Optimal Control of a High-Volume Assemble-to-Order System

Erica L. Plambeck
Graduate School of Business, Stanford University, Stanford, California 94305, plambeck_erica@gsb.stanford.edu

Amy R. Ward
Department of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, amy@isye.gatech.edu

We consider an assemble-to-order system with a high volume of prospective customers arriving per unit time. Our objective is to maximize expected infinite-horizon discounted profit by choosing product prices, component production capacities, and a dynamic policy for sequencing customer orders for assembly. We prove that a myopic discrete-review sequencing policy, which allocates scarce components among orders for different products to minimize instantaneous physical and financial holding costs, is asymptotically optimal. Furthermore, we prove that optimal prices and production capacity nearly balance the supply and demand for components (i.e., it is economically optimal to operate the system in heavy traffic), so system performance is characterized by a diffusion approximation. The diffusion approximation exhibits state-space collapse: Its dimension equals the number of components (rather than the number of components plus the number of products). These results complement the existing assemble-to-order literature, which focuses on managing component inventory and assumes FIFO sequencing of orders for assembly.

Key words: assemble-to-order systems; functional limit theorems; diffusion limits; Brownian motion; state-space collapse; discrete-review policy; control policy

MSC2000 subject classification: Primary: 90B05, 60F17; secondary: 90B22, 90B35, 60J70

OR/MS subject classification: Primary: queues, diffusion models, limit theorems, inventory/production, policies, review/lead times; secondary: probability, stochastic model applications


1. Introduction. Consumers are heterogeneous, and demand a variety of products. Manufacturers have difficulty in forecasting consumer choice, and therefore incur high costs in carrying finished-goods inventory in the wrong amounts, for the wrong products. For modular products like computers, an efficient alternative is to assemble to order; i.e., to hold inventories of components that can be rapidly assembled into a wide variety of finished products in response to customer orders.

As the Internet and new information technologies have enabled manufacturers to sell directly to customers rather than through retail outlets, assemble-to-order manufacturing has become increasingly prevalent. Dell, a leader in direct sales and assemble-to-order (ATO) manufacturing, has grown by 40% per year in recent years, although the PC industry as a whole has grown by less than 20% per year (Economist [14]). In addition to Dell, companies as diverse as General Electric, American Standard (Bylinsky [10]), BMW (Economist [14]), Toyota (Economist [15], Forbes [17]), Timbuk2 (Plambeck [30]), and National Bicycle (Agrawal and Cohen [2]) either have adopted or are considering adopting an ATO approach.

In an assemble-to-order system, pricing, capacity management, and dynamic execution are very challenging. To maximize profit, the first-order challenge is to choose product prices (which determine revenue and the demand for each component) and to choose the production capacity (or supply contract) for each component. However, due to stochastic fluctuations in supply and demand, component shortages will sometimes occur. Then, the manufacturer must dynamically ration scarce components among customer orders for various products, and/or pay to expedite component production. In practice, most firms adopt simple static rules to sequence customer orders for assembly, such as FIFO or proportional allocation (Agrawal and Cohen [2]) and expedite production on an ad hoc basis (Perman [29]). Even Dell has only recently begun to sequence orders for assembly dynamically, using real-time information about component availability (Perman [29]). Dynamic control is complex because Dell offers thousands of different products (computer configurations) from hundreds of different components. Therefore, theory is needed to guide business practice.

In this paper and its sequel, we undertake a holistic analysis of ATO manufacturing. Specifically, we show how to set product prices and component production capacity, and then dynamically sequence customer orders for assembly and manage component inventory, in order to maximize expected infinite-horizon discounted profit. In this paper, the component production capacity equals the component production rate. However, in our sequel paper (Plambeck and Ward [31]) we allow for expediting extra components and salvaging excess components, so that the effective production rate of components differs from the component production capacity.

Our analysis begins with the deterministic analog of the system, ignoring stochastic variability in supply and demand for components. We solve a static planning problem: Set product prices and component production
capacities to maximize profit in the deterministic system. At the optimal solution to the static planning problem, the long-run average production rate for each component is equal to the long-run average rate of demand. Hence, with stochastic variability, customers will experience delays and components will be held in inventory. Discounted expected profit is reduced from the optimal objective value in the static planning problem because components are purchased before they are needed, and customers will not pay until their products are assembled. However, assuming a high demand rate, we prove that the optimal prices and component production capacities are close to the solution of the static planning problem. That is, heavy traffic is the optimal operating regime.

Then, we can employ a diffusion approximation to the dynamic control problem. First, we characterize a simple, near-optimal rule for allocating scarce components to outstanding customer orders (dynamic sequencing). Next, we slightly perturb the static planning problem solution to increase expected infinite-horizon discounted profit.

The key assumption in our analysis is that a high volume of potential customers arrives per unit of time. Under this assumption, because the optimal production rate for each component approximately equals the optimal demand rate, the functional central limit theorem dictates that the inventory position of each component (the number in stock minus the number required to assemble outstanding orders) changes gradually. However, the system manager can quickly adjust the customer order queue lengths. Consider Dell computer as a concrete example. Its OptiPlex assembly plant assembles more than 20,000 computers per day (an order for a computer arrives every 4.3 seconds), but the queue for a specific configuration of computer will typically be less than 100. By prioritizing that specific configuration for assembly, Dell could eliminate the queue in minutes. However, queues for other products would increase.

We propose to review the system and release orders for assembly at short intervals of time. If one or more components are in shortage, products are prioritized for assembly according to price and component requirements, so that the resulting configuration of queues minimizes the instantaneous financial “holding cost” for delaying assembly (and hence payment). Because the inventory position changes slowly, queue lengths will almost continuously track the minimum cost configuration. This discrete-review approach to dynamic control follows that of Harrison [21] and Maglaras [24, 25], except that the decision at each review time point pairs available components with outstanding orders instead of allocating server time amongst customer classes.

The following summarizes the main contributions of this paper.

(i) We prove that for a high-volume assemble-to-order system, optimal prices and component production capacity are close to the solution of the static planning problem; see Theorem 5.1. In particular, as in the systems studied in Maglaras [26], Maglaras and Zeevi [28], and Plambeck [30], heavy traffic is the optimal operating regime, meaning that diffusion approximations are relevant.

(ii) With dynamic assembly sequencing, the system exhibits state-space collapse; see Proposition 4.1. Hence, in managing component production and inventory, one need not track the customer order queue length for every product, only the component inventory positions. This reduction in problem dimensionality is very important, because a typical assemble-to-order system is designed to support thousands of different products from tens of different components.

(iii) We provide an asymptotically optimal policy for statically setting prices and component production capacity, and dynamically sequencing orders for assembly; see Theorem 5.1.

The remainder of this paper is organized as follows. We first review some relevant literature. Next, in §2, we discuss the model formulation. Section 3 sets up the static planning problem and introduces our assembly policy. Section 4 analyzes the asymptotic behavior of the system under the “first-attempt” policy specified in §3. Finally, we show that nearly balanced systems are optimal, and provide the appropriate “stochastic adjustment” factor to prices and component production capacity in §5. Appendix A contains all of our lemma proofs, and Appendix B provides a table of notation.

