta_1\theta^*_2, \ldots, \beta_{M-1}\theta^*_M),$$

where θ^*_i, $i = 1, \ldots, M$, is the largest root of the equation $\sup_{(\xi_1,\ldots,\xi_i)\in\mathscr{O}_i}(\Lambda_{D^1}(\theta) + \xi_1\Lambda_{B^1}(-\theta) + \cdots + \xi_i\Lambda_{B^i}(-\theta)) = 0$ (cf. (33)). Consider the case $\theta^*_{G,1} = \beta_{i-1}\theta^*_i$ for some $i = 1, \ldots, M$, where $\beta_0 \triangleq 1$. To avoid degenerate cases assume that all production processes B^i have distinct limiting log-moment generating functions and that $1 < \beta_1 < \cdots < \beta_{M-1}$. Let ξ^*_j, $j = 1, \ldots, i$, be the optimal solution of the optimization problem $\sup_{(\xi_1,\ldots,\xi_i)\in\mathscr{O}_i}(\Lambda_{D^1}(\theta) + \xi_1\Lambda_{B^1}(-\theta) + \cdots + \xi_i\Lambda_{B^i}(-\theta))$ at $\theta = \theta^*_i$. It can be seen that one of the ξ^*_j's, $j = 1, \ldots, i$, is equal to 1. In particular, ξ^*_i is equal to 1; otherwise, i.e., if $\xi^*_j = 1$ for some $j < i$, $\theta^*_j = \theta^*_i$ and $\beta_{j-1}\theta^*_j$ will be the minimizer in the definition of $\theta^*_{G,1}$ because $\beta_{j-1}\theta^*_j < \beta_{i-1}\theta^*_i$. Therefore, θ^*_i is the largest root of the equation $\Lambda_{D^1}(\theta) + \Lambda_{B^i}(-\theta) = 0$ and the stockout probability at stage 1 behaves as the exponential

$$e^{-\beta_{i-1}\theta^*_i w_1} = e^{-\theta^*_i w_i}.$$

Considering the single stage result (cf. Theorem 1) we can say that stage i production capacity is the "bottleneck" and characterizes the stockout probability at stage 1.

(3) Suppose that $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{M-1}) \to \infty$. Then from (32) we have $\lim_{\boldsymbol{\beta}\to\infty}\theta^*_{G,1} = \theta^*_{L,1}$, where $\theta^*_{L,1}$ is the largest root of the equation $\Lambda_{D^1}(\theta) + \Lambda_{B^1}(-\theta) = 0$. This is consistent with the result of Theorem 3. Essentially, as $\boldsymbol{\beta} \to \infty$, the various stages decompose and stage 1 is not affected by the upstream material requirement constraint, which makes Theorem 3 accurate. In general, (32) shows $\theta^*_{G,1} \leqslant \theta^*_{L,1}$, and Theorem 3 underestimates the stockout probability.

The result of Theorem 4 can be easily generalized to yield the steady-state stockout probability of the echelon inventory X^i at stages $i = 2, \ldots, M$. More specifically, we can think of echelon inventory X^i at stage i, $i = 1, \ldots, M$, as the echelon 1 inventory of an $(M + 1 - i)$-stage supply chain starting at the ith stage of the original system. Hence, generalizing Theorem 4 we obtain the following corollary.

COROLLARY 1. *Assume the base-stock levels* $w_i, w_{i+1}, \cdots, w_M$, *for* $i = 1, \ldots, M$, *in the multiechelon system satisfy* $w_{i+k} = \beta^i_{i+k-1} w_i$, $k = 1, \ldots, M-i$, *where* β^i_{i+k-1} *are constants and* $1 \leqslant \beta^i_i \leqslant \cdots \leqslant \beta^i_{M-1}$. *The steady-state shortfall* Y^i *of echelon* i *satisfies*

$$\lim_{w_i \to \infty} \frac{1}{w_i} \log \mathbf{P}\left[Y^i \geqslant w_i\right] = -\theta^*_{G,i},$$

where $\theta^*_{G,i}$ *is determined by*

$$\theta^*_{G,i} = \min\Bigg[\inf_{a>0} \frac{1}{a} \inf_{x_0 - x_i = a} \left(\Lambda^{*+}_{D^1}(x_0) + \Lambda^{*-}_{B^i}(x_i)\right),$$

$$\inf_{a>0} \frac{1}{a} \inf_{\substack{x_0 - \xi_i x_i - \xi_{i+1} x_{i+1} = a\beta^i_i \\ (\xi_i, \xi_{i+1}) \in \mathcal{O}_2}} \left(\Lambda^{*+}_{D^1}(x_0) + \xi_i \Lambda^{*-}_{B^i}(x_i) + \xi_{i+1} \Lambda^{*-}_{B^{i+1}}(x_{i+1})\right), \ldots,$$

$$\inf_{a>0} \frac{1}{a} \inf_{\substack{x_0 - \xi_i x_i - \cdots - \xi_M x_M = a\beta^i_{M-1} \\ (\xi_i, \ldots, \xi_M) \in \mathcal{O}_{M-i+1}}} \left(\Lambda^{*+}_{D^1}(x_0) + \xi_i \Lambda^{*-}_{B^i}(x_i) + \cdots + \xi_M \Lambda^{*-}_{B^M}(x_M)\right) \Bigg].$$

5.2. Refining the Large Deviations Asymptotics

Next we will discuss heuristics for refining the large deviations asymptotics. Without loss of generality we will concentrate on stage 1. The discussion easily extends to the remaining stages based on Corollary 1.

Theorem 4 provides us with the asymptotic decay rate for the stockout probability of the echelon-1 inventory as its base-stock level goes to infinity. This leads to the following approximation:

$$\mathbf{P}[Y^1 \geqslant w_1] \sim e^{-\theta^*_{G,1} w_1}. \tag{38}$$

To improve the accuracy of the approximation, especially for relatively large stockout probabilities (i.e., small safety stock w_1), we will introduce a prefactor in front of the exponential. This is in accordance with the development in §4.2, where we used a constant prefactor (cf. (14)). A constant prefactor was also used in improving the large deviations approximation in the multiclass single-stage case considered in Bertsimas and Paschalidis (2001). Here, instead, we will use the following refined approximation:

$$\mathbf{P}[Y^1 \geqslant w_1] \approx f_1(w_1, \boldsymbol{\beta}) e^{-\theta^*_{G,1} w_1}, \tag{39}$$

where the prefactor $f_1(w_1, \boldsymbol{\beta})$ is a function of w_1 and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{M-1}) = (w_2/w_1, \ldots, w_M/w_1)$. As we commented in §5.1, as $\boldsymbol{\beta} \to \infty$, different stages decouple and the upstream material requirement constraint becomes insignificant. Thus, to be consistent with the analysis of §4 we will select a function $f_1(w_1, \boldsymbol{\beta})$ that converges to a constant as $\boldsymbol{\beta} \to \infty$. In particular, $\lim_{\boldsymbol{\beta} \to \infty} f_1(w_1, \boldsymbol{\beta}) = c_1$ and $\lim_{\boldsymbol{\beta} \to \infty} \theta^*_{G,1} = \theta^*_{L,1}$, where c_1 is equal to the constant prefactor used in (14).

We will be using a function $f_1(w_1, \boldsymbol{\beta})$, which is piecewise linear in w_1 and $\boldsymbol{\beta}$. More specifically, we will evaluate the stockout probability $\mathbf{P}[Y^1 \geqslant w_1]$ at several sample points $\mathbf{w} = (w_1, \ldots, w_M)$ by simulation and then find a piecewise linear function $f_1(w_1, \boldsymbol{\beta})$ so that $f_1(w_1, \boldsymbol{\beta}) e^{-\theta^*_{G,1} w_1}$ matches the true value of $\mathbf{P}[Y^1 \geqslant w_1]$ at those sample points. This requires two main steps: (i) selecting appropriate sample points $\mathbf{w} = (w_1, \ldots, w_M)$, or equivalently points $(w_1, \boldsymbol{\beta}) = (w_1, w_2/w_1, \ldots, w_M/w_1)$, and (ii) given a "data set" of $((w_1, \boldsymbol{\beta}); f_1(w_1, \boldsymbol{\beta}))$ pairs, "fit" a function $f_1(w_1, \boldsymbol{\beta})$ to the points in the data set. We will start from step (i).

We are interested in selecting sample points $(w_1, \boldsymbol{\beta})$ such that the values of w_1 are scattered in $\mathbb{R}_+$, $\boldsymbol{\beta}$s are in the feasible set $\mathcal{B} = \{\boldsymbol{\beta} \in \mathbb{R}^{M-1}_+ : 1 \leqslant \beta_1 \leqslant \cdots \leqslant \beta_{M-1}\}$, and $(w_1, \boldsymbol{\beta})$'s adequately "explore" the feasible set $\mathbb{R}_+ \times \mathcal{B}$. Recall from the proof of Theorem 4 that (cf. (32)) $\theta^*_{G,1} = \min(\theta^*_1, \beta_1 \theta^*_2, \ldots, \beta_{M-1} \theta^*_M)$, where θ^*_i's are defined in (33). This characterization of $\theta^*_{G,1}$ separates the feasible set $\mathcal{B}$ into up to M polyhedral regions depending on which term is the minimizer in (32). Thus, to achieve enough "variety," we will pick a number of points in each of those M regions. Suppose now that we have selected a set of $K \times N$ sample points $(w^k_1, \boldsymbol{\beta}^i)$, $k = 1, \ldots, K$, $i = 1, \ldots, N$, $w^1_1 \leqslant \cdots \leqslant w^K_1$. Let $\mathbf{w}^{k,i} = (w^k_1, w^k_1 \beta^i_1, \ldots, w^k_1 \beta^i_{M-1})$ be the hedging point vector corresponding to $(w^k_1, \boldsymbol{\beta}^i)$. We simulate the system with each sample point $\mathbf{w}^{k,i}$ and obtain the stockout probability $\mathbf{P}[Y^1 \geqslant w^k_1]$. We compute

$$f_1(w^k_1, \boldsymbol{\beta}^i) = \frac{\mathbf{P}[Y^1 \geqslant w^k_1]}{e^{-\theta^*_{G,1} w^k_1}}.$$

Thus, we construct a data set consisting of pairs $((w^k_1, \boldsymbol{\beta}^i); f_1(w^k_1, \boldsymbol{\beta}^i))$. To reduce the required simulation time we can select sample points $\mathbf{w}^{1,1}, \ldots, \mathbf{w}^{K,N}$ with relatively small safety stocks which do not lead to very small stockout probabilities (such probabilities require long simulation running times). The key point here is that we use analysis to obtain the exponent of the stockout probability. We only use simulation to refine the approximation and for that it suffices to use relatively small safety stocks. The numerical results of §7 demonstrate that the proposed procedure leads to rather accurate approximations.

