Proceedings of the 40th IEEE
Conference on Decision and Control
Orlando, Florida USA, December 2001

Threshold-Based Control for Make-to-Stock Models: A Synergy between Large Deviations and Perturbation Analysis[1]

Ioannis Ch. Paschalidis[2] Yong Liu[3] Christos G. Cassandras[4] Ping Zhang[5]

Abstract

We consider a model of a make-to-stock manufacturing system. External demand is met from the finished goods inventory; unsatisfied demand is backlogged. We adopt a base-stock production policy which produces if inventory falls below a certain threshold and idles otherwise. We set this threshold to guarantee stockout or delay probabilities to stay below given constants (*service level constraints*). These can be set by solving a utility maximization problem which trades-off Quality of Service with expected inventory costs. We combine analytical (large deviations) and simulation-based (perturbation analysis) techniques. We demonstrate that there is a natural synergy between these two approaches.

Keywords: *Make-to-stock systems, Service levels, Large Deviations, Perturbation Analysis.*

1 Introduction

In this paper we will focus on a make-to-stock manufacturing system. In such systems, demand is met from a *finished goods inventory (FGI)* and the production facility strives to maintain this inventory nonempty to avoid stockouts, which lead either to backordered demand or lost sales. A large variety of products are manufactured in this fashion. In these systems the fundamental trade-off is between *producing*, which accumulates inventory and incurs inventory costs, and *idling*, which leads to stockouts and unsatisfied demand. The objective is to devise a production policy optimizing some measure of the system's performance, which should incorporate both *inventory costs* and *backorder costs*, i.e., costs associated with not being able to timely meet demand.

In modern manufacturing, *Quality of Service (QoS)* is gaining significance in acquiring and maintaining market share. To that end, we introduce constraints that ensure that probabilities of stockout events or delays stay bounded below given desirable levels. We believe that such *service-level* constraints provide a more natural representation of customer satisfaction than expected backorder costs. The latter are the norm in the literature but are hard to quantify.

The make-to-stock problem we are considering has been studied extensively in the literature (see [1] and references therein). In a variety of settings (e.g., [2]) it has been established that a *base-stock policy* (produce when inventory falls below a certain threshold and idle otherwise) is optimal. The multiclass version of the problem is more involved; apart from idling, a production policy consists of scheduling decisions as well. Combining fluid and large deviations techniques the multiclass problem was analyzed in [3].

In this paper we will combine Large Deviations (LD) (employed in [3]) and Perturbation Analysis (PA) techniques [4]. We will demonstrate that there is a natural synergy between these rather distinct approaches.

LD analysis is an *off-line* approach, based on asymptotic results, intended to evaluate performance measures of interest that involve "rare events". A case in point arises with stockout probabilities which should be relatively small. One can obtain asymptotically tight approximations of such probabilities (as they become small) for a variety of models, including models that capture dependencies in the demand and production processes. LD analysis is computationally fast, characterizes the most likely way that stockouts occur, and can be used to answer key "what-if" questions. On the other hand, it does require detailed statistical characterization of the stochastic processes involved. PA is an *on-line* approach intended to estimate performance measures by observing an actual (or simulated) system sample path. It estimates performance over multiple parameter settings from a single sample path. In contrast to the LD approach, it requires data, the collection of which may be time-consuming (e.g., estimating small probabilities requires long sample paths). On the other hand, PA does not rely on detailed knowledge of

[1]Research partially supported by the NSF under under a CAREER award ANI-9983221 and grants NCR-9706148, ACI-9873339, and EEC-0088073, by the ARO under the ODDR&E MURI2001 Program Grant DAAD19-01-1-0465 to the Center for Networked Communicating Control Systems, by the AFOSR under grant F49620-01-0056, by the AFRL under contract F30603-99-C-0057, and by the EPRI/DOD under contract WO8333-03.