1.1. Literature review. Song and Zipkin [36] provide an excellent survey of the literature on managing ATO systems. This literature is focused on the management of component inventory, taking the demand process and distribution of component lead times as given, and assuming that customer orders must be filled FIFO. The subset of this literature that adopts a discrete-time formulation must further specify how to allocate scarce components among orders that arrive simultaneously, i.e., in the same period. Each of the following three papers assumes a different rule. In Agrawal and Cohen [2], scarce components are allocated “fairly” so that in each period, the fraction of demand that is satisfied is the same for every product. In Zhang [41], orders that arrive in the same period are prioritized according to price. Akçay and Xu [1] optimize the allocation of components (among orders that arrive in the same period) to maximize the fraction of orders assembled within the quoted maximum delay. All three papers assume independent base-stock control of component inventory,
and optimize the base-stock levels. They find very different optimal base-stock levels, which suggests that the optimal inventory policy is very sensitive to the rule for allocating scarce components among orders for various products. Akçay and Xu [1] recommend that “inventory replenishment and component allocation optimizations must be made jointly” (p. 110). Clearly, these decisions are also sensitive to product prices and component lead times. This motivates our model formulation, which incorporates product prices, component production capacity, and the dynamic sequencing of orders for assembly as decision variables. We find that the optimal allocation of components to products (sequencing) collapses the state space, so the joint optimization recommended by Akçay and Xu [1] becomes more tractable than inventory optimization in isolation.

In an ATO system with only one component in shortage, the assembly-sequencing decision is similar to the decision of which class to serve in a multiclass, single-server queue. Much of the research in this area combines sequencing with dynamic lead-time quotation. Two representative publications are Wein [39] and Duenyas [13], and each of these papers provides an extensive review of related literature. When customers must be served within a class-specific maximum delay but production cannot be expedited (our follow-up paper Plambeck and Ward [31] has delay constraints and expediting), then some customers must be turned away. Maglaras and Van Mieghem [27] and Plambeck et al. [32] consider dynamic sequencing and admission control in this setting.

In practice, many firms accept all customer orders, but set a lower bound on the fill rate (fraction of customer orders assembled within the quoted maximum delay). Researchers have provided methods for optimizing the base-stock level for each component in order to minimize long-run average inventory holding costs, subject to the lower bound on the fill rate. For example, when component lead times are i.i.d. random variables (the relevant model formulation when transportation comprises most of component lead time), Song [35] and Lu et al. [23] show how to calculate fill rates, and Cheng et al. [12] go on to optimize base-stock levels. When component production capacity is limited by supply contracts and/or physical capacity constraints, lead time increases in the number of components ordered. To capture this, as in our formulation, Song et al. [37], Glasserman and Wang [19], and Wang [38] explicitly model the component production facilities as single-server queues. Glasserman and Wang [19] undertake an asymptotic analysis as the fill rate approaches one, and establish a linear relationship between the target maximum delay and the component base-stock levels.

Our follow-up paper (Plambeck and Ward [31]) assumes a fill rate of exactly one. To make this feasible, we allow the system manager to expedite components instantaneously at a high per-unit cost. In this setting, we provide an asymptotically optimal policy for statically setting prices and component production capacity, and dynamically sequencing orders for assembly and expediting components. Our control problem becomes even more challenging when the system manager can dynamically sell excess components at a low “salvage” price. We heuristically derive and numerically solve an appropriate approximating diffusion control problem.

2. Model formulation. Consider a system in which \( J \) different components are assembled into \( K \) different finished products, as shown in Figure 1. Product \( k = 1, \ldots, K \) requires a positive, integer amount of type \( j = 1, \ldots, J \) components equal to \( a_{kj} \). Every component \( j \) is required for the assembly of at least one product, so that for any \( k \), \( a_{kj} > 0 \) for at least one \( j \). Assembly is instantaneous, given the necessary components.
At time $t = 0$, the system manager chooses product prices $p_1, \ldots, p_K$ and component production capacity $\gamma_1, \ldots, \gamma_J$. Then, given the vector of product prices $p$, customers order product $k$ according to a renewal process $O_k$ with rate $\lambda_k \equiv \lambda_k(p)$. Specifically, $O_k(t)$ denotes the cumulative number of orders for product $k$ that arrive before time $t$. A customer pays $p_k$ when his order for product $k$ is filled. Components of type $j$ arrive at the assembly facility according to a renewal process $C_j$ with rate $\gamma_j$ (the component $j$ production capacity). Specifically, $C_j(t)$ denotes the cumulative number of components that arrive before time $t$. Each component has an associated unit production cost, $c_j > 0$, paid upon the delivery of the component. Pricing and component production capacity decisions are irreversible, and remain fixed throughout the time horizon. Observe that our formulation allows for customer migration between products; i.e., customer demand for any one product may increase as the prices of other products increase.

Next, the system manager must dynamically determine when and in what sequence to assemble outstanding product orders. Depending on relative prices, holding costs, outstanding orders, and component inventory levels, it may be advantageous to prioritize assembly of one product over another. Without loss of generality, we can assume that customer orders for each product are filled FIFO. Therefore, specification of the cumulative number of type $k$ orders assembled in $[0, t]$, which we denote by $A_k(t)$, uniquely determines the order queue lengths at time $t$,

$$Q_k(t) \equiv O_k(t) - A_k(t), \quad k = 1, \ldots, K,$$

and the component inventory levels at time $t$,

$$I_j(t) \equiv C_j(t) - \sum_{k=1}^K a_{kj} A_k(t) \geq 0, \quad j = 1, \ldots, J.$$

Component inventory incurs a linear physical holding cost. Let $h_j$ denote the physical holding cost per component of type $j$ per unit of time.

The objective is to maximize the expected value of the infinite-horizon discounted profit\(^2\)

$$\Pi \equiv \sum_{k=1}^K \int_0^\infty p_k e^{-\delta t} dA_k(t) - \sum_{j=1}^J \int_0^\infty e^{-\delta t} (c_j dC_j(t) + h_j I_j(t)) dt,$$

which, using the Reimann-Stieltjes integration by parts theorem and the definition of the queue length and inventory processes in (1) and (2), can also be written in the following form, which is more convenient for analysis

$$\Pi = \int_0^\infty \delta e^{-\delta t} \left( \sum_{k=1}^K p_k O_k(t) - \sum_{j=1}^J c_j C_j(t) \right) dt - \int_0^\infty e^{-\delta t} \left( \sum_{k=1}^K \delta p_k Q_k(t) + \sum_{j=1}^J h_j I_j(t) \right) dt. \quad (3)$$

Because each customer pays when his order is filled, delay in order assembly decreases discounted revenue. The system manager could avoid this loss of revenue by holding large component inventories. However, he would then purchase components long before using them, causing him both to tie up a large amount of capital in component inventory, and to incur excessive physical inventory holding costs.

An admissible policy $u = (p_u, \gamma_u, A_u)$ specifies the product prices, component production capacity, and the sequencing rule for assembly. We require that $p_u$ and $\gamma_u$ are nonnegative. Also, the process $A_u$ is integer-valued, nondecreasing, nonanticipating, and has $A_u(0) = 0$ for all $t > 0$. For clarity of presentation, we do not subscript the system processes associated with a particular policy whenever the policy under consideration is clear from the context. The reader is to understand the implicit dependence of the system processes on the policy.

In general, our stated goal of maximizing (3) by choosing product prices, component production capacity, and the sequencing rule for assembly appears intractable. However, under the high-volume conditions introduced in §4.1, our goal is achievable.