We will now turn our attention to step (ii) mentioned above. That is, we assume we have a "data set" consisting of $K \times N$ pairs $((w_1, \boldsymbol{\beta}); f_1(w_1, \boldsymbol{\beta}))$ and wish to fit a function $f_1(w_1, \boldsymbol{\beta})$ to the points in the data set. We will be using a function $f_1(w_1, \boldsymbol{\beta})$ which for fixed w_1 is a piecewise linear and convex function of $\boldsymbol{\beta}$ in each of the M polyhedral regions comprising $\mathcal{B}$, and for a fixed $\boldsymbol{\beta}$ is a piecewise linear function of w_1. The selection of such a function is motivated by (38). Note that due to (32)

$$e^{-\theta^*_{G,1} w_1} = \max\left(e^{-\theta^*_1 w_1}, e^{-\beta_1 \theta^*_2 w_1}, \ldots, e^{-\beta_{M-1} \theta^*_M w_1}\right),$$

which is convex in β_i in each of the M polyhedral regions, for all $i = 1, \ldots, M-1$. Of course, the proposed functional form of $f_1(w^1_1, \boldsymbol{\beta})$ is just one potential candidate that

yields satisfactory numerical results. Alternatively, given a data set consisting of $K \times N$ sample points in $\mathbb{R}_+ \times \mathcal{B}$ we can approximate $f_1(w_1, \boldsymbol{\beta})$ by some parametric form (e.g., some polynomial function or even a neural network) and then use a least squares procedure to "fit" the parametric form on the data set.

5.3. Approximating the Expected Inventory Cost

The main motivation for analyzing the echelon inventory policy was to acquire the flexibility to reduce expected inventory costs by trading off inventory between various stages while at the same time maintaining service level constraints. To that end, we need to assess expected inventory costs.

We will assume linear inventory costs. Let h_i be the holding cost for echelon-i inventory for all $i = 1, \ldots, M$. Noting that the expected echelon-i inventory is given by $\mathbf{E}[I^i] + \cdots + \mathbf{E}[I^2] + \mathbf{E}[(I^1)^+]$, where $(I^1)^+ = \max(I^1, 0)$, the total expected inventory cost is given by

$$h_1\mathbf{E}[(I^1)^+] + h_2(\mathbf{E}[(I^1)^+] + \mathbf{E}[I^2]) + \cdots + h_M(\mathbf{E}[(I^1)^+] + \mathbf{E}[I^2] + \cdots + \mathbf{E}[I^M]). \tag{40}$$

We have

$$\mathbf{E}[(I^1)^+] = \mathbf{E}[(w_1 - Y^1)^+] = w_1 - \mathbf{E}[Y^1] + \mathbf{E}[\max(0, Y^1 - w_1)]. \tag{41}$$

Using the tail distribution of Y^1 given in (39), we obtain

$$\mathbf{E}[\max(0, Y^1 - w_1)] = \int_0^\infty \mathbf{P}[Y^1 - w_1 > y]\, dy \approx f_1(w_1, \boldsymbol{\beta}) \frac{e^{-\theta^*_{G,1} w_1}}{\theta^*_{G,1}}. \tag{42}$$

For all $i \geqslant 2$, we have $I^i = (w_i - Y^i) - (w_{i-1} - Y^{i-1})$, which implies

$$\mathbf{E}[I^i] = (w_i - \mathbf{E}[Y^i]) - (w_{i-1} - \mathbf{E}[Y^{i-1}]). \tag{43}$$

Thus, combining (40), (41), (42), and (43), the total expected inventory cost can be approximated by the following expression:

$$\sum_{i=1}^{M} h_i(w_i - \mathbf{E}[Y^i]) + (h_1 + \cdots + h_M)\mathbf{E}[(Y^1 - w_1)^+] = \sum_{i=1}^{M} h_i(w_i - \mathbf{E}[Y^i]) + (h_1 + \cdots + h_M) f_1(w_1, \boldsymbol{\beta}) \frac{e^{-\theta^*_{G,1} w_1}}{\theta^*_{G,1}}. \tag{44}$$

To obtain an analytical approximation for the inventory cost we are now left with computing $\mathbf{E}[Y^i]$. This is hard to do analytically; instead we will use an approach similar to the one used in obtaining $f_1(w_1, \boldsymbol{\beta})$. We will first establish some structural properties for $\mathbf{E}[Y^i]$.

PROPOSITION 4. *Consider the multiechelon system (cf. (19), (20)) and let $0 \leqslant w_1 \leqslant w_2 \leqslant \cdots \leqslant w_M$ be the corresponding hedging points. Define $\Delta_i \triangleq w_{i+1} - w_i$, for $i = 1, \ldots, M-1$. Then $\mathbf{E}[Y^M]$ is a constant function of $(\Delta_1, \ldots, \Delta_{M-1})$. Furthermore, for all $i = 1, \ldots, M-1$, $\mathbf{E}[Y^i]$ is a function of $(\Delta_i, \ldots, \Delta_{M-1})$, which is convex and monotonically nonincreasing in every coordinate. In addition, as $\Delta_i, \ldots, \Delta_{M-1} \to \infty$, $\mathbf{E}[Y^i]$ converges to a constant.*

PROOF. Recall from (19) and (20) that the shortfalls satisfy the following evolution equations:

$$Y^i_{n+1} = \max\{Y^i_n + D^1_n - B^i_n, 0, Y^{i+1}_n + D^1_n - \Delta_i\}, \quad i = 1, \ldots, M-1, \tag{45}$$

$$Y^M_{n+1} = \max\{Y^M_n + D^1_n - B^M_n, 0\}. \tag{46}$$

Due to (1), a steady-state distribution exists for each Y^i_n, $i = 1, \ldots, M$. In particular, Y^i_n converges as $n \to \infty$ to Y^i. From the evolution equations above it is clear that $\mathbf{E}[Y^M]$ is a constant function of $(\Delta_1, \ldots, \Delta_{M-1})$.

Next, consider the echelon inventory at stage $M-1$ in three distinct systems A, B, and C. System A operates with hedging points satisfying $\Delta_{M-1} = \Delta_A$. System B operates with hedging points satisfying $\Delta_{M-1} = \Delta_B$. System C operates with hedging points satisfying $\Delta_C = \alpha\Delta_A + (1-\alpha)\Delta_B$, where $0 \leqslant \alpha \leqslant 1$. Assume without loss of generality that $\Delta_A < \Delta_B$. Let $Y^{M-1}_{A,n}$, $Y^{M-1}_{B,n}$, and $Y^{M-1}_{C,n}$ be the echelon shortfall at stage $M-1$ for systems A, B, and C, respectively, during time slot n. We define the demand and production processes for all systems A, B, and C on the same probability space so that they are driven by identical sample paths. As a result, the echelon-M shortfall in all three systems is identical for all time slots n; we will denote it by Y^M_n. We have

$$Y^{M-1}_{A,n+1} = \max\{Y^{M-1}_{A,n} + D^1_n - B^{M-1}_n, 0, Y^M_n + D^1_n - \Delta_A\},$$
$$Y^{M-1}_{B,n+1} = \max\{Y^{M-1}_{B,n} + D^1_n - B^{M-1}_n, 0, Y^M_n + D^1_n - \Delta_B\},$$
$$Y^{M-1}_{C,n+1} = \max\{Y^{M-1}_{C,n} + D^1_n - B^{M-1}_n, 0, Y^M_n + D^1_n - \Delta_C\}.$$

At time 0, let $Y^{M-1}_{A,0} = Y^{M-1}_{B,0} = Y^{M-1}_{C,0} = 0$, which trivially satisfy $\alpha Y^{M-1}_{A,0} + (1-\alpha)Y^{M-1}_{B,0} \geqslant Y^{M-1}_{C,0}$. At time slot n, suppose $\alpha Y^{M-1}_{A,n} + (1-\alpha)Y^{M-1}_{B,n} \geqslant Y^{M-1}_{C,n}$. At time $n+1$,

$$\begin{aligned}
&\alpha Y^{M-1}_{A,n+1} + (1-\alpha)Y^{M-1}_{B,n+1} \\
&\quad = \alpha \max\{Y^{M-1}_{A,n} + D^1_n - B^{M-1}_n,\ 0,\ Y^M_n + D^1_n - \Delta_A\} \\
&\qquad + (1-\alpha)\max\{Y^{M-1}_{B,n} + D^1_n - B^{M-1}_n,\ 0,\ Y^M_n + D^1_n - \Delta_B\} \\
&\quad \geqslant \max\{\alpha(Y^{M-1}_{A,n} + D^1_n - B^{M-1}_n) \\
&\qquad\quad + (1-\alpha)(Y^{M-1}_{B,n} + D^1_n - B^{M-1}_n),\ 0, \\
&\qquad\quad \alpha(Y^M_n + D^1_n - \Delta_A) + (1-\alpha)(Y^M_n + D^1_n - \Delta_B)\} \\
&\quad \geqslant \max\{Y^{M-1}_{C,n} + D^1_n - B^{M-1}_n, 0, Y^M_n + D^1_n - \Delta_C\} \\
&\quad = Y^{M-1}_{C,n+1}.
\end{aligned}$$

Therefore, for all time slots n, $\alpha Y_{A,n}^{M-1} + (1-\alpha)Y_{B,n}^{M-1} \geq Y_{C,n}^{M-1}$, which implies

$$\alpha \mathbf{E}[Y_A^{M-1}] + (1-\alpha)\mathbf{E}[Y_B^{M-1}] \geq \mathbf{E}[Y_C^{M-1}].$$

Thus, $\mathbf{E}[Y^{M-1}]$ is a convex function of Δ_{M-1}. Furthermore, from (45) it can be easily seen that $\mathbf{E}[Y^{M-1}]$ is nonincreasing in Δ_{M-1} and as $\Delta_{M-1} \to \infty$ converges to a constant. In particular, it converges to the expected shortfall of a single-stage system with demand D^1 and capacity B^{M-1} (decoupled system).

Similarly, it can be shown that $\mathbf{E}[Y^i]$, which is a function of $(\Delta_i, \ldots, \Delta_{M-1})$, is convex and nonincreasing in Δ_i and that it converges to a constant as $\Delta_i \to \infty$. Following a similar procedure, it can also be shown that for all sample paths and all time slots n, Y_{n+1}^i is a convex and nondecreasing function of Y_n^{i+1}, which by its turn is convex and nonincreasing in Δ_{i+1}, and convex and nondecreasing in Y_{n-1}^{i+2}. Therefore, $\mathbf{E}[Y^i]$ is convex and nonincreasing in Δ_{i+1}. Continuing in this fashion, we conclude that $\mathbf{E}[Y^i]$ is a function of $(\Delta_i, \ldots, \Delta_{M-1})$ that is convex and nonincreasing in every coordinate. Furthermore, as $\Delta_i, \ldots, \Delta_{M-1} \to \infty$, $\mathbf{E}[Y^i]$ converges to a constant. In particular, it converges to the expected shortfall of a single-stage system with demand D^1 and capacity B^i (decoupled system). □

Motivated by these properties of $\mathbf{E}[Y^i]$ we will approximate it using a piecewise linear convex function $g_i(\Delta_i, \ldots, \Delta_{M-1})$, using a similar approach to the one used in approximating $f_1(w_1, \boldsymbol{\beta})$ in §5.2. More specifically, we will be using the following approximation:

$$\mathbf{E}[Y^i] = g_i(w_{i+1} - w_i, \ldots, w_M - w_{M-1}), \quad i = 1, \ldots, M-1.$$

As in §5.2 we can select a number of sample points $\mathbf{w}^j$, $j = 1, \ldots, N$, and construct a piecewise linear convex function that matches $\mathbf{E}[Y^i]$ at those sample points. Note that one can evaluate $\mathbf{E}[Y^i]$ from the same simulation run used to evaluate $\mathbf{P}[Y^i \geq w_i]$, thus, the same set of sample points and simulation runs can be used to construct both $g_i(\cdot)$ and $f_1(\cdot)$.