[2]Corresponding author. Department of Manufacturing Engineering, Boston University, Boston, MA 02215, e-mail: yannisp@bu.edu, url: http://ionia.bu.edu/.

[3]Department of Manufacturing Engineering, Boston University, Boston, MA 02215, e-mail: liuyong@bu.edu.

[4]Department of Manufacturing Engineering, Boston University, Boston, MA 02215, e-mail: cgc@bu.edu, url: http://vita.bu.edu/cgc.

[5]Department of Manufacturing Engineering, Boston University, Boston, MA 02215, e-mail: pzhang@bu.edu

the model statistics.

We believe that there is a natural synergy between PA and LD, arising in several domains:

- PA provides accurate estimates for relatively large stockout probabilities (e.g., 5%) within reasonable time, and LD approximations become very reliable for relatively small ones (e.g., 10^{-4}). Hence, combining LD and PA one can accurately estimate a large range of such probabilities.
- LD analysis can be used to quickly obtain a control policy based on some initial estimates on demand and production processes and then rely on PA to fine-tune the control policy based on the actual realization of these processes.
- PA can also be of direct use in the LD analysis. As we will see in the sequel, we use the expected inventory position to fine-tune our large deviations asymptotics. PA can be naturally used to provide this expectation.

Our analysis is general enough to handle rather sophisticated models of demand and production that can represent inherent dependencies in these processes. This enables us to model realistic demand scenarios and *failure-prone* manufacturing facilities. To that end, we will allow demand and production to be modeled by *autocorrelated* stochastic processes.

Our motivation is to develop this line of analysis to handle the more general supply chain case (multiple stages). In this conference paper, however, we chose to focus on the simpler, yet nontrivial, single stage case to demonstrate our key ideas on the synergy between LD and PA techniques.

The remainder of this paper is organized as follows: Section 2 introduces the system we will study using both LD analysis (Section 3) and PA (Section 4). In Section 5, we present numerical results using both methods. Conclusions are in Section 6.

2 The Model

We consider the make-to-stock manufacturing system of Figure 1. Demand is met from the *finished goods inventory (FGI)*; unsatisfied demand is backordered. We assume a discrete-time model, where time is slotted and the state of the system is examined at the beginning of each time slot $n \in \mathbb{Z}$ (periodic review policy). Let D_n denote the demand arriving during time slot n, and B_n denote the amount of goods that can be produced (capacity) during the same time slot. Let also $X_n \in \mathbb{R}$ denote the available inventory at the beginning of time slot n. We allow X_n to take negative values; when nonnegative it is equal to the amount of available inventory, and when negative it is equal to the amount of backordered demand. We will be measuring D_n, B_n, and X_n in units of work content of a reference machine (identical to ours) which produces goods at a given constant reference rate.

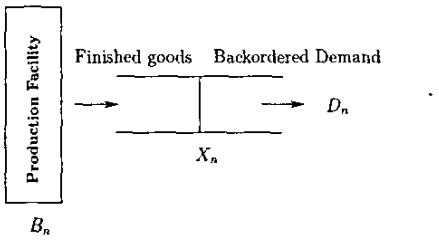

Figure 1: The model of a make-to-stock system.

We assume that the demand and service processes $\{D_n, B_n;\ n \in \mathbb{Z}\}$ are arbitrary, stationary, and mutually independent stochastic processes, that satisfy certain mild technical conditions (a large deviations principle, see [3] for details). These assumptions are satisfied by a fairly large class of stochastic processes, which includes renewal processes, Markov-modulated processes, and general stationary processes with mild mixing conditions. For stability purposes, we further assume that

$$\mathbf{E}[B_1] > \mathbf{E}[D_1], \tag{1}$$

which by stationarity carries over to all time slots n.