Our analysis requires a few standard assumptions on the demand function $\lambda$. First, $\lambda(p)$ is continuously differentiable, and the Jacobian matrix $[\partial \lambda_k/\partial p_m, m = 1, \ldots, K]$ is invertible everywhere. The inverse function theorem then guarantees that the inverse demand function $p(\lambda)$ is unique, continuous, and differentiable. Second, customer demand for any one product is strictly decreasing in the price of that product, but is nondecreasing.

\(^1\) We do not explicitly use notation, such as $\vec{v}$, to indicate $v$ is a vector; the reader should be able to differentiate between scalars and vectors from the context.

\(^2\) All integrals appearing in this paper should be interpreted as Riemann-Stieltjes integrals in the usual sense.
in the price of different products, so $\partial \lambda_k(p)/\partial p_k < 0$ while $\partial \lambda_k(p)/\partial p_m > 0$, $m \neq k$. Third, $\sum_{m=1}^{K} \partial \lambda_k/\partial p_m < 0$ for each $k = 1, \ldots, K$, which means that demand for each product decreases when all products’ prices increase by the same amount. Finally, we assume that the revenue rate $r(\lambda) \equiv \sum_{k=1}^{K} \lambda_k p_k(\lambda)$ is strictly concave, as in Gallego and van Ryzin [18].

We conclude by stating a lemma useful for later analysis. We delay its proof, which mimics that of Lemma 2 in Bernstein and Federgruen [8], until the appendix, where the reader can find the proofs of all the lemmas in this paper.

**Lemma 2.1.** The assumptions

$$\frac{\partial \lambda_k(p)}{\partial p_k} < 0, \quad \frac{\partial \lambda_k(p)}{\partial p_m} \geq 0, \quad m \neq k \quad \text{and} \quad \sum_{m=1}^{K} \frac{\partial \lambda_k}{\partial p_m} < 0 \quad \text{for each} \quad k = 1, \ldots, K \quad (4)$$

imply

$$\frac{\partial p_k}{\partial \lambda_k} < 0 \quad \text{and} \quad \frac{\partial p_k}{\partial \lambda_m} \leq 0, \quad m \neq k.$$

3. The static planning problem and the proposed assembly policy. Subsection 3.1 derives prices and component production capacities that maximize the profit rate, assuming that demand and production flow at their respective mean rates. This yields a first-order approximation to the optimal prices and production capacities in the system with stochastic variability. In §3.2, we propose a discrete-review assembly policy that minimizes instantaneous financial holding costs at each review point by distributing components to product orders.

3.1. An initial policy for setting prices and component production capacities. Suppose that in initially setting product prices $p$ and component production capacity $\gamma$, the system manager ignores the discrete and stochastic nature of customer orders and component production, and simply assumes that demand and production flow at their long-run average rates. Then, to maximize the profit rate, he chooses prices and component production capacities according to the solution of the following static planning problem:

$$\bar{\pi} \equiv \max_{p \geq 0, \gamma \geq 0} \sum_{k=1}^{K} p_k \lambda_k(p) - \sum_{j=1}^{J} \gamma_j c_j, \quad (5)$$

subject to the constraint that all customer orders must be filled

$$\sum_{k=1}^{K} a_{kj} \lambda_k(p) \leq \gamma_j, \quad j = 1, \ldots, J. \quad (6)$$

In preparation for later analysis, we state the following lemma, which establishes properties of the solution to (5)–(6).

**Lemma 3.1.** The unique solution to the static planning problem $(p^*, \gamma^*)$ has

$$p_k^* > \sum_{j=1}^{J} c_j a_{kj} > 0 \quad \text{for every} \quad k = 1, \ldots, K, \quad \text{and} \quad \gamma^* > 0. \quad (7)$$

Furthermore, suppose we relax the constraint (6) to yield the following perturbed problem:

$$\bar{\pi}(\theta) \equiv \max_{p \geq 0, \gamma \geq 0} \sum_{k=1}^{K} p_k \lambda_k(p) - \sum_{j=1}^{J} \gamma_j c_j, \quad (8)$$

subject to

$$\sum_{k=1}^{K} a_{kj} \lambda_k(p) \leq \gamma_j + \theta_j, \quad j = 1, \ldots, J. \quad (9)$$

If $\theta_j \leq \gamma_j$, $j = 1, \ldots, J$, the perturbed problem (8)–(9) has a unique optimal solution $(p^*(\theta), \gamma^*(\theta))$ such that

$$p^*(\theta) = p^*, \quad \gamma^*(\theta) = \gamma^* - \theta, \quad \text{and} \quad \bar{\pi}(\theta) - \bar{\pi} = \sum_{j=1}^{J} c_j \theta_j.$$

The optimal objective value in (5) upper bounds the expected profit rate, which implies $\delta^{-1}\bar{\pi}$ upper bounds expected infinite-horizon discounted profit in the stochastic system. Due to stochastic variability, customers will experience delay, and components will sit in inventory, meaning the upper bound is not, in general, achieved. However, §5 establishes that under high-volume conditions, the optimal prices and component production capacities are close to $(p^*, \gamma^*)$, and infinite-horizon discounted expected profit is close to $\delta^{-1}\bar{\pi}$.
3.2. The proposed assembly policy. The assembly policy we propose is a discrete-review policy that releases orders for assembly at review time points, and does nothing at all other times.\footnote{For analytic convenience, as is standard in the assemble-to-order literature, we assume that assembly is instantaneous. If assembly time is a constant $\tau_k$ for each product $k$, then our analysis goes through with $e^{-\tau_k p_k}$ substituted for $p_k$.} We let $l, 2l, 3l, \ldots$ be the discrete-review time points, where the review-period length

$$I \equiv \left( \frac{1}{|A^*|} \right)^{2/3}$$

depends upon the average time between order arrivals, and $|\cdot|$ is the Euclidean norm. Hence, the review-period length will become short in high volume.

Initially, the number of orders assembled by time 0, $A^*(0)$, is zero. At each subsequent review time point, we allocate available inventory to product orders so as to minimize instantaneous holding costs, given the shortage of each component. The shortage\footnote{The shortage is the negative of the inventory position, the term commonly used in the operations management literature. We use the term “shortage” to focus attention on how scarce components are allocated to outstanding orders.} process tracks the difference between the number of components required to assemble all outstanding orders and the number of components in inventory, defined as

$$S_j(t) \equiv \sum_{k=1}^{K} a_{kj} Q_k(t) - C_j(t), \quad j = 1, \ldots, J,$$

$$= \sum_{k=1}^{K} a_{kj} Q_k(t) - I_j(t), \quad j = 1, \ldots, J.\tag{10}$$

Let the functions (which we derive in the paragraphs following) $Q^*(S, O, A)$ and $I^*(S, O, A)$ denote the arrangement of queue lengths and inventory levels that minimize instantaneous holding costs when the shortage process is $S$, the cumulative number of order arrivals is $O$, and $A$ is the number of orders that have been assembled so far. Then, the assembly process $A^*$ is defined recursively as follows:

$$A^*(il) = O(il) - Q^*(S(il), O(il), A^*((i-1)l)),\tag{12}$$

recalling that $A^*(0) = 0$. No order assembly occurs between review time points. Our discrete-review policy contrasts with those in the work of Harrison [21] and Ata and Kumar [4] because a time allocation is not relevant to our situation.

When the solution to the linear program

$$\min_{Q, I \geq 0} \delta \sum_{k=1}^{K} p_k^* Q_k + \sum_{j=1}^{J} h_j I_j\tag{13}$$

subject to

$$I_j \equiv \sum_{k=1}^{K} a_{kj} Q_k - S_j \geq 0, \quad j = 1, \ldots, J,$$

$$O - Q \geq A \tag{14}$$

is unique, $Q^*$ and $I^*$ are exactly the solution to (13)–(15). Here, constraint (14) requires that inventory levels be nonnegative. Constraint (15) prevents already-assembled orders from being disassembled.