We now have all the ingredients to pose the problem of optimizing expected inventory costs subject to maintaining service level constraints. Using the approximating expression for the expected inventory cost in (44), we have the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{M} h_i(w_i - g_i(w_{i+1} - w_i, \ldots, w_M - w_{M-1})) \\ & + \left(\sum_{i=1}^{n} h_i\right) f_1\left(w_1, \frac{w_2}{w_1}, \ldots, \frac{w_M}{w_1}\right) \frac{e^{-\theta_{G,1}^* w_1}}{\theta_{G,1}^*} \\ \text{subject to} \quad & \mathbf{P}[Y^i \geq w_i] = f_i(w_i, w_{i+1}/w_i, \ldots, w_M/w_i) \\ & \cdot e^{-\theta_{G,i}^* w_i} \leq \epsilon_i, \quad i = 1, \ldots, M, \\ & w_M \geq \cdots \geq w_2 \geq w_1 \geq 0. \end{aligned} \tag{47}$$

This problem can be solved analytically using standard nonlinear programming techniques. Because there are a number of approximations involved in this formulation, it is of interest to assess the accuracy of the solution when compared with "brute-force" simulation. We will see in §7 that the solution predicted by the problem in (47) is accurate. The very significant advantage of our approach is that we can set the proper hedging points analytically, which leads to drastic computational savings.

6. EXTENSIONS: THE MULTICLASS CASE AND LOST SALES

In this section we discuss two simple extensions: (1) to a supply-chain model that can accommodate multiple classes, and (2) to a model where unsatisfied demand at stage 1 is lost instead of backlogged.

6.1. The Multiclass Case

In the multiclass case, a production policy consists of scheduling decisions as well. That is, at each point in time and at each facility in the chain, we have not only to decide whether the facility will be working or not, but also to decide which products it will producing, if any. Finding an optimal production policy to minimize expected inventory costs subject to service level constraints appears rather hard, even in a single-stage system (see the discussion in the Introduction). Bertsimas and Paschalidis (2001) have proposed production policies in the multiclass, single-stage problem by using fluid model analysis to obtain a scheduling policy and large deviations analysis for the idling policy. In this paper we will use a scheduling policy that is motivated by fairness considerations and ease of analysis.

We extend the model depicted in Figure 1 as follows. We assume that instead of a single product class the system produces K products. We will maintain separate inventory buffers for each product class. We let $I_n^{k,j}$ denote the class k inventory at stage j and time slot n, for $k = 1, \ldots, K$ and $j = 1, \ldots, M$. We also let $D_n^{k,1}$ denote the amount of external orders arriving at stage 1 for class k during time slot n. We will implement a scheduling policy which allocates a constant fraction of the capacity of each facility to every class. In particular, we will let $\phi_{k,i}$ denote the fraction of the stage-i capacity B_n^i allocated to class k at time slot n, for all $k = 1, \ldots, K$ and $i = 1, \ldots, M$, where $\sum_{k=1}^{K} \phi_{k,i} = 1$. Note that $\phi_{k,i}$ is constant for all time slots. This policy will be referred to as the *generalized processor sharing* policy (GPS) and has in fact been analyzed in the large deviations regime by Bertsimas et al. (1999) for the two-class case; an approximate analysis for more than two classes can be found in Paschalidis (1999). The same policy has been used by Glasserman (1996) in a multiclass make-to-stock system. The GPS policy is attractive because it guarantees a minimum fraction of the capacity to every class. Thus, it can be viewed as fair because the performance of a class cannot be compromised at times that other classes are congested, as might happen for example with a priority policy.

Notice next that, according to the GPS policy, the capacity allocated to a class k can be distributed to the remaining classes during times that class k has no work to be

done. This allocation of the unutilized capacity can be done according to the weights $\phi_{k,i}$. As a result, classes are "coupled," which leads to a rather involved large deviations analysis (see Bertsimas et al. 1999). To facilitate the analysis in our supply-chain model we will decompose the system across classes and ignore the unutilized capacity allocated to a class during times that other classes are not busy. A similar decomposition assumption has been made in Glasserman (1996). Hence, the multiclass supply chain is decomposed in K single-class chains, and the results we have developed in this paper are immediately applicable. In particular, our single-class asymptotics and hedging points can be derived for each class k by using capacity $\phi_{i,k}B_n^i$ at each stage i during time slot n. The limiting log-moment generating function and the corresponding large deviations rate function of the process $\{\phi_{i,k}B_n^i;\ n \in \mathbb{Z}\}$ can be easily derived from $\Lambda_{B^i}(\theta)$ and $\Lambda^*_{B^i}(a)$, respectively. Of course, for stability purposes we have to assume $\mathbf{E}[D_n^{k,1}] < \min_{i=1,\dots,m} \phi_{i,k}\mathbf{E}[B_n^i]$, for all $k = 1, \dots, K$. This simply implies that the policy will allocate sufficient capacity to each product.

6.2. A Model with Lost Sales

We next turn our attention to a model where if inventory is not available, external demand is lost and not backordered. We will start the discussion with the multiechelon inventory model.

Consider the supply-chain model of Figure 1 operating under the echelon inventory policy of §5. Assume that unsatisfied demand is lost. Our notation for the lost sales system will parallel the one we used in §5. Let $\tilde{X}_n^i$ and $\tilde{Y}_n^i$ denote the echelon inventory and shortfall, respectively, at stage i and time slot n for $i = 1, \dots, M$. Let also $\mathbf{w} = (w_1, \dots, w_M)$ denote the hedging point vector. We can obtain an evolution equation for $\tilde{X}_n^i$ (respectively, $\tilde{Y}_n^i$) by introducing a reflecting boundary at zero (respectively, w_i) in (17) and (18) (respectively, (19) and (20)). In particular, for $\tilde{Y}_n^i$ we have

$$\tilde{Y}_{n+1}^i = \min\Big\{\max\big\{\tilde{Y}_n^i + D_n^1 - B_n^i,\ 0,\ \tilde{Y}_n^{i+1} + D_n^1 - (w_{i+1} - w_i)\big\},\ w_i\Big\}, \quad i = 1, \dots, M-1, \tag{48}$$

$$\tilde{Y}_{n+1}^M = \min\big\{\max\{\tilde{Y}_n^M + D_n^1 - B_n^M,\ 0\},\ w_i\big\}. \tag{49}$$

In the lost sales system the steady-state stockout probability is $\mathbf{P}[\tilde{X}^i = 0]$ or, equivalently, $\mathbf{P}[\tilde{Y}^i = w_i]$. As in §5, our objective is to minimize the expected inventory cost subject to maintaining these probabilities below given thresholds ϵ_i for each stage i.

Note that the steady-state stockout probability at stage 1 can be interpreted as the long-term average fraction of time that the system has no stock (under ergodicity assumptions). This can be connected with the percentage of orders that are lost. Consider the case of a Bernoulli demand process (i.e., D_n^1 is one with probability p and zero otherwise at each time slot n and independently of anything else in the system). Then the steady-state probability that an order is lost is $p\mathbf{P}[\tilde{Y}^i = w_i]$, which is the same as the expected amount of lost sales (in product units). The same reasoning does not apply for arbitrary demand processes because they may not see "time averages"; the probability that an order is lost will depend on the distribution of the demand process. To avoid such complications and have a measure that depends on the system we opted for the steady-state stockout probability to construct service-level constraints.

The main result in this subsection is that the stockout probability (or equivalently, the probability that the shortfall equals the safety stock) in the lost sales model has the exact same tail behavior as the stockout probability (or equivalently, the probability that the shortfall crosses the safety stock level) in the model with backorders of §5. More specifically, under somewhat more restrictive assumptions on the demand and production processes (some form of a sample path large deviations principle) we obtain the following theorem. The proof and a detailed description of the required assumptions is given in the Appendix.

THEOREM 5. *Assume the hedging points $w_1, w_2, \dots, w_M$ in the multiechelon lost sales system satisfy $w_i = \beta_{i-1}w_1$, $i = 2, \dots, M$, where β_i are constants and $1 \leqslant \beta_1 \leqslant \cdots \leqslant \beta_{M-1}$. The steady-state shortfall $\tilde{Y}^1$ of echelon 1 satisfies*

$$\lim_{w_1 \to \infty} \frac{1}{w_1} \log \mathbf{P}[\tilde{Y}^1 = w_1] = -\theta^*_{G,1}, \tag{50}$$

where $\theta^*_{G,1}$ is given by (22).

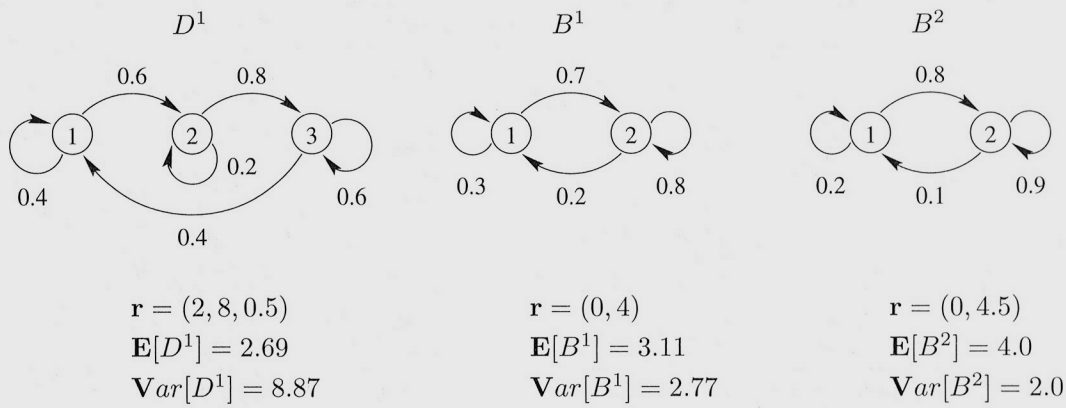
Following an analysis that parallels the one in §5.3 we can also obtain the total expected inventory cost in the lost sales system. It is given by

$$\sum_{i=1}^{M} h_i(w_i - \mathbf{E}[\tilde{Y}^i]).$$

Thus, to obtain the hedging point vector we can construct an optimization problem similar to (47).

A lost sales extension to the local information case of §4 appears to be more involved. The single-stage result of Theorem 1 can be readily extended to the lost sales model by taking the limit $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{M-1}) \to \infty$ in Theorem 5 (cf. Remark 3 in §5.1). To use our decomposition approach, though, for handling the multistage case requires characterizing the departure process of a G/G/1 queue (Theorem 2). In the lost sales model, one needs to extend that result and characterize the departure process of a finite buffer G/G/1 queue. We conjecture that this is doable along the lines of the result in Bertsimas et al. (1998b); confirming the conjecture is, however, beyond the scope of the present paper. A simple approximation can be easily developed by using the departure process of an infinite buffer G/G/1 queue to obtain a bound on the large deviations rate function of the departure process of a finite buffer queue. Using such an approximation one can apply our results in §4 to treat the local information case with lost sales.

Figure 5. The models for demand and production in Example 1, a two-stage system. We denote by $\mathbf{r}$ the vector of demand or production amounts at each state of the corresponding Markov chain.



$\mathbf{r} = (2, 8, 0.5)$
$\mathbf{E}[D^1] = 2.69$
$\mathbf{V}ar[D^1] = 8.87$

$\mathbf{r} = (0, 4)$
$\mathbf{E}[B^1] = 3.11$
$\mathbf{V}ar[B^1] = 2.77$

$\mathbf{r} = (0, 4.5)$
$\mathbf{E}[B^2] = 4.0$
$\mathbf{V}ar[B^2] = 2.0$

7. NUMERICAL RESULTS

In this section, we present numerical results to evaluate the performance of the proposed large deviations approximations. We will consider a two-stage system and we will (a) use the decomposition approach developed in §4 to derive a base-stock policy for each stage under a variety of service-level requirements, and (b) use the echelon base-stock policy analyzed in §5 to optimize the expected inventory cost subject to service-level constraints. We will also obtain the echelon base-stock policy for a three-stage system and, finally, present an example demonstrating that detailed distributional information on the demand and production processes is critical in making inventory control decisions.