We will implement a *base-stock* policy which maintains a safety stock or hedging point of w: the system produces when the inventory is below w and idles otherwise. We have

$$X_{n+1} = \min\{X_n - D_n + B_n, w\}. \tag{2}$$

We quantify customer dissatisfaction by the probability, $\mathbf{P}[X_n \leq 0]$, of not being able to meet incoming demand immediately (*stockout probability*). Alternatively, we can quantify customer dissatisfaction by the probability of exceeding a promised delivery time (*delay probability*). Let us denote by $\mathbf{P}$[customer dissatisfaction] either the stockout, or the delay probability. Let us now assume the existence of a "utility function" $U(x, y)$ that quantifies the desirability of a given QoS level x, and a given level y of inventory costs. Let finally, ϵ be a desirable upper bound on the QoS level. We are interested in selecting a hedging point w that solves the following optimization problem:

$$\begin{aligned} \max \quad & U(\mathbf{P}[\text{customer dissatisfaction}], \text{Cost}) && (3)\\ \text{s.t.} \quad & \mathbf{P}[\text{customer dissatisfaction}] < \epsilon, && (4)\\ & \text{Cost} = h\mathbf{E}[X_n^+], && (5) \end{aligned}$$

where h is a given scalar and X_n^+ denotes $\max\{X_n, 0\}$. We will be referring to (4) as the *service-level* constraint. To achieve our goal we need to compute $\mathbf{P}$[customer dissatisfaction]. An exact analytic expression is impossible to obtain, especially in view of the complicated, autocorrelated, models for the demand and production processes. To that end, we will resort either to simulation or to asymptotic techniques.

3 Large Deviations Analysis

Given a process $\{X_i\}$, where X_i, $i \geq 1$, are identically distributed, possibly interdependent, random variables, let $S_n \stackrel{\Delta}{=} \sum_{i=1}^{n} X_i$, and $\Lambda(\theta) \stackrel{\Delta}{=} \lim_{n\to\infty} \frac{1}{n} \log \mathbf{E}[e^{\theta S_n}]$. We will refer to $\Lambda(\cdot)$ as the *limiting* log-*moment generating function*. In the sequel, we will be 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 $\{X_i\}$ (see [3] for some background in large deviations).

We start by an asymptotic large deviations analysis. Define the shortfall L_n, during time slot n, as:

$$L_n \stackrel{\Delta}{=} w - X_n.$$

Eq. (2) becomes

$$L_{n+1} = \max\{L_n + D_n - B_n, 0\}. \tag{6}$$

We can interpret L_n as the queue length of a discrete-time G/G/1 queue with D_n arrivals and at most B_n departures during time slot n. We will refer to this queue as the make-to-order system corresponding to the make-to-stock system we are studying. Using this equivalence, we have $\mathbf{P}[X_n \leq 0] = \mathbf{P}[L_n \geq w]$. Under general assumptions on the demand processes and production processes, we have the following result ([3]).

Proposition 3.1 *The steady-state queue length process L_n satisfies*

$$\lim_{w\to\infty} \frac{1}{w} \log \mathbf{P}[L_n \geq w] = -\theta^*, \tag{7}$$

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

Intuitively, for large enough w we have

$$\mathbf{P}[X_n \leq 0] = \mathbf{P}[L_n \geq w] \sim e^{-w\theta^*}. \tag{8}$$

Thus, the minimum w that guarantees $\mathbf{P}[X_n \leq 0]$ to be below ϵ is $w = -\log(\epsilon)/\theta^*$. Notice that $\Lambda_D(\theta) + \Lambda_B(-\theta)$ is zero at the origin and has negative derivative at the same point (cf. (1)). When $\Lambda_D(\theta) + \Lambda_B(-\theta) < 0$ for all $\theta > 0$ we will say that $\theta^* = \infty$. In this case, no stockouts occur and a safety stock of zero should be maintained (*Just in Time (JIT)* policy).