Otherwise, when multiple solutions to (13)–(15) exist, we examine constraint (15). If (15) binds for any one of the solutions, we arbitrarily choose a solution to use to define $Q^*$ and $I^*$. Otherwise, if (15) is slack for every solution, we carefully select a solution $(q^*, i^*)$ to the linear program in (13)–(14), ignoring constraint (15) as follows.

**Assumption 3.1.** The function $q^*$ is chosen as a Lipschitz continuous function of the shortage process $S$. In particular, for any $S^1 \in \mathbb{R}^J$ and $S^2 \in \mathbb{R}^J$,

$$\max_{k=1, \ldots, K} \left| q^*_k(S^1) - q^*_k(S^2) \right| \leq \kappa \max_{j=1, \ldots, J} \left| S^1_j - S^2_j \right|,\tag{16}$$

where $\kappa$ is a finite constant.
In the case that the solution to (13)–(14) is unique for every $S$, Theorem 10.5 in Schrijver [34] (which is based on Hoffman’s lemma) establishes (16). To see when uniqueness for every $S$ occurs, observe that the objective function (13) can be written as
\[ f(S) = \sum_{j=1}^{J} h_j S_j, \]
where
\[ f(S) \equiv \min_{Q \geq 0} \delta \sum_{k=1}^{K} \left( p^*_k + \sum_{j=1}^{J} h_j a_{kj} \right) Q_k \]
subject to
\[ \sum_{k=1}^{K} a_{kj} Q_k \geq S_j, \quad j = 1, \ldots, J. \]
The linear program defining $f$ has a unique solution whenever the vector of “cost coefficients”
\[ \left( p^*_1 + \sum_{j=1}^{J} h_j a_{1j}, \ldots, p^*_K + \sum_{j=1}^{J} h_j a_{Kj} \right) \]
is not parallel to $(a_{kj})_{k=1,\ldots,K}$, the vector of products requiring component $j$, for every $j = 1, \ldots, J$.
If the vector in (17) is parallel to $(a_{kj})_{k=1,\ldots,K}$ for some $j$, then the solution to (12)–(13) is not unique for some $S$. Nevertheless, one can identify a Lipschitz solution $q^*(S)$ by perturbing the price vector $p^*$ very slightly, so that the vector in (17) is not parallel to $(a_{kj})_{k=1,\ldots,K}$ for any $j = 1, \ldots, J$, yet the (unique) optimal solution to the perturbed linear program is an optimal solution to the unperturbed linear program. Once equipped with the $q^*(S)$, one should proceed with the original (unperturbed) $p^*$ for the remainder of the analysis. Existence of a Lipschitz $q^*(S)$ is proven rigorously in Proposition 2 of Bassamboo et al. [5]. For additional explanation of the issue regarding the construction of a Lipschitz continuous function, see Berens et al. [7].

The linear program in (13)–(15) does not account for the effect of assembly in the current period on future holding costs due to constraint (15). Nevertheless, the myopia of the proposed policy does not cause problems in the high-volume limit. Intuitively, when the system experiences a high volume of demand and component production, the functional central limit theorem dictates that the shortage process in (10) changes gradually in the high-volume limit. In other words, they are almost deterministic functions of the shortage process; see Proposition 4.1.

4. System behavior under the static planning problem solution and the proposed assembly policy. To analyze the performance of a policy that sets prices and component production capacities using the solution to the static planning problem in §3.1, and assembles orders according to the proposed assembly policy in §3.2, we undertake an asymptotic analysis as the customer arrival rate increases. We use the colloquial term “high volume” to refer to a system with a high order arrival rate as defined in §4.1. Subsection 4.2 shows that our proposed assembly policy forces queue lengths and inventory levels to track deterministic functions of the shortage process with very high probability. Finally, §4.3 establishes the asymptotic behavior of the shortage, queue length, and inventory processes, and of expected infinite-horizon discounted profit.

4.1. High-volume conditions. The one modelling assumption necessary for our asymptotic analysis is that the system experiences a high volume of demand. Specifically, consider the sequence $n = 1, 2, \ldots$. Let order arrival rates tend to infinity in a manner that preserves the structure of the demand functions, as follows:
\[ \lambda^*_k(p) \equiv n \lambda_k(p), \quad k = 1, \ldots, K. \]
Henceforth, when we wish to refer to any process or other quantity associated with the assemble-to-order system having order arrival rate function $\lambda^*$, we superscript the appropriate symbol by $n$. An admissible policy refers to an entire sequence, $u = (p^n_*, \gamma^n, A^n)$, that specifies an admissible policy for each $n$. We let $\lambda^n_{*,k} \equiv \lambda^*_k(p^n_*)$. In high volume, the review-period length is
\[ l^n = \left( \frac{1}{\rho^*} \right)^{2/3}. \]
The following technical specifications are needed. Let \((\Omega, \mathcal{F}, P)\) be a probability space. For each positive integer \(i\), let \(D^i\) be the space of all functions \(\omega: [0, \infty) \to \mathbb{R}^i\) that are right continuous with left limits. Consider \(D^i\) to be endowed with the usual Skorohod-\(J_1\) topology (see, for example, Ethier and Kurtz [16] or Billingsley [9]), and let \(M^i\) denote the Borel \(\sigma\)-algebra on \(D^i\) associated with this topology. All stochastic processes in this paper are measurable functions from \((\Omega, \mathcal{F}, P)\) into \((D^i, M^i)\) for the appropriate dimension \(i\). For a sequence of stochastic processes \(\{\xi^n\}_n\), each of dimension \(i\), the notation \(\xi^n \Rightarrow \xi\) means the probability measures induced by \(\xi^n\) on \((D^i, M^i)\) converge weakly to the probability measure induced by \(\xi\) on \((D^i, M^i)\).

We represent the renewal processes \(O_1, \ldots, O_K\) and \(C_1, \ldots, C_J\) in terms of the \(K + J\) independent sequences of mean nonnegative random variables \(\{x_k(i), i = 1, 2, \ldots\}, k = 1, \ldots, K\) and \(\{y_j(i), i = 1, 2, \ldots\}, j = 1, \ldots, J\) having

\[
\text{Var}(x_k(i)) = \sigma^2_{\Omega, k} \quad \text{and} \quad \text{Var}(y_j(i)) = \sigma^2_{C, j}
\]
as follows. Let

\[
\tilde{\mathcal{X}}_k(m) = \sum_{i=1}^{m} x_k(i), \quad k = 1, \ldots, K \quad \text{and} \quad \tilde{\mathcal{Y}}_j(m) = \sum_{i=1}^{m} y_j(i), \quad j = 1, \ldots, J,
\]
and define

\[
O^n_{\alpha, k}(t) = \max\{m \geq 0: \tilde{\mathcal{X}}_k(m) \leq \lambda^n_{\alpha, k} t\} \quad \text{and} \quad C^n_{\alpha, j}(t) = \max\{m \geq 0: \tilde{\mathcal{Y}}_j(m) \leq n \gamma^n_{\alpha, j} t\}.
\]

We assume that \(E|x_k(1)|^{2+\varepsilon} < \infty, k = 1, \ldots, K\) and \(E|y_j(1)|^{2+\varepsilon} < \infty, j = 1, \ldots, J\) for \(\varepsilon = 3\). Normal, truncated normal, uniform, and Erlang random variables satisfy this condition, as do Weibull and gamma random variables for a restricted set of parameters. We use the assumption of finite \(2 + \varepsilon\) moments in the proof of Lemma 4.1 to obtain a probabilistic bound on the number of orders and components arriving within a review period. The work of Ata and Kumar [4] implies that the minimal assumption necessary for our analysis is \(\varepsilon > 0\), but we set \(\varepsilon = 3\) to avoid defining the review-period length in terms of \(\varepsilon\), which simplifies and shortens our presentation.