Throughout this section we consider Markov-modulated demand and production processes. Figure 5 depicts the model of the demand and production processes in a two-stage supply chain. We will be referring to this system as *Example 1*. Notice that according to the mean production capacities the bottleneck is the first stage. We construct *Example 2* by exchanging the order of the two production facilities. The three-stage example we will consider will be referred to as *Example 3*.

7.1. A Two-Stage Supply Chain with the Decomposition Approach

For both examples we use the approach developed in §4.2. Using the result of Theorem 3 we compute the asymptotic decay rate of the stockout probability at each stage, namely, $(\theta^*_{L,1}, \theta^*_{L,2}) = (0.093, 0.334)$ for Example 1 and $(\theta^*_{L,1}, \theta^*_{L,2}) = (0.258, 0.093)$ for Example 2. To compute analytically the hedging points we use the expression in (16). To compute the prefactor c_i (cf. (15)) we simulated the system to obtain the expected shortfalls, which are independent from the hedging points (due to the decomposition). In Table 1 we compare analytical and simulation results for Example 1. Results for Example 2 are in Table 2. In both tables, the first two columns list the desired service-level requirements for stages 1 (final product) and 2, respectively. The third and fourth columns list the analytically computed hedging points, for stages 1 and 2, respectively. We simulated the system with these hedging points. In both tables, in columns 5–8 we report the simulated values for expected inventory and service levels. Finally, and to make comparisons with the multiechelon approach later on (cf. Tables 3 and 4), we report in the last column the simulated value

Table 1. Numerical results from the decomposition approach for Example 1.

Analytical Results				Simulation Results						
ϵ_1	ϵ_2	w_1	w_2	$\mathbf{E}[(I^1)^+]$	$\mathbf{E}[I^2]$	$\mathbf{P}[I^1 \leqslant 0]$	$\mathbf{P}[I^2 = 0]$	h_1	h_2	$\mathbf{E}[C]$
0.2	10^{-2}	15.73	10.79	8.23	9.86	0.227	1.35×10^{-2}	1	1	26.32
0.15	10^{-2}	18.82	10.79	10.70	9.86	0.175	1.35×10^{-2}	1	1	31.26
0.1	10^{-2}	23.17	10.79	14.42	9.86	0.120	1.35×10^{-2}	1	1	38.70
0.05	10^{-2}	30.21	10.79	21.20	9.86	0.063	1.35×10^{-2}	1	1	52.26
10^{-2}	10^{-3}	47.87	17.70	38.61	16.63	1.06×10^{-2}	1.08×10^{-3}	1	10	591.01
10^{-2}	10^{-3}	47.87	17.70	38.61	16.63	1.06×10^{-2}	1.08×10^{-3}	1	1	93.85
10^{-2}	10^{-3}	47.87	17.70	38.61	16.63	1.06×10^{-2}	1.08×10^{-3}	5	1	248.29
10^{-2}	10^{-3}	47.87	17.70	38.61	16.63	1.06×10^{-2}	1.08×10^{-3}	600	1	23,221.24
10^{-3}	10^{-3}	72.58	17.70	63.22	16.63	1.05×10^{-3}	1.08×10^{-3}	1	1	143.07
10^{-3}	10^{-4}	72.58	24.60	63.28	23.51	1×10^{-3}	0.96×10^{-4}	1	1	150.07
10^{-4}	10^{-4}	97.29	24.60	87.98	23.51	1.04×10^{-4}	0.98×10^{-4}	1	1	199.47
10^{-4}	10^{-5}	97.29	31.50	87.99	30.40	1.04×10^{-4}	0.94×10^{-5}	1	1	206.38

The simulated values for the expected shortfalls, which are used in computing the prefactors c_i, are $\mathbf{E}[L^1] = 9.297$ and $\mathbf{E}[L^2] = 1.098$.

Table 2. Numerical results from the decomposition approach for Example 2.

Analytical Results				Simulation Results						
ϵ_1	ϵ_2	w_1	w_2	$\mathbf{E}[(I^1)^+]$	$\mathbf{E}[I^2]$	$\mathbf{P}[I^1 \leqslant 0]$	$\mathbf{P}[I^2 = 0]$	h_1	h_2	$\mathbf{E}[C]$
0.2	10^{-2}	4.41	46.55	2.78	38.48	0.196	0.94×10^{-2}	1	1	44.04
0.15	10^{-2}	5.53	46.55	3.69	38.48	0.156	0.94×10^{-2}	1	1	45.86
0.1	10^{-2}	7.09	46.55	5.03	38.48	0.111	0.94×10^{-2}	1	1	48.54
0.05	10^{-2}	9.78	46.55	7.49	38.48	0.061	0.94×10^{-2}	1	1	53.46
10^{-2}	10^{-3}	16.01	71.25	13.61	63.06	1.02×10^{-2}	0.98×10^{-3}	1	1	90.28
10^{-2}	10^{-3}	16.01	71.25	13.61	63.06	1.02×10^{-2}	0.98×10^{-3}	5	1	144.72
10^{-2}	10^{-3}	16.01	71.25	13.61	63.06	1.02×10^{-2}	0.98×10^{-3}	1	10	780.31
10^{-3}	10^{-3}	24.92	71.25	22.48	63.06	1.35×10^{-3}	0.98×10^{-3}	1	1	108.02
10^{-3}	10^{-4}	24.92	95.96	22.50	87.75	1.14×10^{-3}	1.26×10^{-4}	1	1	132.75
10^{-4}	10^{-4}	33.83	95.96	31.41	87.74	1.16×10^{-4}	0.97×10^{-4}	1	1	150.56
10^{-4}	10^{-5}	33.83	120.67	31.41	112.45	1.08×10^{-4}	0.9×10^{-5}	1	1	175.27

The simulated values for the expected shortfalls, which are used in computing the prefactors c_i, are $\mathbf{E}[L^1] = 2.420$ and $\mathbf{E}[L^2] = 8.216$.

for the expected inventory cost under holding costs $h_1 + h_2$ and h_2 for stages 1 and 2, respectively.

In most cases, we selected the service-level requirement of the second stage to be same as, or one order of magnitude less, than ϵ_1. For relatively large stockout probabilities ($\epsilon_1 > 0.01$), we set $\epsilon_2 = 0.01$. The numerical results suggest that this suffices to make the decomposition approach valid. In particular, we observe that the proposed large deviations asymptotics are fairly accurate; they capture the exponent of the stockout probability and get fairly close in the first significant digit. They do so even for relatively large stockout probabilities, that is, away from the large deviations limit. Of course, there are many combinations of w_1 and w_2 that would lead to the same service level. Our decomposition approach yields one possible combination. In particular, the decomposition approach minimizes the required safety stock for stage 1, w_1, since it assumes that no upstream material requirement constraints are in effect. In the next subsection we explore how we can select the best such combination to minimize expected inventory costs.

7.2. A Two-Stage System with the Multiechelon Approach

Next we apply the multiechelon approach to both Example 1 and 2. We start with Example 1. Using the results of Theorem 4, Corollary 1, and the characterization of $\theta^*_{G,1}$ in (32) we obtain $(\theta^*_1, \theta^*_2) = (0.0932, 0.0932)$ and $\theta^*_{G,1} = \min(\theta^*_1, w_2\theta^*_2/w_1) = 0.0932$, for all $w_2 \geqslant w_1$. Similarly, $\theta^*_{G,2} = 0.2584$.

We selected $w_1 = 10$, 30, 50, and for each w_1, selected w_2 such that $\beta_1 = w_2/w_1 = 1$, 1.5, 2, 2.5, 3.5. We simulated the system at those sample points (w_1, w_2) and used the approach in §5.2 to construct the prefactor $f_1(w_1, \beta_1)$ for the stockout probability and the approximation $g_1(\Delta_1)$ for the expected shortfall. By simulation we also obtained $\mathbf{E}[Y^2] = 2.42$, and $c_2 = \theta^*_{G,2}\mathbf{E}[Y^2] = 0.63$, which were used as a prefactor in the echelon-2 stockout probability (as in (16)).

We solved the nonlinear programming problem in (47) for a variety of service-level requirements ϵ_1 (we imposed no service-level requirement on stage 2, i.e., $\epsilon_2 = 1$) and holding costs $\mathbf{h} = (h_1, h_2)$ for stages 1, 2, respectively. The

Table 3. Numerical results for Example 1 operated under the multiechelon policy.

		Analytical Results		Simulated Values		Simulation Results		
ϵ_1	$\mathbf{h}$	$\mathbf{w}^*_A$	$\mathbf{E}[C]$	$\mathbf{P}[X^1 \leqslant 0]$	$\mathbf{E}[C]$	$\mathbf{w}^*_S$	$\mathbf{P}[X^1 \leqslant 0]$	$\mathbf{E}[C]$
0.20	(1, 1)	(16.57, 20.71)	28.63	0.208	29.21	(17, 21)	0.199	29.81
0.15	(1, 1)	(23.41, 23.41)	34.32	0.157	34.4	(20, 24)	0.150	34.67
0.10	(1, 1)	(27.76, 27.76)	41.94	0.103	42.0	(26, 28)	0.099	42.30
0.05	(1, 1)	(35.25, 35.25)	55.84	0.052	55.86	(32, 36)	0.049	56.43
10^{-2}	(1, 10)	(52.51, 52.51)	541.25	1.01×10^{-2}	541.3	(50, 53)	0.99×10^{-2}	546.20
10^{-2}	(1, 1)	(52.57, 52.57)	89.61	1.01×10^{-2}	89.62	(50, 53)	0.99×10^{-2}	89.99
10^{-2}	(5, 1)	(52.50, 52.50)	246.67	1.01×10^{-2}	246.64	(50, 53)	0.99×10^{-2}	247.19
10^{-2}	(600, 1)	(47.72, 95.44)	23,218.00	1.03×10^{-2}	23,211.00	(48, 57)	1.00×10^{-2}	23,257.00
10^{-3}	(1, 1)	(77.21, 77.21)	138.72	1.03×10^{-3}	138.70	(74, 78)	0.99×10^{-3}	139.37
10^{-4}	(1, 1)	(101.93, 101.93)	188.13	9.96×10^{-5}	188.12	(99, 102)	9.82×10^{-5}	187.77

We denote by $\mathbf{w}^*_A$ (3rd column) the hedging vector obtained by solving the optimization problem in (47). Similarly, $\mathbf{w}^*_S$ (7th column) denotes the hedging vector obtained by brute force simulation over integer points. The 4th column ($\mathbf{E}[C]$) lists the optimal value of the optimization problem in (47), that is, our analytical approximation of the total expected inventory cost of the policy in column 3. Columns 5 and 6 list the stockout probability and expected inventory cost, respectively, obtained by simulating the policy of column 3. Columns 8 and 9 list the corresponding values obtained by simulating the policy of column 7.

Table 4. Numerical results for Example 2 under the multiechelon policy.