We can improve the accuracy of (8) by introducing a constant in front of the exponential, i.e.,

$$\mathbf{P}[X_n \leq 0] \approx \alpha e^{-w\theta^*}. \tag{9}$$

Thus, the hedging point satisfies

$$w = -\frac{\log(\epsilon/\alpha)}{\theta^*}. \tag{10}$$

Using an idea from [3] we obtain

$$\alpha = \theta^* \mathbf{E}[L_n]. \tag{11}$$

Thus, to find the asymptotic constant we need (apart from θ^*) the expectation of the queue length process L_n in a G/G/1 queue, which is independent of w and can be obtained either by analytical approximations or by simulation (see [3]). In particular, $\mathbf{E}[L_n]$ can be obtained as a byproduct of the PA work (discussed in Section (4)). A key point here is that large deviations analysis is used to determine the exponent, and simulation might be used only to estimate $\mathbf{E}[L_n]$. This is beneficial since it is much easier to obtain a reliable estimate for $\mathbf{E}[L_n]$ than one for $\mathbf{P}[X_n \leq 0]$.

We next turn our attention to the delay probability. Consider an order arriving at time slot n. We will call *delay*, and denote by d_n, the time that it takes until the order starts to get filled. As with inventory, we will measure delay in units of work content, that is, time units required by the reference machine to produce it. Thus, when $X_n \geq 0$, the incoming order can start to get filled immediately, which by our definition implies $d_n = 0$. When, however, $X_n < 0$, the machine has a backlog of $-X_n$ units of work content, which implies $d_n = -X_n$. Using the correspondence with the make-to-order system, we obtain

$$d_n = (L_n - w)^+ = \max\{L_n - w, 0\}. \tag{12}$$

We are interested in obtaining an expression for $\mathbf{P}[d_n \geq r]$, at an arbitrary time slot n, where r is a positive scalar. Note that

$$\mathbf{P}[d_n \geq r] = \mathbf{P}[\max\{L_n - w, 0\} \geq r] = \mathbf{P}[L_n \geq w + r].$$

Using the result of Proposition 3.1 we establish

Proposition 3.2 *For any $r > 0$, the steady-state delay probability satisfies*

$$\lim_{w\to\infty} \frac{1}{w + r} \log \mathbf{P}[d_n \geq r] = -\theta^*.$$

Intuitively, for large values of w,

$$\mathbf{P}[d_n \geq r] \sim e^{-\theta^*(w+r)}.$$

A refined estimate can be obtained using (9), i.e.,

$$\mathbf{P}[d_n \geq r] \approx \alpha e^{-\theta^*(w+r)}, \tag{13}$$

where α is given by (11).

We finally consider the inventory cost. Let $C(w)$ be the expected inventory cost, when we fix the hedging point to w. As in (5), $C(w) = h\mathbf{E}[X_n^+]$, where h is a given constant. $C(w)$ can be approximated by (the proof is omitted due to space limitations)

$$C(w) \approx h(w - \mathbf{E}[L_n] + \mathbf{E}[L_n]e^{-w\theta^*}). \tag{14}$$

4 Perturbation Analysis

Whereas the LD analysis of Section 3 was based on a *time-driven* model for the inventory X_n (or shortfall L_n), for the purpose of Perturbation Analysis (PA) an

event-driven model is more convenient. As we will see, the two models are equivalent. There are two event processes whose evolution affects the inventory level. First, there is an exogenous demand process $\{R_k; k \in \mathbb{Z}^+\}$ where $R_k \in \mathbb{R}^+$ denotes the time when the kth demand request is made (it is possible that an entire order for N parts is placed at one time, i.e., $R_k = R_{k+1} = \ldots = R_{k+N-1}$). Let $\{\rho_k\}$ denote an arbitrary inter-request time stochastic process ($\rho_k \in \mathbb{R}^+$). In addition, there is a production process represented by $\{C_k; k \in \mathbb{Z}^+\}$, where C_k denotes the time when the kth part completes processing and is added to the *FGI*. Let $\{\pi_k\}$ denote an arbitrary processing time stochastic process ($\pi_k \in \mathbb{R}^+$).