It is convenient for our analysis to also define the scaled and centered processes

\[
\tilde{\mathcal{O}}_{\alpha, k}(t) = \sqrt{n}(n^{-1}O^n_{\alpha, k}(t) - \lambda_k(p^n)t)
\]
and

\[
\tilde{\mathcal{C}}_{\alpha, j}(t) = \sqrt{n}(n^{-1}C^n_{\alpha, j}(t) - \gamma^n_j t).
\]

4.2. Reduction in problem dimensionality. If the constraint (15) that orders cannot be disassembled did not exist, our assembly policy would not be worrisome. Then, queue lengths and inventory levels would minimize instantaneous holding costs at all review points \(i^n, i = 0, 1, 2, \ldots\), meaning they would equal \((q^*(S^n(i^n)), i^*(S^n(i^n)))\). In high volume, our next proposition shows that with very high probability, this is exactly the case. In particular, our assembly policy achieves a reduction in problem dimensionality because queue lengths and inventory levels are, with very high probability, deterministic functions of the short process. Define \(I^n = \lfloor 1/l^n \rfloor = \lfloor |n\lambda^n|^{2/3} \rfloor\).

**Proposition 4.1.** Under any policy \(u\) with \(p^n_u \to p^*\) and \(\gamma^n_u \to \gamma^*\) as \(n \to \infty\), there exists a constant \(\beta\) such that under the proposed assembly policy

\[
P(Q^n(i^n) = q^*(S^n(i^n)) \text{ for all } i = 1, 2, \ldots, I^n) \geq 1 - \beta n^{-1/6}.
\]

The proof of Proposition 4.1 requires the following lemma, which bounds the number of orders and components arriving during each review period.

**Lemma 4.1 (Ata and Kumar [4]).** Under any policy \(u\) with \(p^n_u \to p^*\) and \(\gamma^n_u \to \gamma^*\) as \(n \to \infty\), for each \(k = 1, \ldots, K\) and \(j = 1, \ldots, J\), and for any finite constant \(\alpha > 0\), there exists a constant \(\beta > 0\) such that

\[
P\left(\max_{j=1}^{m} \max_{k=1}^{K} |O^n_k((i+1)I^n) - O^n_k(iI^n) - n\lambda_k(p^n_u)I^n| < \alpha n^{1/3}\right) \geq 1 - \beta n^{-1/6}
\]
and

\[
P\left(\max_{j=1}^{m} \max_{j=1}^{J} |C^n_j((i+1)I^n) - C^n_j(iI^n) - n\gamma^n_{u,j}I^n| < \alpha n^{1/3}\right) \geq 1 - \beta n^{-1/6}.
\]
Proof of Proposition 4.1. Fix \( n \), let

\[
\alpha \equiv \min_{k=1, \ldots, K} \min \frac{1}{2} \frac{\lambda_k^2}{|\lambda|^2/3} 1 + \kappa \max_{j=1, \ldots, J} \left( 1 + \sum_{k=1}^K a_{kj} \right),
\]

and consider a sample path \( \omega \in \Omega \) on which

\[
\max_{i=0,1, \ldots, l^n-1} \max_{k=1, \ldots, K} \left| O_k^n((i+1)l^n) - O_k^n(il^n) - \lambda_k(p_k^n)nl^n \right| < \alpha n^{1/3}\tag{21}
\]

and

\[
\max_{i=0,1, \ldots, l^n-1} \max_{j=1, \ldots, J} \left| C_j^n((i+1)l^n) - C_j^n(il^n) - \gamma_{u,n} nl^n \right| < \alpha n^{1/3}.\tag{22}
\]

Because for every \( i = 0, 1, \ldots, l^n - 1 \)

\[
\sum_{k=1}^K p_k^n q_k^n(S^n(il^n)) \leq \sum_{k=1}^K p_k^n Q_k^n(S^n(il^n), O^n(il^n), A^n((i-1)l^n)),
\]

if

\[
O_k^n(il^n) - q_k^n(S^n(il^n)) \geq A_k^n((i-1)l^n)\tag{23}
\]

for each \( k = 1, \ldots, K \) and every \( i = 1, \ldots, l^n \), so that constraint (15) is not violated at \( q_k^n(S^n(il^n)) \), then

\[
Q^n(il^n) = Q^*(S^n(il^n), O^n(il^n), A^n((i-1)l^n)) = q^*(S^n(il^n)).
\]

We use mathematical induction to show (23) holds for every \( i = 0, 1, \ldots, l^n - 1 \). For \( i = 1 \), observing that \( q^*(0) = 0 \), use (21), (22), and the Lipschitz continuity of \( q^* \) to find

\[
O_k^n(il^n) - q_k^n(S^n(il^n)) \geq \lambda_k(p_k^n)nl^n - \alpha n^{1/3} - \kappa \max_{j=1, \ldots, J} |S^n_j(il^n)|
\]

\[
\geq \lambda_k(p_k^n)nl^n - \alpha n^{1/3} - \kappa \alpha n^{1/3} \max_{j=1, \ldots, J} \left( 1 + \sum_{k=1}^K a_{kj} \right)
\]

\[
= n^{1/3} \left( \frac{\lambda_k(p_k^n)}{|\lambda|^2/3} - \alpha \left( 1 + \kappa \max_{j=1, \ldots, J} \left( 1 + \sum_{k=1}^K a_{kj} \right) \right) \right)
\]

\[
> 0 = A_k^n(0) \quad \text{(for large enough } n)\]

for \( k = 1, \ldots, K \). Hence, \( Q^*(l^n) = q^*(S^n(l^n)) \). Next, assuming \( Q^*(il^n) = q^*(S^n(il^n)) \) so that \( A^n(il^n) = O^n(il^n) - q^*(S^n(il^n)) \), the following similar argument

\[
O_k^n((i+1)l^n) - q_k^n(S^n((i+1)l^n))
\]

\[
= O_k^n((i+1)l^n) - O_k^n(il^n) + O_k^n(il^n) + q_k^n(S^n(il^n)) - q_k^n(S^n((i+1)l^n)) - q_k^n(S^n((i+1)l^n)) + A_k^n(il^n)
\]

\[
\geq n^{1/3} \left( \frac{\lambda_k(p_k^n)}{|\lambda|^2/3} - \alpha \left( 1 + \kappa \max_{j=1, \ldots, J} \left( 1 + \sum_{k=1}^K a_{kj} \right) \right) \right) + A_k^n(il^n)
\]

\[
> A_k^n(il^n) \quad \text{(for large enough } n)
\]

establishes \( Q^*((i+1)l^n) = q^*(S^n((i+1)l^n)) \). Lemma 4.1 guarantees there exists a constant \( \beta \) such that the probability of having a sample path that satisfies (21) and (22) exceeds \( 1 - \beta n^{-1/6} \), which completes the proof.