		Analytical Results		Simulated Values		Simulation Results		
ϵ_1	**h**	$\mathbf{w}_A^*$	**E**[C]	$\mathbf{P}[X^1 \leqslant 0]$	**E**[C]	$\mathbf{w}_S^*$	$\mathbf{P}[X^1 \leqslant 0]$	**E**[C]
0.20	(1, 1)	(8.82, 22.04)	20.87	1.98×10^{-1}	20.13	(10, 21)	0.200	19.85
0.15	(1, 1)	(10.06, 25.19)	24.83	1.52×10^{-1}	24.03	(10, 25)	0.150	23.81
0.10	(1, 1)	(11.62, 29.91)	30.65	0.102	29.93	(13, 29)	0.099	30.07
0.05	(1, 1)	(14.18, 38.05)	40.92	5.06×10^{-2}	40.37	(15, 37)	0.05	40.01
10^{-2}	(1, 1)	(21.74, 54.61)	64.73	1.03×10^{-2}	64.37	(22, 54)	1.00×10^{-2}	63.99
10^{-2}	(5, 1)	(17.75, 62.15)	129.74	1.05×10^{-2}	129.12	(18, 61)	1.00×10^{-2}	129.12
10^{-2}	(1, 10)	(26.92, 52.81)	460.32	1.03×10^{-2}	459.93	(22, 54)	1.00×10^{-2}	467.19
10^{-3}	(1, 1)	(29.26, 81.17)	98.73	1.07×10^{-3}	98.65	(30, 81)	9.95×10^{-4}	99.20
10^{-4}	(1, 1)	(38.21, 106.00)	132.49	1.22×10^{-4}	132.47	(40, 107)	1.00×10^{-4}	135.46

The notation and the structure of the table are the same as in Table 3.

results are reported in Table 3. We also obtained the optimal policy (w_1, w_2) by brute-force simulation. There are two main observations we can make:

(1) Our analytical approximation for the stockout probability and the expected inventory cost is very accurate. To see that compare (i) the actual stockout probability (column 5) achieved by the "analytical" optimal solution $\mathbf{w}_A^*$ with the corresponding service level requirement ϵ_1, and (ii) the actual inventory cost of $\mathbf{w}_A^*$ (column 6) with its analytical approximation (column 4). Our results are accurate even for relatively large stockout probabilities, that is, away from the large deviations limiting regime.

(2) The performance of our analytically obtained policy $\mathbf{w}_A^*$ is rather close to the optimal policy (obtained by simulation). In fact, our policy is within at most 2% of the optimal (difference of columns 6 and 9), which drops to at most 1% if we ignore the first row of the table.[2]

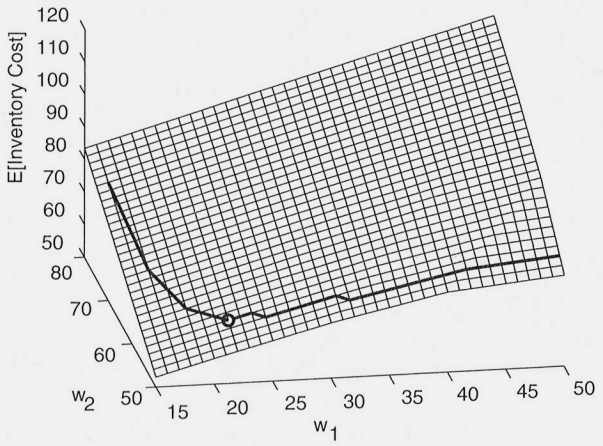
To assess the efficiency of the analytical approach, note that to optimize the expected inventory cost by simulation we need to simulate for all possible integer combinations of w_1 and w_2 and select the one that yields the lowest cost. Moreover, simulating small stockout probabilities requires very long sample paths. It usually takes from several hours to several days to find the optimal by brute-force simulation, depending on the length of sample paths (as dictated by the service level requirements) and the number of (w_1, w_2) points. In fact, these running times of the brute-force simulation were achieved by using information from our analytical solution, that is, starting our search in a set "centered" around the analytical solution where we expect the optimal to be. Brute-force simulation with no information at all would take much longer and would be computationally intractable for the smaller ϵ_1 (last rows of Table 3). The nonlinear programming problem can typically be solved within 1 minute (for the instances we considered), while "pre-processing" (i.e., obtaining the prefactors) took on the order of 30 minutes. It is evident that the proposed analytical approach leads to huge computational savings at a modest performance cost. It should also be noted that in finding $\mathbf{w}_S^*$ we only considered integer valued hedging points w_1 and w_2. As a result, the gap between $\mathbf{w}_A^*$ and $\mathbf{w}_S^*$ in Table 3 contains this quantization error and thus overestimates the actual error of the analytical approach.

To compare the solution obtained by the multiechelon approach with the one obtained by the decomposition approach consider the inventory cost of the latter given in Table 1 (last column).[3] Because we ignore coupling among stages the decomposition approach is not as accurate in approximating $\mathbf{P}[I^1 \leqslant 0]$. Thus, in some cases it leads to solutions that violate the service-level requirement (more in rows 1–4 of Table 1 and slightly in some of the remaining). In terms of inventory cost, the multiechelon approach leads, in general, to more efficient solutions (except in rows 1–4 and 8 of Table 1 in which we end up with less inventory cost because we violate the service-level requirement).

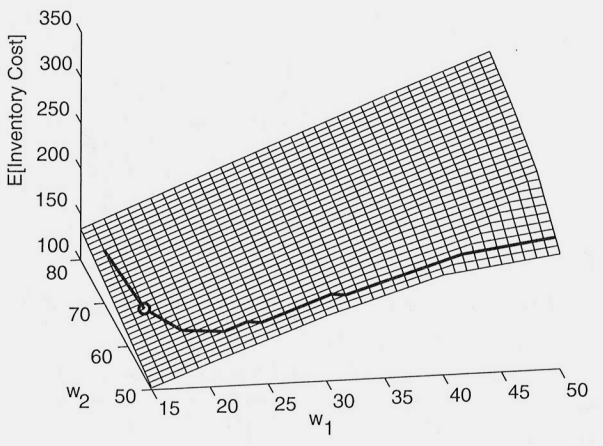
Next, we consider Example 2. We compute $(\theta_1^*, \theta_2^*) = (0.2584, 0.0932)$, $\theta_{G,1}^* = \min(\theta_1^*, w_2\theta_2^*/w_1) = \min(0.2584, 0.0932w_2/w_1)$, and $\theta_{G,2}^* = 0.0932$. As in Example 1, we simulated the system with sample points (w_1, w_2), using $w_1 = 10, 20, 30, 40$, and selecting w_2 such that $\beta_1 = w_2/w_1 = 1, 2, 2.5, 2.7736, 3, 4$, for each w_1. We constructed the prefactor $f_1(w_1, \beta_1)$ for the stockout probability and the approximation $g_1(\Delta_1)$ for the expected shortfall using the approach of §5.2. By simulation we also obtained $\mathbf{E}[Y^2] = 9.28$, and $c_2 = \theta_{G,2}^* \mathbf{E}[Y^2] = 0.865$, which we used as a prefactor in the stockout probability at stage 2.

Solving the optimization problem in (47) we obtain the results of Table 4, which are very similar in nature to the ones we obtained for Example 1. That is, our approximations are accurate for both stockout probabilities and expected inventory costs and the analytical solution is within at most 2.2% of the optimal. Figure 6 depicts how the expected inventory cost (obtained by simulation) changes with the hedging vector for the cases in rows 5 and 6 of Table 4. It can be seen that the policy obtained by our analytical approach is very close to optimal; deviating from $\mathbf{w}_A^*$ can lead to significantly larger expected inventory cost, which stresses the significance of optimization. Finally, as in Example 1, by comparing the inventory costs in Tables 2 (last column) and Table 4 we conclude that the multiechelon policy leads to more economic solutions.

Figure 6. The optimal multiechelon policies for Example 2 obtained by simulation. The thick curve is the boundary of the (feasible) set of (w_1, w_2) satisfying the service-level constraints. The optimal policy obtained by our analytical approach is marked with a circle.



(a) $(h_1, h_2) = (1, 1)$, $\epsilon_1 = 10^{-2}$.



(b) $(h_1, h_2) = (5, 1)$, $\epsilon_1 = 10^{-2}$.

7.3. A Three-Stage System with the Multiechelon Approach

We next consider a three-stage system (Example 3), with the Markov-modulated demand and production processes depicted in Figure 7. We apply the multiechelon approach to Example 3. Using the results of §5 (cf. (32), (33)) we obtain $\theta_1^* = 0.1276$, $\theta_2^* = 0.0866$, $\theta_3^* = 0.0866$, and $\theta_{G,1}^* = \min(\theta_1^*, w_2\theta_2^*/w_1, w_3\theta_3^*/w_1) = \min(\theta_1^*, w_2\theta_2^*/w_1)$. Therefore, the feasible set of $\boldsymbol{\beta} = (\beta_1, \beta_2) = (w_2/w_1, w_3/w_1)$ has two regions determined by β_1: $1 \leqslant \beta_1 < \theta_1^*/\theta_2^*$ and $\beta_1 \geqslant \theta_1^*/\theta_2^*$. We select $w_1 = 20, 40, 60, 80$, and for each w_1, select a set of (w_2, w_3) to include sample points on the boundary and inside the two regions of $\boldsymbol{\beta}$. Using the approach of §5.2 we obtained $f_1(w_1, \boldsymbol{\beta})$ and $g_i(\cdot)$, $i = 1, 2, 3$, and, as a result, the stockout probability and the expected inventory cost.

We solved the nonlinear programming problem in (47) for a variety of service-level requirements ϵ_1 and holding costs $\mathbf{h} = (h_1, h_2, h_3)$. The results are reported in Table 5. Again, we observe that the analytical results are very close to the ones obtained by simulation. That is, our approximations are accurate for both stockout probabilities and expected inventory costs and the analytical solution is within at most 2.5% of the optimal. Regarding running times, as we noted in §7.2 the nonlinear programming problem takes on the order of a couple of minutes, "preprocessing" took on the order of a couple of hours for this three-stage example, while the brute-force simulation approach takes on the order of several days, depending on the service-level requirement and the number of hedging point vectors it goes through. Again, as in §7.2, this is the case for a simulation that uses information from our analytical solution.

7.4. Significance of Distributional Information

As our final example we present a two-stage supply chain model operated under the multiechelon inventory policy. We will demonstrate that distributional information on the demand and service processes is critical in making inventory control decisions. In particular, we will try to identify the bottleneck stage that determines the stockout probability at stage 1.

The demand and production processes are all discrete-time Markov modulated processes. Letting $\mathbf{P}$ and $\mathbf{r}$ denote the transition probabilities and the vector of demand or production amounts in each state of the corresponding Markov

Figure 7. The models of demand and production in Example 3.

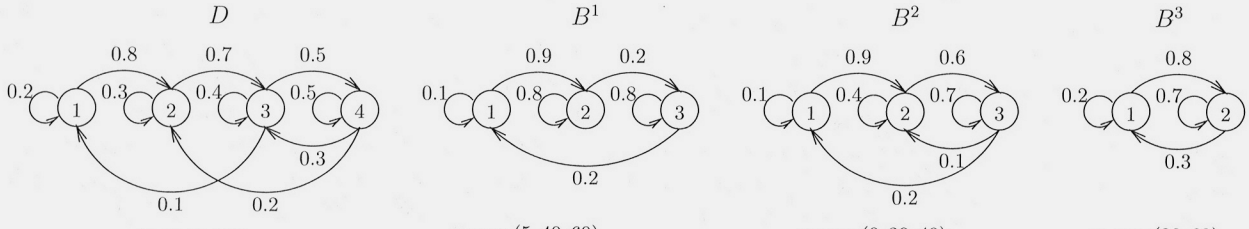


Table 5. Numerical results for Example 3 operated under the multiechelon policy.