As before, we shall use a hedging point to control production. However, whereas in the LD analysis the hedging point is *real-valued* because of its interpretation as a safety stock measured in time units, in an event-driven model it is *integer-valued*: it is a counter of finished parts in the FGI. With this observation in mind, we shall retain the same symbol w to denote the hedging point. Note that C_k satisfies

$$C_k = \max\{C_{k-1}, R_{k-w}\} + \pi_k, \tag{15}$$

where $R_{k-w} = 0$ for all k such that $k - w \leq 0$. This model is virtually the same as one used to represent kanban-based manufacturing systems with w corresponding to the number of kanban assigned to the workcenter modeled through (15), as in [5].

In this model, a "stockout event" is one such that $R_k < C_k$. Thus, if a sample path under w is available, an estimate of the stockout probability, denoted by $\mathbf{P}[w]$, after N demand requests are observed is given by

$$\hat{\mathbf{P}}[w] = \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}\{R_k < C_k\}. \tag{16}$$

Next we turn our attention to the delay probability. Recalling the definition of delay, we obtain

$$d_k = (C_k - R_k)^+ = \max\{C_k - R_k, 0\}. \tag{17}$$

Thus, if a sample path under w is available, an estimate of the delay probability, denoted by $\mathbf{P}[d_k \geq r]$, after N demand requests are observed is given by

$$\hat{\mathbf{P}}[d_k \geq r] = \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}\{C_k - R_k \geq r\}. \tag{18}$$

The expected delay $\mathbf{E}[d_k]$ can be estimated as follows:

$$\hat{\mathbf{E}}[d_k] = \frac{1}{N}\sum_{k=1}^{N} d_k. \tag{19}$$

Finally, we can also estimate the expected inventory:

$$\hat{\mathbf{E}}[X_k^+] = \frac{1}{T_N}\sum_{k=0}^{N} X_k^+ \cdot (t_k - t_{k-1}), \tag{20}$$

where T_N is the length of the observed sample path and t_k is occurrence time of an event (production or demand request) which changes X_k. Thus, if a sample path under w is available, an estimate of inventory cost $C(w)$ is given by

$$\hat{C}(w) = h\hat{\mathbf{E}}[X_k^+]. \tag{21}$$

4.1 Completion Time Perturbation Dynamics

For our purposes, any sample path observed under a hedging point w is referred to as the *nominal* sample path. Then, a *perturbed* sample path is one that would have resulted if the exact same nominal one had been reproduced under a different hedging point $\tilde{w} \neq w$. In the sequel, we will use a tilde ($\tilde{}$) for all perturbed variables. We define completion time perturbations as follows:

$$\Delta C_k(\tilde{w}) = \tilde{C}_k - C_k, \quad k = 1, 2, \ldots, \tag{22}$$

and $\Delta C_0(\tilde{w}) = 0$. The purpose of PA is to derive a recursive relationship for $\Delta C_k(\tilde{w})$, $k = 1, 2, \ldots$, which involves only quantities directly observable along the nominal sample path. If this is possible, then at each completion time C_k one can evaluate $\Delta C_k(\tilde{w})$ as well for as many values of $\tilde{w}$ as desired. In the sequel, we shall only write ΔC_k to denote perturbed completion times corresponding to a given $\tilde{w}$. Let us also define $I_k = R_{k-w} - C_{k-1}$, and observe that if $I_k > 0$, it represents the length of an idle period starting with the completion of the $(k-1)$th part and ending with the $(k-w)$th demand request. Furthermore, set $Q_k = R_{k-\tilde{w}} - R_{k-w}$, which represents the time difference in ending an idle period (positive or negative) due to the change in hedging point. The following proposition characterizes ΔC_k; the proof is omitted due to page limitations.