4.3. Asymptotic behavior of the system. Proposition 4.1 is the key to showing the proposed assembly policy achieves a reduction in problem dimensionality in high volume. In particular, our first theorem, which establishes the asymptotic behavior of the system, shows the following.

(i) Queue lengths and inventory levels viewed under diffusion scaling are approximately deterministic functions of the shortage process in high volume.

(ii) Expected infinite-horizon discounted profit in the \( n \)th system is within \( \sqrt{n} \) of the upper bound from the static planning problem, \( \delta^{-1} \sqrt{n} \).
In preparation for the statement of Theorem 4.2 and later results, define $\Gamma$ to be the $J \times J$ matrix whose $(i, j)$th entry is

$$
\Gamma_{i,j} \equiv \sum_{k=1}^{K} a_{ik}a_{kj}\lambda_k^2 \sigma_{\theta,k}^2 + \gamma_j^2 \sigma_{C,j}^2 \mathbf{1}[i=j].
$$

(24)

The matrix $\Gamma$ is the covariance matrix for the Brownian motion appearing in Proposition 4.2 as the heavy-traffic limit for the shortage process. The following uniform integrability result, which employs arguments similar to those necessary to establish (155) in the proof of Theorem 5.3 in Bell and Williams [6], is also useful.

**Lemma 4.2.** Under any policy $u$ with $p^n_u \to p^*$ and $\gamma_u^n \to \gamma^*$ as $n \to \infty$, for each $k = 1, \ldots, K$ and $j = 1, \ldots, J$,

$$
\left\{ \int_{0}^{\infty} e^{-\delta t} \sup_{0 \leq s \leq t} |\tilde{\Theta}^n_{u,k}(s)| \, dt, \, n \geq 0 \right\} \quad \text{and} \quad \left\{ \int_{0}^{\infty} e^{-\delta t} \sup_{0 \leq s \leq t} |\tilde{C}^n_{u,j}(s)| \, dt, \, n \geq 0 \right\}
$$

are uniformly integrable families. Furthermore, as $n \to \infty$, for each $k = 1, \ldots, K$ and $j = 1, \ldots, J$,

$$
E \int_{0}^{\infty} e^{-\delta t} \tilde{\Theta}^n_k(t) \, dt \to 0 \quad \text{and} \quad E \int_{0}^{\infty} e^{-\delta t} \tilde{C}^n_{u,j}(t) \, dt \to 0.
$$

Recall that $\Pi^n_{[p^*, \gamma^*, \lambda^n]}$ is the infinite-horizon discounted profit (defined in (3)) for the system having arrival rate $n\lambda^n_k$, operating under the policy $(p^*, \gamma^*, \lambda^n)$.

**Proposition 4.2.** Under the policy $(p^*, \gamma^*, \lambda^n)$, for $B$ a Brownian motion with 0 drift and covariance matrix $\Gamma$, as $n \to \infty$,

(a) $(S^n, Q^n, I^n)/\sqrt{n} \Rightarrow (B, q^*(B), i^*(B))$,

(b) $E \left[ \frac{\Pi^n_{[p^*, \gamma^*, \lambda^n]} - \delta^{-1} n \bar{P}}{\sqrt{n}} \right] \to -E \left[ \int_{0}^{\infty} e^{-\delta t} \left( \sum_{k=1}^{K} \lambda_k^2 \sigma_{\theta,k}^2 B^O_k(t) + \sum_{j=1}^{J} \gamma_j^2 \sigma_{C,j}^2 B^C_j(t) \right) \, dt \right].$

**Proof.** Proof of (a): By the functional central limit theorem (see, for example, Theorem 5.11 in Chen and Yao [11]),

$$
\tilde{\Theta}^n_k \Rightarrow \sqrt{\lambda_k^2 \sigma_{\theta,k}^2} B^O_k \quad \text{and} \quad \tilde{C}^n_j \Rightarrow \sqrt{\gamma_j^2 \sigma_{C,j}^2} B^C_j,
$$

where $B^O_k, k = 1, \ldots, K$ and $B^C_j, j = 1, \ldots, J$ are $K + J$ independent standard Brownian motions. Therefore, by the continuous-mapping theorem, any linear combination of the components of $n^{-1/2} S^n$ weakly converges as follows:

$$
\sum_{j=1}^{J} x_j \frac{S^n_j}{\sqrt{n}} = \sum_{j=1}^{J} x_j \left( \sum_{k=1}^{K} a_{kj} \tilde{\Theta}^n_k - \tilde{C}^n_j \right)
$$

$$
\Rightarrow \sum_{j=1}^{J} x_j \left( \sum_{k=1}^{K} a_{kj} \sqrt{\lambda_k^2 \sigma_{\theta,k}^2} B^O_k - \sqrt{\gamma_j^2 \sigma_{C,j}^2} B^C_j \right)
$$

for each $(x_1, \ldots, x_J) \in \mathbb{R}^J$. Furthermore, for any $i = 1, \ldots, J$ and $j = 1, \ldots, J$,

$$
\text{Cov} \left( \sum_{k=1}^{K} a_{ik} \sqrt{\lambda_k^2 \sigma_{\theta,k}^2} B^O_k - \sqrt{\gamma_j^2 \sigma_{C,j}^2} B^C_j, \sum_{k=1}^{K} a_{kj} \sqrt{\lambda_k^2 \sigma_{\theta,k}^2} B^O_k - \sqrt{\gamma_j^2 \sigma_{C,j}^2} B^C_j \right)
$$

$$
= \sum_{k=1}^{K} a_{ik} a_{kj} \lambda_k^2 \sigma_{\theta,k}^2 + \gamma_j^2 \sigma_{C,j}^2 \mathbf{1}[i = j].
$$

Theorem 7.7 in Billingsley [9] (the Cramer-Wold device) now establishes

$$
\frac{S^n}{\sqrt{n}} \Rightarrow B,
$$

(25)

where $B$ is a $J$-dimensional Brownian motion with mean 0 and covariance matrix $\Gamma$ defined in (24).
Because assembly only occurs at discrete-review time points, for any \( k = 1, \ldots, K \),
\[
\sup_{0 \leq t \leq 1} \left| n^{-1/2} Q^n_k(t) - q^*_k \left( \frac{S^n(t)}{\sqrt{n}} \right) \right| \\
= \sup_{0 \leq t \leq 1} \left| n^{-1/2} \left( Q^n_k \left( \left\lfloor \frac{t}{I^n} \right\rfloor I^n \right) + O^n_k(t) - O^n_k \left( \left\lfloor \frac{t}{I^n} \right\rfloor I^n \right) \right) - q^*_k \left( \frac{S^n \left( \left\lfloor \frac{t}{I^n} \right\rfloor I^n \right)}{\sqrt{n}} \right) \right| \\
\leq \max_{i=0,1,\ldots,1/I^n} \left| n^{-1/2} Q^n_k(i I^n) - q^*_k \left( \frac{S^n(i I^n)}{\sqrt{n}} \right) \right| + n^{-1/2} \max_{i=0,1,\ldots,1/I^n} \left| O^n_k(i I^n) - O^n_k(i I^n) - n \lambda^*_k I^n \right| \\
+ \sup_{0 \leq t \leq 1} q^*_k \left( \frac{S^n(t)}{\sqrt{n}} \right) - q^*_k \left( \frac{S^n(t)}{\sqrt{n}} \right) \right| + n^{-1/2} \frac{\lambda^*_k}{\lambda^*_k} \\
\rightarrow 0
\]
in probability, as \( n \to \infty \), by Proposition 4.1, the observation that for any \( S \in \mathfrak{N} \), \( q^*(n^{-1/2} S) = n^{-1/2} q^*(S) \), Lemma 4.1, and the fact that \( i I^n, i = 0, 1, 2, \ldots, (I^n)^{-1} \) becomes dense in \((0, 1)\). Because by the continuous-mapping theorem \( q^*(n^{-1/2} S) \Rightarrow q^*(B) \), we conclude
\[
\frac{Q^n}{\sqrt{n}} \Rightarrow q^*(B),
\]
as \( n \to \infty \).