		Analytical Results		Simulated Values		Simulation Results		
ϵ_1	**h**	$\mathbf{w}_A^*$	$\mathbf{E}[C]$	$\mathbf{P}[X^1 \leqslant 0]$	$\mathbf{E}[C]$	$\mathbf{w}_S^*$	$\mathbf{P}[X^1 \leqslant 0]$	$\mathbf{E}[C]$
10^{-1}	(1, 1, 1)	(27.5, 53.7, 68.7)	129.8	1.14×10^{-1}	132.6	(23, 51, 74)	0.98×10^{-1}	136.0
7×10^{-2}	(1, 1, 1)	(29.3, 59.0, 74.0)	142.8	7.55×10^{-2}	145.0	(23, 56, 80)	6.82×10^{-2}	147.6
3×10^{-2}	(1, 1, 1)	(35.2, 68.1, 88.1)	175.1	3.36×10^{-2}	176.4	(28, 64, 92)	2.97×10^{-2}	173.4
10^{-2}	(1, 1, 1)	(42.5, 83.7, 103.7)	215.6	1.15×10^{-2}	215.8	(34, 88, 104)	1.01×10^{-2}	211.5
10^{-2}	(1, 3, 1)	(43.9, 78.5, 108.5)	360.2	1.20×10^{-2}	362.7	(34, 85, 106)	1.01×10^{-2}	364.1
10^{-3}	(1, 1, 1)	(59.7, 116.2, 130.0)	289.9	1.08×10^{-3}	290.1	(57, 116, 130)	1.01×10^{-3}	287.5
10^{-4}	(1, 1, 1)	(75.9, 134.4, 162.6)	362.3	0.97×10^{-4}	363.5	(69, 135, 166)	1.00×10^{-4}	360.6

The notation and the structure of the table are the same as in Table 3.

chain we set $\mathbf{r}_D = (5, 10)$, $\mathbf{r}_{B^1} = (0, 25)$, $\mathbf{r}_{B^2} = (0, 14)$,

$$\mathbf{P}_D = \begin{bmatrix} 0.2 & 0.8 \\ 0.4 & 0.6 \end{bmatrix}, \quad \mathbf{P}_{B^1} = \begin{bmatrix} 0.2 & 0.8 \\ 0.3 & 0.7 \end{bmatrix},$$

$$\mathbf{P}_{B^2} = \begin{bmatrix} 0.15 & 0.85 \\ 0.05 & 0.95 \end{bmatrix},$$

which implies $\mathbf{E}[D] = 8.33$, $\mathbf{E}[B^1] = 18.18$, and $\mathbf{E}[B^2] = 13.22$. Applying the results of §5 (cf. (32), (33)) we obtain $\theta_1^* = 0.1785$, $\theta_2^* = 0.1785$. We compute $\theta_{G,1}^* = \min(\theta_1^*, w_2\theta_2^*/w_1) = \theta_1^* = 0.1785$, which (according to the discussion in Remark 2 of §5) implies that the "bottleneck" is stage 1 in the sense that process B^1 and not B^2 characterizes the stockout probability at stage 1. This seems to contradict the naive intuition that the "bottleneck" is stage 2 because $\mathbf{E}[B^1] > \mathbf{E}[B^2]$! The conclusion that the "bottleneck" is stage 1 is explained by noting that B^1 is more bursty than B^2.

8. CONCLUSIONS

We proposed two production policies in a multistage, single-class supply chain. Demand and service processes are general, potentially autocorrelated processes, which makes it possible to model complex demand scenarios and failure-prone production facilities. Both policies emphasize quality of service, which is becoming important in modern manufacturing, by maintaining desirable service level constraints. The first policy is a base-stock policy that uses only local inventory information. The second policy is an echelon-base stock policy. In both cases we relied upon large deviations techniques for analysis. This led to asymptotically tight approximations for the stockout probabilities which allows us to analytically obtain appropriate hedging points that maintain the desirable service level constraints. Our analysis under the echelon policy provides particular insight on how stockouts occur. In particular, it identifies a "bottleneck" stage whose production capacity is "responsible" for stockouts at stage 1. But, this "bottleneck" stage is not necessarily the one with the smallest mean production capacity; it depends on the full distribution of production processes.

The echelon base-stock policy enables optimization among all possible hedging point vectors that satisfy the service-level constraints; by solving a nonlinear optimization problem we select the one with minimum expected inventory cost. Numerical results show that the solutions obtained by analysis are very close to the ones obtained by brute-force simulation. Our analytic approach for selecting appropriate hedging points leads to dramatic computational savings when compared to the time needed to obtain them by simulation.

APPENDIX

Here we provide the proof of Theorem 5. As we indicated in §6.2 we will need somewhat more restrictive assumptions on the demand and production processes. We will require that they satisfy the following version of a *sample path large deviations principle* (SPLDP) (see Bertsimas et al. 1999 for an extended discussion on SPLDPs). More specifically, let $\{X_j; j \in \mathbb{Z}\}$ denote any of the (demand or production) processes $\{D_j^1; j \in \mathbb{Z}\}$, $\{B_j^i; j \in \mathbb{Z}\}$, $i = 1, \ldots, M$, and let $S_{j,k}^X = \sum_{r=j}^k X_r$ denote the partial sums of the process X. We will be assuming that for all $m \in \mathbb{N}$, for every $\epsilon_1, \epsilon_2 > 0$, and for every set of scalars $a_0, \ldots, a_{m-1}$, there exists $K > 0$ such that for all $n \geqslant K$ and all $k_0, \ldots, k_m$ with $0 = k_0 \leqslant k_1 \leqslant \cdots \leqslant k_m = n$,

$$e^{-(n\epsilon_2 + \sum_{i=0}^{m-1}(k_{i+1}-k_i)\Lambda_X^*(a_i))} \leqslant \mathbf{P}\big[|S_{k_i+1,k_{i+1}}^X - (k_{i+1}-k_i)a_i| \leqslant \epsilon_1 n, \ i = 0, \ldots, m-1\big]. \tag{51}$$

Intuitively, this assumption deals with the probability of sample paths being constrained within a tube around a "polygonal" path made up of linear segments of slopes $a_0, \ldots, a_{m-1}$. This assumption is satisfied by a large class of processes, including, renewal, Markov-modulated, and stationary processes with mild mixing conditions (see Chang 1995). It can also be seen from the derivation of the large deviations rate function in Chang (1995) that the time-reversed process $\hat{X}$ (see §5) has the same large deviations rate function as the forward process. To prove Theorem 5 we first need an alternative expression for $\theta_{G,1}^*$ in (22).

LEMMA A.1. *It holds that*

$$\theta^*_{G,1} = \min\Bigg[\inf_{a>0}\frac{1}{a}\inf_{x_0-x_1=a}\big(\Lambda^*_{D^1}(x_0)+\Lambda^*_{B^1}(x_1)\big),$$
$$\inf_{a>0}\frac{1}{a}\inf_{\substack{x_0-\xi_1x_1-\xi_2x_2=a\beta_1\\(\xi_1,\xi_2)\in\mathscr{O}_2}}\big(\Lambda^*_{D^1}(x_0)$$
$$+\xi_1\Lambda^*_{B^1}(x_1)+\xi_2\Lambda^*_{B^2}(x_2)\big),\ldots,$$
$$\inf_{a>0}\frac{1}{a}\inf_{\substack{x_0-\xi_1x_1-\cdots-\xi_Mx_M=a\beta_{M-1}\\(\xi_1,\ldots,\xi_M)\in\mathscr{O}_M}}\big(\Lambda^*_{D^1}(x_0)$$
$$+\xi_1\Lambda^*_{B^1}(x_1)+\cdots+\xi_M\Lambda^*_{B^M}(x_M)\big)\Bigg]. \tag{52}$$

PROOF. Comparing (22) and (52), it suffices to show

$$\inf_{a>0}\frac{1}{a}\inf_{\substack{x_0-\xi_1x_1-\cdots-\xi_ix_i=a\beta_{i-1}\\(\xi_1,\ldots,\xi_i)\in\mathscr{O}_i}}\big(\Lambda^{*+}_{D^1}(x_0)$$
$$+\xi_1\Lambda^{*-}_{B^1}(x_1)+\cdots+\xi_i\Lambda^{*-}_{B^i}(x_i)\big)$$
$$=\inf_{a>0}\frac{1}{a}\inf_{\substack{x_0-\xi_1x_1-\cdots-\xi_ix_i=a\beta_{i-1}\\(\xi_1,\ldots,\xi_i)\in\mathscr{O}_i}}\big(\Lambda^*_{D^1}(x_0)$$
$$+\xi_1\Lambda^*_{B^1}(x_1)+\cdots+\xi_i\Lambda^*_{B^i}(x_i)\big)$$

for some arbitrary $i=1,\ldots,M$. We will denote by LHS (respectively, RHS) the expression on the left-hand side (respectively, right-hand side) of the the above. First observe that LHS $\leqslant$ RHS, because for any process X and any a we have $\Lambda^{*-}_X(a)\leqslant\Lambda^*_X(a)$ and $\Lambda^{*+}_X(a)\leqslant\Lambda^*_X(a)$ (cf. (6)).

Next, consider an optimal solution $\mathbf{y}=(a^*,x_0^*,\ldots,x_i^*,\xi_1^*,\ldots,\xi_i^*)$ of the optimization problem in the LHS. Without loss of generality assume that $\xi_1^*,\ldots,\xi_i^*>0$; otherwise some terms will be eliminated from the objective function and the rest of the proof carries through. Fix $\epsilon>0$, sufficiently small. We will construct a feasible solution $\mathbf{y}'(\epsilon)$ of the LHS that is also optimal. We will distinguish several cases:

(1) Suppose $x_0^*\geqslant\mathbf{E}[D^1]$ and $x_j^*\leqslant\mathbf{E}[B^j]$ for all $j=1,\ldots,i$. Then set $\mathbf{y}'=\mathbf{y}$.

(2) Suppose $x_0^*<\mathbf{E}[D^1]$. Note that by feasibility $x_0^*-\xi_1^*x_1^*-\cdots-\xi_i^*x_i^*\geqslant 0$. This implies that for some $j=1,\ldots,i$, $x_j^*<\mathbf{E}[B^j]$. Otherwise, i.e., if $x_j^*\geqslant\mathbf{E}[B^j]$ for all $j=1,\ldots,i$, due to (1) we have

$$x_0^*-\xi_1^*x_1^*-\cdots-\xi_i^*x_i^*$$
$$<\mathbf{E}[D^1]-\xi_1^*\mathbf{E}[B^1]-\cdots-\xi_i^*\mathbf{E}[B^i]$$
$$\leqslant\mathbf{E}[D^1]-\min_{j\in\{1,\ldots,i\}}\mathbf{E}[B^j]<0.$$

Then set $\mathbf{y}'=(a^*,x_0^*+\epsilon,x_1^*,\ldots,x_j^*+\epsilon/\xi_j^*,\ldots,x_i^*,\xi_1^*,\ldots,\xi_i^*)$, which is feasible. Note that because $\Lambda^{*+}_{D^1}(\cdot)$ is zero below the mean $\mathbf{E}[D^1]$ and $\Lambda^{*-}_{B^j}(\cdot)$ is nonincreasing below the mean $\mathbf{E}[B^j]$ the objective value of the problem in the LHS at $\mathbf{y}'$ is no more than the corresponding value at $\mathbf{y}$. Hence, $\mathbf{y}'$ is also optimal.