Proposition 4.1 *For any perturbed value $\tilde{w}$, the completion time perturbation is given by ($k = 1, 2, \ldots$)*

$$\Delta C_k = \max\{\Delta C_{k-1}, I_k + Q_k\} - \max\{I_k, 0\}. \tag{23}$$

In order to use (23) on-line, we must ensure that at time C_k, I_k and Q_k are indeed available, i.e., we need to ensure that $C_k \geq \max\{R_{k-w}, R_{k-\tilde{w}}, C_{k-1}\}$. It is obvious from (15) that $C_k > \max\{C_{k-1}, R_{k-w}\}$. Moreover, if $\tilde{w} > w$, then $R_{k-w} \geq R_{k-\tilde{w}}$.

In a simulation environment, we conclude that it is always advantageous to simulate the system at the smallest feasible value of w and apply PA to infer the effect of all desirable $\tilde{w} > w$.

The PA algorithm operates by updating $\Delta C_k(\tilde{w})$ over a given set of values for $\tilde{w}$, after observing the kth completion event at time C_k along an observed sample path under a given hedging point w. If we are interested in estimating $\mathbf{P}[w]$ and $\mathbf{P}[\tilde{w}]$, this is accomplished using (16). If, in addition, our objective is to determine the smallest hedging point that provides a stockout probability below a given threshold ϵ, this is done by simply comparing all stockout probability estimates to deter-

mine the smallest $\tilde{w}$ satisfying $\hat{\mathbf{P}}(\tilde{w}) < \epsilon$. In so doing, we are exploiting the fact that $\mathbf{P}[w]$ is monotonically nonincreasing in w.

Algorithm:

1. Initialize: $N_0(w) = 0$, $\Delta C_0(\tilde{w}) = N_0(\tilde{w}) = 0, \forall \tilde{w}$.

2. Whenever a completion event occurs at time C_k:

   2a. Evaluate $\Delta C_k(\tilde{w})$ for all $\tilde{w}$ using (23).

   2b. If $C_k > R_k$, increase the number of nominal sample path stockout events: $N_k(w) = N_{k-1}(w) + 1$.

   2c. If $C_k + \Delta C_k(\tilde{w}) > R_k$, increase the number of perturbed sample path stockout events: $N_k(\tilde{w}) = N_{k-1}(\tilde{w}) + 1$.

3. After N demand request events are observed, STOP and estimate the optimal hedging point:

   3a. For every $\tilde{w}$ (including $\tilde{w} = w$), estimate the stockout probability: $\hat{\mathbf{P}}(\tilde{w}) = \frac{N_N(\tilde{w})}{N}$.

   3b. Estimate the optimal hedging point: $w^* = \arg\min_{\tilde{w}}\{\hat{\mathbf{P}}(\tilde{w}) < \epsilon\}$.

Clearly, estimates of the delay probability, expected delay, and expected inventory can also be obtained under Step 3a above by using (18), (19), and (20) respectively.

In a simulation environment, the advantage of the algorithm is that it can provide performance metric estimates over multiple hedging points from a single simulation run at some selected hedging point values. Typical numerical results we have obtained show a "speedup" realized by the PA algorithm of about 5.5, while the overhead imposed by the algorithm was approximately 2.5%.

The PA algorithm described above can be adapted to a discrete-time model conforming to the one presented for the LD analysis. This will allow us to use both LD and PA-based hedging point determination on a common model for comparison purposes and for achieving the synergy that has motivated this work. We assume a periodic review policy where time is divided into time slots of equal duration and the system is examined at the beginning of each time slot. Let N_{π_i} be the number of production events and N_{ρ_i} the number of demand requests observed during the ith time slot. The demand process $\{N_{\rho_i}, i \in Z^+\}$ and the production process $\{N_{\pi_i}, i \in Z^+\}$ are arbitrary stochastic processes. Our objective is to recover π_k and ρ_k (needed to drive the event-driven model based on which PA was carried out) from $\{N_{\pi_i}\}$ and $\{N_{\rho_i}\}$. Once we recover these integer values of π_k and ρ_k (details are omitted), we can return to the model (15) and the PA algorithm obtained through (23), the only difference being that all quantities involved are now integer-valued. The optimal hedging point obtained in this fashion is based on a discrete-time model of the same underlying system.