Recall from (11) that scaled inventory levels are represented as a linear combination of scaled queue lengths and the scaled shortage process. Hence, the continuous-mapping theorem establishes
\[
\frac{I^n}{\sqrt{n}} \Rightarrow \sum_{k=1}^K a_k q^*(B) - B = i^*(B)
\]
as \( n \to \infty \). Finally, the stated joint convergence follows because \( q^* \) and \( i^* \) are deterministic functions.

**Proof of (b):** The representation for infinite-horizon discounted profit in (3) shows
\[
\Pi^n_{i^*, \gamma^*, \lambda^*} = \delta^{-1} n \mathbb{E} = \int_0^\infty e^{-\delta t} \left( \sum_{k=1}^K p_k^* \frac{O^n_k(t) - n \lambda^*_k t}{\sqrt{n}} - \sum_{j=1}^J c_j \frac{C^n_j(t) - n \gamma^*_j t}{\sqrt{n}} \right) dt \\
- \int_0^\infty e^{-\delta t} \left( \sum_{k=1}^K \frac{h_k^*}{\sqrt{n}} \right) dt.
\]

From our assembly policy, because \( a_{kj} \) assumes integer values, for any \( t > 0 \),
\[
Q^n_k(t) \leq \max_{j=1,\ldots,J} \left\{ S^n_j \left( \left\lfloor \frac{t}{I^n} \right\rfloor I^n \right) \right\} + O^n_k(t) - O^n_k \left( \left\lfloor \frac{t}{I^n} \right\rfloor I^n \right) \\
\leq J \max_{j=1} \left\{ S^n_j \left( \left\lfloor \frac{t}{I^n} \right\rfloor I^n \right) \right\} + O^n_k(t) - O^n_k \left( \left\lfloor \frac{t}{I^n} \right\rfloor I^n \right) \\
\leq J \left( \sum_{k=1}^K a_{kj} \sup_{0 \leq s \leq t} |O^n_k(s) - n \lambda^*_k s| \right) + \sup_{0 \leq s \leq t} |C^n_j(s) - n \gamma^*_j s| \\
+ 2 \sup_{0 \leq s \leq t} |O^n_k(s) - n \lambda^*_k s| + n^{1/3} \frac{\lambda^*_k}{\lambda^*_k}. \tag{29}
\]
Hence, Lemma 4.2 guarantees \( \int_0^\infty e^{-\delta t} n^{-1/2} Q^n_k(t) dt \) is less than the sum of uniformly integrable families, and so forms a uniformly integrable family for each \( k = 1, \ldots, K \). Also, because
\[
I^n_j(t) \leq \sum_{k=1}^K a_{kj}Q^n_k \left( \left\lfloor \frac{t}{I^n} \right\rfloor I^n \right) - S^n_j \left( \left\lfloor \frac{t}{I^n} \right\rfloor I^n \right) + C^n_j(t) - C^n_j \left( \left\lfloor \frac{t}{I^n} \right\rfloor I^n \right),
\]
the bound in (29) and Lemma 4.2 establish that \( \{ \int_0^\infty e^{-\delta t} n^{-1/2} I^x_j(t) \, dt \} \) forms a uniformly integrable family for each \( j = 1, \ldots, J \). Therefore, part (a), the continuous-mapping theorem, and the now-justified interchange of limit and expectation show, as \( n \to \infty \),

\[
E \left[ \int_0^\infty e^{-\delta t} \frac{Q^x_j(t)}{\sqrt{n}} \, dt \right] \to E \left[ \int_0^\infty e^{-\delta t} q^*_k(B(t)) \, dt \right], \quad k = 1, \ldots, K, \tag{30}
\]

\[
E \left[ \int_0^\infty e^{-\delta t} \frac{I^x_j(t)}{\sqrt{n}} \, dt \right] \to E \left[ \int_0^\infty e^{-\delta t} i^*_j(B(t)) \, dt \right], \quad j = 1, \ldots, J. \tag{31}
\]

Lemma 4.2, and the convergences in (30) and (31), establish the convergence of the expected value of (28), i.e., that

\[
\mathbb{E} \left[ \Pi^\pi_{(\rho^*, \gamma^*, \Lambda^x)} - \delta^{-1} n \bar{\pi} \right] \to \mathbb{E} \left[ \int_0^\infty e^{-\delta t} \left( \sum_{k=1}^K p^*_k \delta q^*(B(t)) + \sum_{j=1}^J h_j i^*_j(B(t)) \right) \, dt \right]. \tag{32}
\]

Observe that the proof of Proposition 4.2 goes through with only a small technical modification when prices and component arrival rates are adjusted slightly in the \( n \)th system, but the assembly policy \( A^x_n \) is unchanged. Define the capacity imbalance of the system under a given policy \( u^* \):

\[
\theta^x_{u^*, j} \equiv \sum_{k=1}^K a_{kj} \lambda_k (p^x_n) - \gamma^x_{u^*, j}, \quad j = 1, \ldots, J.
\]

Because

\[
\frac{S^x_{u^*, j}(t)}{\sqrt{n}} = \sum_{k=1}^K a_{kj} \tilde{O}^x_k(t) - \tilde{C}^x_j(t) + \sqrt{n t} \theta^x_{u^*, j}, \tag{33}
\]

provided

\[
\sqrt{n \theta^x_{u^*, j}} \to \theta_j,
\]

as \( n \to \infty \), the same arguments used to establish the weak convergence in (25) show

\[
\frac{S^x_n}{\sqrt{n}} \to B_\theta,
\]

as \( n \to \infty \), where \( B_\theta \) is a Brownian motion with drift \( \theta \) and covariance matrix \( \Gamma \) defined in (24). When \( p^x_n \to p^* \) and \( \gamma^x_n \to \gamma^* \) as \( n \to \infty \), Proposition 4.1 and Lemma 4.1 hold. Therefore, the arguments establishing the weak convergence of the scaled queue length and inventory processes in (26) and (27) remain valid. Furthermore, the arguments establishing the convergence of expected values in (30) and (31) remain valid, and we can conclude

\[
E \left[ \int_0^\infty e^{-\delta t} \left( \sum_{k=1}^K p^*_k \delta q_k^*(B_\theta(t)) + \sum_{j=1}^J h_j i_j^*(B_\theta(t)) \right) \, dt \right] \to \mathcal{H}(\theta), \tag{34}
\]

where \( \mathcal{H}(\theta) \) represents the expected discounted “holding cost” caused by stochastic imbalance in supply and demand for components

\[
\mathcal{H}(\theta) \equiv E \left[ \int_0^\infty e^{-\delta t} \left( \sum_{k=1}^K p^*_k \delta q_k^*(B_\theta(t)) + \sum_{j=1}^J h_j i_j^*(B_\theta(t)) \right) \, dt \right]. \tag{35}
\]