(3) Suppose that for some j, $j=1,\ldots,i$, $x_j^*>\mathbf{E}[B^j]$. We distinguish two cases:

(a) Suppose $x_0^*>\mathbf{E}[D^1]$. Then set $\mathbf{y}'=(a^*,x_0^*-\xi_j^*\epsilon,x_1^*,\ldots,x_j^*-\epsilon,\ldots,x_i^*,\xi_1^*,\ldots,\xi_i^*)$, which is feasible. Note that because $\Lambda^{*+}_{D^1}(\cdot)$ is nondecreasing above the mean $\mathbf{E}[D^1]$ and $\Lambda^{*-}_{B^j}(\cdot)$ is zero above the mean $\mathbf{E}[B^j]$ the objective value of the problem in the LHS at $\mathbf{y}'$ is no more than the corresponding value at $\mathbf{y}$. Hence, $\mathbf{y}'$ is also optimal.

(b) Finally, suppose that $x_0^*=\mathbf{E}[D^1]$. Then, as in Case 2 above, for some $j'=1,\ldots,i$ not equal to j we have $x_{j'}^*<\mathbf{E}[B^{j'}]$. Then set $\mathbf{y}'=(a^*,x_0^*,x_1^*,\ldots,x_j^*-\epsilon/\xi_j^*,\ldots,x_{j'}^*+\epsilon/\xi_{j'}^*,\ldots,x_i^*,\xi_1^*,\ldots,\xi_i^*)$, which is feasible. Note that because $\Lambda^{*-}_{B^{j'}}(\cdot)$ is nonincreasing below the mean $\mathbf{E}[B^{j'}]$ and $\Lambda^{*-}_{B^j}(\cdot)$ is equal to zero above the mean $\mathbf{E}[B^j]$ the objective value of the optimization problem in the LHS at $\mathbf{y}'$ is no more than the corresponding value at $\mathbf{y}$. Hence, $\mathbf{y}'$ is also optimal.

Given $\mathbf{y}$ we keep repeating the procedure in Cases 2 and 3 above until we construct a new optimal solution $\mathbf{y}'=(a',x_0',\ldots,x_i',\xi_1',\ldots,\xi_i')$ that satisfies $x_0'\geqslant\mathbf{E}[D^1]$ and $x_j'\leqslant\mathbf{E}[B^j]$ for all $j=1,\ldots,i$. Such a $\mathbf{y}'$ is feasible for the optimization problem in the RHS and achieves the same objective value for both RHS and LHS. The optimal value of the RHS can be no worse. Thus, RHS $\leqslant$ LHS. □

PROOF OF THEOREM 5. Recall that for each time slot n the lost sales system satisfies the evolution equations (48) and (49), while the system with backorders satisfies (19) and (20). We define demand and production processes on the same probability space for both systems so that they are driven by identical sample paths. It holds that $\tilde{Y}_n^1\leqslant Y_n^1$, for all n. Hence, by using Propositions 2 and 3 we obtain

$$\limsup_{w_1\to\infty}\frac{1}{w_1}\log\mathbf{P}[\tilde{Y}^1=w_1]\leqslant\limsup_{w_1\to\infty}\frac{1}{w_1}\log\mathbf{P}[Y^1\geqslant w_1]$$
$$\leqslant-\theta^*_{G,1}.$$

For the lower bound we will mimic the proof of Proposition 1. The key of that proof is that we identified M scenarios (Cases $1,\ldots,M$) which led to $Y^1\geqslant w_1$. The probabilities of these scenarios provide M lower bounds; we selected the tightest by maximizing over those. Here we will establish that these scenarios are also feasible scenarios in the lost sales model and lead to $\tilde{Y}_n^1=w_1$. Using the same notation as in the proof of Proposition 1, let m be large enough, choose $a>0$, set $w_1=ma$ and consider the following M scenarios:

Scenario 1.

$$\{|S^{\hat{D}^1}_{1,j}-jx_0|\leqslant\epsilon_0 m,\ j=1,\ldots,m\},$$

$$\{|S^{\hat{B}^1}_{1,j}-jx_1|\leqslant\epsilon_1 m,\ j=1,\ldots,m\},$$

where $x_0,x_1\geqslant 0$, $\epsilon_0,\epsilon_1>0$, $x_0-x_1=a+\epsilon'$, and $\epsilon'=\epsilon_0+\epsilon_1$.

Scenarios $i=2,\ldots,M$.

$$\big\{|S^{\hat{D}^1}_{1,j}-jx_0|\leqslant\epsilon_0 m,\ j=1,\ldots,m\big\},$$

$$\left\{|S^{\hat{B}^1}_{1,j} - jx_1| \leqslant \epsilon_1 m_1,\ j = 1, \ldots, m_1\right\},$$

$$\left\{|S^{\hat{B}^i}_{k_{i-1}+1,k_{i-1}+j} - jx_i| \leqslant \epsilon_i m_i,\ j = 1, \ldots, m_i\right\},$$

where $x_0, x_1, \ldots, x_i \geqslant 0$, $\epsilon_0, \ldots, \epsilon_i > 0$, $x_0 - \xi_1 x_1 - \cdots - \xi_i x_i = (a + \epsilon')\beta_{i-1}$, $\xi_j = m_j/m$ for $j = 1, \ldots, i$, $(\xi_1, \ldots, \xi_i) \in \mathcal{O}_i$, $\epsilon' = \epsilon_0 + \cdots + \epsilon_i$, and $k_{i-1} = (i-1) + \sum_{j=1}^{i-1} m_j$.

Using the same arguments as in the proof of Proposition 1, according to scenario i the shortfall in the system with backorders builds up linearly with m at a rate of $a+\epsilon'$, where $\epsilon' \to 0$ as $\epsilon_0, \ldots, \epsilon_i \to 0$. It reaches $m(a+\epsilon')$ in m time slots. Now from (48) and (49) note that starting from zero, $\tilde{Y}^1_n$ and Y^1_n follow identical sample paths until they hit w_1. Hence, $\tilde{Y}^1_n$ reaches $w_1 = ma$ in m time slots. Thus, using the same notation as in the proof of Proposition 1, for every $i = 1, \ldots, M$, we have

$$\begin{aligned}
&\mathbf{P}[\tilde{Y}^1 = ma] \\
&\quad \geqslant \mathbf{P}[\min\{G_m, ma\} = ma] \\
&\quad \geqslant \mathbf{P}[|S^{\hat{D}^1}_{1,j} - jx_0| \leqslant \epsilon_0 m,\ j = 1, \ldots, m] \\
&\qquad \times \mathbf{P}[|S^{\hat{B}^1}_{1,j} - jx_1| \leqslant \epsilon_1 m_1,\ j = 1, \ldots, m_1] \\
&\qquad \times \cdots \times \mathbf{P}[|S^{\hat{B}^i}_{k_{i-1}+1,k_{i-1}+j} - jx_i| \leqslant \epsilon_i m_i,\ j = 1, \ldots, m_i] \\
&\quad \geqslant e^{-m[\Lambda^*_{D^1}(x_0) + \xi_1 \Lambda^*_{B^1}(x_1) + \cdots + \xi_i \Lambda^*_{B^i}(x_i) + \epsilon]},
\end{aligned}$$

where m is large enough, $\epsilon \to 0$ as $\epsilon_0, \epsilon_1, \ldots, \epsilon_i \to 0$, and the last inequality above is due to the SPLDP assumption in (51). As in the proof of Proposition 1 we optimize over all parameters of scenario i to obtain

$$\begin{aligned}
\liminf_{w_1 \to \infty} \frac{1}{w_1} \log \mathbf{P}[\tilde{Y}^1 = w_1] \\
\geqslant -\inf_{a>0} \frac{1}{a} \inf_{x_0 - \xi_1 x_1 - \cdots - \xi_i x_i = a\beta_{i-1}(\xi_1, \ldots, \xi_i) \in \mathcal{O}_i} \big[\Lambda^*_{D^1}(x_0) \\
+ \xi_1 \Lambda^*_{B^1}(x_1) + \cdots + \xi_i \Lambda^*_{B^i}(x_i)\big].
\end{aligned}$$

By using Lemma A.1 and selecting the tightest bound among all scenarios $1, \ldots, M$ we obtain

$$\liminf_{w_1 \to \infty} \frac{1}{w_1} \log \mathbf{P}[\tilde{Y}^1 = w_1] \geqslant -\theta^*_{G,1}.$$

ENDNOTES

1. That is, for the existence of a unique stationary distribution to which the system converges from all initial conditions.

2. In some cases in Table 3, $\mathbf{w}^*_A$ achieves less inventory cost than $\mathbf{w}^*_S$ because in these instances $\mathbf{w}^*_A$ slightly violates the service level requirements (due to the large deviations approximation).

3. Note that holding costs h_1, h_2 for echelon 1 and 2, respectively, correspond to holding costs $h_1 + h_2$ and h_2 for stage 1 and 2, respectively, in the decomposition approach.

ACKNOWLEDGMENTS

The authors thank Dimitris Bertsimas for several useful discussions. They also thank the associate editor and two anonymous referees whose comments have significantly improved the exposition and prompted them to consider the lost sales model of §6.2. Research partially supported by the NSF under a CAREER award ANI-9983221 and grants NCR-9706148 and ACI-9873339 and by the ARO under the ODDR&E MURI2001 Program Grant DAAD19-01-1-0465 to the Center for Networked Communicating Control Systems.

REFERENCES

Akella, R., P. R. Kumar. 1986. Optimal control of production rate in a failure prone manufacturing system. *IEEE Trans. Automatic Control* **AC-31** 116–126.

Asmussen, S. 1987. *Applied Probability and Queues*. Wiley, New York.

Baccelli, F., Z. Liu. 1992. On a class of stochastic recursive sequences arising in queueing theory. *Ann. Probab.* **20** 350–374.

Bertsimas, D., I. Ch. Paschalidis. 2001. Probabilistic service level guarantees in make-to-stock manufacturing systems. *Oper. Res.* **49**(1) 119–133.

——, ——, J. N. Tsitsiklis. 1998a. Asymptotic buffer overflow probabilities in multiclass multiplexers: An optimal control approach. *IEEE Trans. Auto. Control* **43**(3) 315–335.

——, ——, ——. 1998b. On the large deviations behaviour of acyclic networks of G/G/1 queues. *Ann. Appl. Probab.* **8**(4) 1027–1069.

——, ——, ——. 1999. Large deviations analysis of the generalized processor sharing policy. *Queueing Systems* **32** 319–349.

Chang, C. S. 1995. Sample path large deviations and intree networks. *Queueing Systems* **20** 7–36.

Chen, F., J.-S. Song. 2001. Optimal policies for multiechelon inventory problems with Markov-modulated demand. *Oper. Res.* **49**(2) 226–234.

Clark, A. J., H. Scarf. 1960. Optimal policies for a multi-echelon inventory problem. *Management Sci.* **6** 475–490.

Cramér, H. 1938. Sûr un nouveau théorème-limite de la théorie des probabilités. *Actualités Scientifiques et Industrielles*. 736 in *Colloque consacré à la théorie des probabilités*. Hermann, Paris, 5–23.

De Véricourt, F., F. Karaesmen, Y. Dallery. 2000. Dynamic scheduling in a make-to-stock system: A partial characterization of optimal policies. *Oper. Res.* **48**(5) 811–819.

Dembo, A., O. Zeitouni. 1998. *Large Deviations Techniques and Applications*, 2nd ed. Springer-Verlag, New York.

The Economist. 1998. A survey of manufacturing: Meet the global factory, June 20.