5 Numerical Results

We next provide numerical results to demonstrate the accuracy of the LD asymptotics and compare them to those obtained by PA. We consider only stockout probabilities; the conclusions for the delay probability would not be much different. The demand and production processes are discrete-time Markov modulated processes (see Fig. 2). By $\mathbf{r}$ we denote the vector of demand or production amounts at each state of the corresponding Markov chain. The load of the system is nearly 0.8.

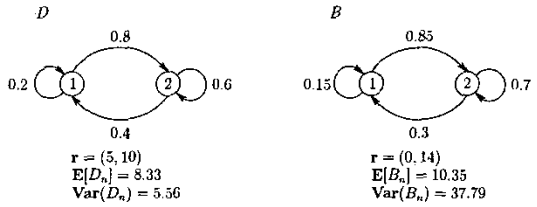

Figure 2: The demand and production processes.

5.1 Performance Evaluation

For the example of Fig. 2, we apply the PA algorithm. For each service level requirement ϵ, the optimal hedging point from PA is the average over 20 independent sample paths of $N = 10^7$ demand request events.

The LD analysis yields a decay rate for the stockout probability $\theta^* = 0.120$. As discussed in Section 3, to refine the LD asymptotic (cf. 9) we use PA to obtain $\mathbf{E}[L_n]$ in the equivalent G/G/1 queue. The hedging point (10) becomes

$$w^* = -\frac{\log(\epsilon/0.768)}{0.120}.$$

Table 1 compares the LD and PA results. Both approaches are very accurate and the results are very close for most ϵ's. It can be seen that when ϵ is extremely small, PA requires longer sample paths to get a reliable estimate, while the LD approximation is still reliable. The refined LD approximation is accurate even for large ϵ's.

	LD	PA	Simulation	
ϵ	w	w	$\tilde{w}$	$\mathbf{P}[X_n \leq 0]$
0.3	7.84	7.60	8	0.326
0.2	11.22	11.00	11	0.194
1.0×10^{-1}	16.99	17.05	17	0.996×10^{-1}
1.0×10^{-2}	36.18	36.10	36	1.040×10^{-2}
1.0×10^{-3}	55.37	54.85	55	1.072×10^{-3}
1.0×10^{-4}	74.56	73.35	75	0.964×10^{-4}
1.0×10^{-5}	93.74	95.50	94	0.983×10^{-5}
1.0×10^{-6}	112.93	106.06	113	1.059×10^{-6}

Table 1: Comparing LD with PA results.

5.2 Robustness Analysis

In general, statistical models of demand and production might not be known a priori, or even if the model structure is known, certain parameters have to be estimated. Estimation leads to errors, which lead to inaccuracies in the LD analysis. We next consider the example of Figure 2 and introduce disturbances in the parameters to test the robustness of the LD results. In certain cases, the discrepancies are substantial which demonstrates the utility of combining LD and PA techniques as outlined in the introduction.

Constant disturbances in the transition probabilities: Suppose there is a disturbance Δ in the transition probability matrix of the underlying Markov chain characterizing the demand process, e.g., consider a matrix $\mathbf{P}_D$ where

$$\mathbf{P}_D = \begin{bmatrix} 0.2 - \Delta & 0.8 + \Delta \\ 0.4 & 0.6 \end{bmatrix},$$

and $\Delta \in [-0.2, 0.2]$. We use PA to estimate optimal hedging points w' for the disturbed system and compare it with the w given by LD for the original system. We found that for $-0.2 \leq \Delta \leq 0.2$, almost all relative errors $|\frac{w'-w}{w}| \cdot 100\%$ are to be within 10%, yet a Δ equal to 0.2 causes a large relative change. On closer inspection, "robustness" is more a property of the system itself rather than the method of determining w, i.e., w' does not change much as we vary Δ. Thus, the stockout probability is not too sensitive in Δ.

Constant disturbance in the demand: Consider next a disturbance in the amount of demand at each state of the underlying Markov chain. In particular, we consider a demand process with $\mathbf{r}_D = (5 + \Delta, 10)$, where $\Delta \in \{-5, -4, \ldots, 4, 5\}$. We observed that when we increase the demand, i.e., Δ is positive, the relative error increases very steeply. In contrast, negative disturbances do not affect the analytically (LD-based) computed w value as much. This can be attributed to the inherent nonlinearity in the system: as the utilization increases the effect on the stockout probability is more dramatic.

Random disturbance in the demand: Finally suppose that the disturbance Δ in $\mathbf{r}_D = (5 + \Delta, 10)$ is a zero mean random variable. This is for example the case when the amount of demand at state 1 is not known and we are estimating it from real data to apply the LD approach. Table 2 gives simulation (PA) results for uniformly distributed disturbances in $[-5, 5]$ and disturbances with a $(0, 1)$ Gaussian distribution. These results show that if our model is accurate, in the sense that the mean of the disturbance is zero, the LD approach provides fairly reliable approximations. Nevertheless, errors of about 5% that appear when Δ is uniformly distributed might be significant in certain circumstances.

Original System			$\Delta \sim U(-5,5)$		$\Delta \sim N(0,1)$	
ϵ	w	w'	w'	$\frac{\lvert w'-w \rvert}{w}$	w'	$\frac{\lvert w'-w \rvert}{w}$
0.1	17.0	17.1	18.0	5.94%	17.0	0.06%
0.01	36.2	36.0	37.7	4.20%	36.1	0.24%
0.001	55.4	54.9	57.6	4.02%	55.4	0.05%

Table 2: Random disturbances in the demand.

6 Conclusions

We considered the model of a make-to-stock manufacturing system under rather complicated, potentially autocorrelated, models of demand and production. We adopted a production policy that sets a certain threshold for inventory (hedging point); the system produces when inventory falls below this threshold and idles otherwise. We analyzed the system using both large deviations (LD) and perturbation analysis (PA) techniques. LD techniques provide asymptotically tight approximations for the stockout or delay probability and allow us to set the hedging point to guarantee desirable service levels. PA techniques find this hedging point efficiently from simulation. We demonstrated that there is a natural synergy between LD and PA. In particular, (i) PA is very reliable and accurate for relatively large stockout probabilities and becomes computational expensive for very small stockout probabilities, while LD becomes more accurate for small stockout probabilities, (ii) PA can be used to refine the LD asymptotics and make them reliable in a wide range of service levels, and (iii) LD can initialize PA and lead to faster convergence in situations where detailed statistical models of demand and production are either imprecisely known or even unknown.

References

[1] R. Kapuscinski and S. Tayur, "Optimal policies and simulation based optimization for capacitated production inventory systems," in *Quantitative Models for Supply Chain Management* (S. Tayur, R. Ganeshan, and M. Magazine, eds.), ch. 2, pp. 7–40, Kluwer, 1999.

[2] A. Federgruen and P. Zipkin, "An inventory model with limited production capacity and uncertain demands I. The average cost criterion," *Mathematics of Operations Research*, vol. 11, no. 2, pp. 193–207, 1986.

[3] D. Bertsimas and I. C. Paschalidis, "Probabilistic service level guarantees in make-to-stock manufacturing systems," *Operations Research*, vol. 49, no. 1, pp. 119–133, 2001.

[4] C. G. Cassandras, *Discrete Event Systems: Modeling and Performance Analysis*. Irwin Publ., 1993.

[5] C. G. Panayiotou and C. G. Cassandras, "Optimization of kanban-based manufacturing systems," *Automatica*, vol. 35, pp. 1521–1533, September 1999.