Evans, R. 1967. Inventory control of a multiproduct system with a limited production resource. *Naval Res. Logist.* 173–184.

Federgruen, A., P. Zipkin. 1984. Computational issues in an infinite horizon multi-echelon inventory model. *Oper. Res.* **32** 818–836.

——, ——. 1986. An inventory model with limited production capacity and uncertain demands I. The average cost criterion. *Math. Oper. Res.* **11**(2) 193–207.

Gavish, B., S. Graves. 1980. A one-product production/inventory problem under continuous review policy. *Oper. Res.* **28** 1228–1236.

Glasserman, P. 1996. Allocating production capacity among multiple products. *Oper. Res.* **44**(5) 724–734.

——. 1997. Bounds and asymptotics for planning critical safety stocks. *Oper. Res.* **45**(2) 244–257.

——, S. R. Tayur. 1994. The stability of a capacitated, multi-echelon production-inventory system under a base-stock policy. *Oper. Res.* **42**(5) 913–925.

——, ——. 1995. Sensitivity analysis for base-stock levels in multiechelon production-inventory systems. *Management Sci.* **45**(2) 263–281.

Ha, A. J. 1997. Optimal dynamic scheduling policy for a make-to-stock production system. *Oper. Res.* **45**(1) 42–53.

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

——, ——. 1999. Optimal policies and simulation based optimization for capacitated production inventory systems. S. R. Tayur, R. Ganeshan, M. Magazine, eds. *Quantitative Models for Supply Chain Management*. Kluwer, Dordrecht, The Netherlands, 7–40.

Paschalidis, I. Ch. 1999. Class-specific quality of service guarantees in multimedia communication networks. *Automatica* (special issue on control methods for communication networks) **35**(12) 1951–1968.

Peña Perez, A., P. Zipkin. 1997. Dynamic scheduling rules for a multiproduct make-to-stock queue. *Oper. Res.* **45**(6) 919–930.

Rockafellar, R. T. 1970. *Convex Analysis.* Princeton University Press, Princeton, NJ.

Sobel, M. 1982. The optimality of full-service policies. *Oper. Res.* **30** 636–649.

Veatch, M. H., L. M. Wein. 1996. Scheduling a make-to-stock queue: Index policies and hedging points. *Oper. Res.* **44**(4) 634–647.

Wein, L. M. 1992. Dynamic scheduling of a multiclass make-to-stock queue. *Oper. Res.* **40** 724–735.

Zheng, Y. S., P. Zipkin. 1990. A queueing model to analyze the value of centralized inventory information. *Oper. Res.* **38**(2) 296–307.

CONTRIBUTORS

Shabbir Ahmed ("An Approximation Scheme for Stochastic Integer Programs Arising in Capacity Expansion") is an Assistant Professor in the School of Industrial and Systems Engineering at Georgia Institute of Technology. His research interests are in the development of algorithms for stochastic programming and global optimization problems. This paper is part of his Ph.D. dissertation at the University of Illinois at Urbana–Champaign.

Cynthia Barnhart ("Improving Crew Scheduling by Incorporating Key Maintenance Routing Decisions") is a Professor of Civil and Environmental Engineering and Engineering Systems, and also the Co-Director of the Center for Transportation and Logistics at the Massachusetts Institute of Technology. Her research interests include developing linear, integer, and network optimization models and methods for large-scale transportation systems.

Fernando Bernstein ("Pricing and Replenishment Strategies in a Distribution System with Competing Retailers") is an Assistant Professor in the Operations Management area of the Fuqua School of Business at Duke University. This paper arose from his Ph.D. dissertation, which was written under the supervision of his coauthor, Awi Federgruen. The work is part of an ongoing project and a series of associated publications analyzing the behavior and efficient management of decentralized supply chains, employing a variety of game-theoretical models.

Dimitris Bertsimas ("Restaurant Revenue Management") is the Boeing Professor of Operations Research at the Sloan School of Management and the Operations Research Center at the Massachusetts Institute of Technology. His research interests include discrete, stochastic, and dynamic optimization, analysis and control of stochastic systems and applications in revenue management, finance, and e-commerce.

Amy Mainville Cohn ("Improving Crew Scheduling by Incorporating Key Maintenance Routing Decisions") is an Assistant Professor in the Industrial and Operations Engineering Department at the University of Michigan. This paper stems from her dissertation research in the Operations Research Center at MIT. She continues to work on integrated airline planning problems, as well as other large-scale applications of discrete optimization.

Robin L. Dillon ("Programmatic Risk Analysis for Critical Engineering Systems Under Tight Resource Constraints") is an Assistant Professor at the McDonough School of Business, Georgetown University. Her research interests are in decision and risk analysis and specifically the management of risk trade-offs in resource-constrained environments. This paper is the result of using programmatic risk models developed as part of her dissertation to analyze risky decisions for the unmanned space program.

Feryal Erhun ("Enterprise-Wide Optimization of Total Landed Cost at a Grocery Retailer") is an Assistant Professor in the Management Science and Engineering Department at Stanford University. Her research interests include supply-chain management and logistics, with an emphasis on modeling and analysis of tactical and strategic issues. This paper is part of her Ph.D. dissertation written under the supervision of Sridhar Tayur.

Awi Federgruen ("Pricing and Replenishment Strategies in a Distribution System with Competing Retailers") is the Charles E. Exley Professor of Management at the Decision, Risk and Operations Division of the Graduate School of Business, Columbia University. This paper arose from Fernando Bernstein's Ph.D. dissertation, which was written under Federgruen's supervision. The work is part of an ongoing project and a series of associated publications analyzing the behavior and efficient management of decentralized supply chains, employing a variety of game-theoretical models.

Seth D. Guikema ("Programmatic Risk Analysis for Critical Engineering Systems Under Tight Resource Constraints") is a Ph.D. candidate in the Department of Management Science and Engineering, Stanford University. His research interests are in probabilistic risk analysis, decision analysis, and Bayesian reliability analysis. His contributions to the paper came through assisting with the application of the method.

Peiqing Huang ("A Note on 'Inventory Models with Cost Changes' ") is a Professor in the School of Management, Shanghai Jiao Tong University, People's Republic of China. His current research interests include operations research, enterprises management and decision theory, business logistics, and supply-chain management of modern enterprises. The paper in this issue is part of an ongoing project on the integration of dynamic pricing and inventory management problems, which Jianwen Luo is completing under the supervision of Peiqing Huang.

Yong Liu ("Large Deviations-Based Asymptotics for Inventory Control in Supply Chains") is an Algorithm Design Engineer at SmartOps Corporation, Pittsburgh Pennsylvania. This research is part of the author's Ph.D. thesis, which addressed issues of resource allocation under Quality of Service requirements in stochastic networks; in particular, communication and production-inventory networks. In his current capacity, the author is seeking the application of these ideas in practice.

0030-364X/03/5103-0505 $05.00
1526-5463 electronic ISSN

Operations Research © 2003 INFORMS
Vol. 51, No. 3, May–June 2003, pp. 505–507

Jianwen Luo ("A Note on 'Inventory Models with Cost Changes' ") is an Associate Professor in the School of Management, Shanghai Jiao Tong University, People's Republic of China. His research interests include business logistics and supply-chain management, capital budgeting, risk analysis, and stochastic programming. The paper in this issue is part of the working paper of an ongoing project on the integration of dynamic pricing and inventory management problems, which Jianwen Luo is completing under the supervision of Peiqing Huang.

Philip J. Neame ("Offer Stack Optimization in Electricity Pool Markets") is a Lecturer in the Department of Mathematical Sciences, University of Technology, Sydney, Australia. This contribution was completed while he was a Postdoctoral Research Fellow at the University of Auckland and is part of ongoing research into the development of optimization models for application in electricity markets, focusing particularly on optimal behavior for generators. His other research interests include stochastic programming and integer programming.

Ioannis Ch. Paschalidis ("Large Deviations-Based Asymptotics for Inventory Control in Supply Chains") is an Associate Professor in the Department of Manufacturing Engineering, and Center for Information and Systems Engineering, Boston University. This paper is part of a larger body of work seeking to incorporate probabilistic Quality of Service constraints into inventory control decisions. The research has parallels with work of the author in quality of service provisioning for multiservice communication networks.

M. Elisabeth Paté-Cornell ("Programmatic Risk Analysis for Critical Engineering Systems Under Tight Resource Constraints") is the Professor and Chair, Department of Management Science and Engineering, Stanford University. Her main research interests are in engineering risk analysis and decision analysis. In her recent research, she has applied the risk analysis method to the management of unmanned space missions and to the threats of a terrorist attack on the United States. She is a member of the NAE and its council, the President's Foreign Intelligence Advisory Board, and the JPL Advisory Council.

Andrew B. Philpott ("Offer Stack Optimization in Electricity Pool Markets") is a Professor in the Department of Engineering Science at the University of Auckland, New Zealand. This contribution is part of ongoing research into the development of optimization models for application in electricity markets, focusing particularly on optimal behavior for generators. His other research interests include stochastic programming and yacht optimization models.

Erica L. Plambeck ("Incentive Efficient Control of a Make-to-Stock Production System") is an Assistant Professor of Operations, Information and Technology at the Graduate School of Business, Stanford University. She is interested in optimal control of assemble-to-order systems, dynamic incentive problems, and relational contracts (informal agreements) in operations and supply-chain management. This paper is a chapter from her dissertation written under the supervision of Stefanos Zenios. An earlier version of the paper won first prize in the Nicholson Student Paper Competition in 2000.

Geoffrey Pritchard ("Offer Stack Optimization in Electricity Pool Markets") is a Senior Lecturer in the Department of Statistics at the University of Auckland, New Zealand. This contribution is part of ongoing research into the development of optimization models for application in electricity markets, focusing particularly on optimal behavior for generators. His interests include Markov chains and stochastic optimization, particularly as applied to electricity markets.

Nikolaos V. Sahinidis ("An Approximation Scheme for Stochastic Integer Programs Arising in Capacity Expansion") is a Professor of Chemical and Biomolecular Engineering at the University of Illinois at Urbana–Champaign, where he is also an affiliate faculty of Industrial Engineering. His research interests are at the interface between computer science and operations research, with applications in chemical, biological, medical, and engineering systems. This paper is part of a continuing line of work on computational complexity and approximation algorithms for problems in process systems engineering.

Serpil Sayin ("A Procedure to Find Discrete Representations of the Efficient Set with Specified Coverage Errors") is an Associate Professor at Koc University. Multiple Criteria Optimization (MCO) has been her primary area of research since her dissertation work. The difficulty of dealing with the solution set of continuous MCO problems has motivated her to focus on the "discrete representations" concept, which also constitutes the core idea of the current paper.

Romy Shioda ("Restaurant Revenue Management") received her B.S. and M.S. degrees from Massachusetts Institute of Technology in 1999 and 2002, respectively. She will finish her Ph.D., under the supervision of Dimitris Bertsimas, at the Operations Research Center of MIT in June 2003. Her research interests include revenue management, computational integer optimization, and data mining.

Hartmut Stadtler ("Multilevel Lot Sizing with Setup Times and Multiple Constrained Resources: Internally Rolling Schedules with Lot-Sizing Windows") is a Professor in the Institute of Business Administration at Darmstadt University of Technology. His current research interests are in supply-chain management, especially in advanced planning systems, their principles (hierarchical planning), their modules (master planning), and solution techniques (time decomposition combined with a standard MIP solver, which is the topic of this paper).

Sridhar Tayur ("Enterprise-Wide Optimization of Total Landed Cost at a Grocery Retailer") is a Professor at