Define

\[
\Pi^\pi_n = \frac{\Pi^\pi_{(\rho^x_n, \gamma^x_n, \Lambda^x_n)} - \delta^{-1} n \bar{\pi}}{\sqrt{n}}
\]

\[
= \int_0^\infty e^{-\delta t} \left( \sum_{k=1}^K p^x_n \tilde{O}^x_k(t) \right) - \sum_{j=1}^J c_j \tilde{C}^x_j(t) \right) \, dt
\]

\[
- \int_0^\infty e^{-\delta t} \left( \delta \sum_{k=1}^K p^x_n \frac{Q^x_k(t)}{\sqrt{n}} + \sum_{j=1}^J h_j \frac{I^x_j(t)}{\sqrt{n}} \right) \, dt
\]

\[
+ \delta^{-1} \sqrt{n} \left( \sum_{k=1}^K p^x_n \lambda_k (p^x_n) - \sum_{j=1}^J c_j \gamma^x_{u^*, j} - \bar{\pi} \right).
\]
When \( p^*_n \) and \( \gamma^*_n \) solve the perturbed static planning problem so that \( p^*_n = \rho^*(\theta^*_n) \) and \( \gamma^*_n = \gamma^*(\theta^*_n) \), by Lemma 3.1,
\[
\sum_{k=1}^{K} p^*_n k \lambda_k (p^*_n) - \sum_{j=1}^{J} c_j \gamma^*_n, - ~ \bar{\pi} = \sum_{j=1}^{J} c_j \theta^*_n, j
\]
and so Lemma 4.2 and the convergence in (34) shows
\[
E[\tilde{\Pi}^*_n] \rightarrow \delta^{-1} \sum_{j=1}^{J} c_j \theta_j - \mathcal{H}(\theta),
\]
as \( n \rightarrow \infty \). Observe from the definition of \( \tilde{\Pi}^*_n \) that \( E[\tilde{\Pi}^*_n] \) and its limiting value are negative because \( \delta^{-1} n \bar{\pi} \) is an upper bound on the expected infinite-horizon discounted profit.

We have now established the following corollary to Proposition 4.2.

Corollary 4.1. Let \( B_\theta \) be a Brownian motion with drift \( \theta \) and covariance matrix \( \Gamma \) defined in (24). Under a policy \( u = (p^*_n, \gamma^*_n, A^*_n) \) having
\[
\sqrt{n} \theta^*_n, j \rightarrow \theta_j, \quad j = 1, \ldots, J
\]
as \( n \rightarrow \infty \), and \( p^*_n = \rho^*(\theta^*_n) \) and \( \gamma^*_n = \gamma^*(\theta^*_n) \) so that \( p^*_n \) and \( \gamma^*_n \) solve the perturbed static planning problem in (8)–(9), as \( n \rightarrow \infty \),
(a) \((S^n, Q^n, I^n)/\sqrt{n} \Rightarrow (B_\theta, q^*(B_\theta), i^*(B_\theta))\),
(b) \( E[\tilde{\Pi}^*_n] \rightarrow \delta^{-1} \sum_{j=1}^{J} c_j \theta_j - \mathcal{H}(\theta) \).

5. An asymptotically optimal policy. Adjusting product prices and component production capacities potentially improves the performance of the policy \((\rho^*, \gamma^*, A^*_n)\) in high volume. In particular, Corollary 4.1 suggests that any asymptotically optimal policy \( u \) should have
\[
\lim_{n \to \infty} E[\tilde{\Pi}^*_n] = \sup_{\theta \in \Theta^*} \left\{ \delta^{-1} \sum_{j=1}^{J} c_j \theta_j - \mathcal{H}(\theta) \right\}. \tag{36}
\]
Unfortunately, guaranteeing the existence of a unique maximizer of the right-hand side of (36) is difficult, because for a nontrivial portion of the parameter space, the function
\[
\delta \sum_{k=1}^{K} p^*_k q^*_j(S) + \sum_{j=1}^{J} h_j i^*_j(S)
\]
is not convex in \( S \). (We show a large region of the parameter space in which convexity does not hold in §5.2 of Plambeck and Ward [31].) However, the following lemma provides the basis for the definition of a policy that is asymptotically optimal.

Lemma 5.1. The set of maximizers
\[
\Theta^* = \left\{ \theta \in \mathcal{N}^J : \delta^{-1} \sum_{j=1}^{J} c_j \theta_j - \mathcal{H}(\theta) \geq \delta^{-1} \sum_{j=1}^{J} c_j \theta_j' - \mathcal{H}(\theta') \quad \text{for all} \quad \theta' \in \mathcal{N}^J \right\}
\]
is nonempty and
\[
\Theta^* \subset \left[ -c_{i_1}^{-1} \delta \mathcal{H}(0) \left( 1 + \max_{l \neq k} \left\{ c_l / \Delta \right\} \right), \delta \mathcal{H}(0)/\Delta \right] \times \cdots \times \left[ -c_{i_J}^{-1} \delta \mathcal{H}(0) \left( 1 + \max_{l \neq k} \left\{ c_l / \Delta \right\} \right), \delta \mathcal{H}(0)/\Delta \right],
\]
where
\[
\Delta = \min_{k=1, \ldots, K} \left\{ \frac{p^*_k - \sum_{j=1}^{J} a_{ij} c_j}{\left( \sum_{j=1}^{J} d_{ij} \right)} \right\}.
\]

Recall from Lemma 3.1 that \( p^*_k = \sum_{j=1}^{J} a_{ij} c_j > 0 \) for all \( k \), so \( \Delta > 0 \).

The main result of this paper both establishes that optimal pricing and capacity decisions create heavy-traffic conditions, and provides an asymptotically optimal policy.

Definition 5.1. A policy \( * \) is said to be asymptotically optimal if it is admissible and
\[
\liminf_{n \to \infty} E[\tilde{\Pi}^*_n] \geq \limsup_{n \to \infty} E[\tilde{\Pi}^*_n] \tag{37}
\]
for any other admissible policy \( u \).
Our definition of asymptotic optimality captures deviations from the upper bound on infinite-horizon expected discounted profit in the stochastic system of order \( \sqrt{n} \). To see that this scaling is the appropriate one, first realize that the functional central limit theorem applied to the shortage process in (32) shows shortage process fluctuations are inherently of order \( \sqrt{n} \). Hence, order queues and component inventory levels also may be of order \( \sqrt{n} \). Therefore, under any admissible policy, infinite-horizon discounted profit will not be closer than order \( \sqrt{n} \) to its upper bound.

**Theorem 5.1.** For \( \theta^* \in \Theta^* \), the policy \( \pi = (p^*, \gamma^* - n^{-1/2} \theta^*, A^*_n) \) having

\[
\lim_{n \to \infty} E[\Pi^*_n] = \sum_{j=1}^J c_j \theta^*_j - \mathcal{H}(\theta^*)
\]

is asymptotically optimal. Furthermore, under any asymptotically optimal policy \( \alpha \) (in particular, under any policy that is optimal in the \( n \)th system for every \( n \)),

\[
p^*_n \to p^*, \quad \gamma^*_n \to \gamma^*, \quad \text{and} \quad \text{dist}[\sqrt{n} \theta^*_n, \Theta^*] \to 0,
\]

where

\[
\theta^*_n = \sum_{k=1}^K a_{k,j} \lambda_k (p^*_n) - \gamma^*_n, \quad j = 1, \ldots, J.
\]

The notation

\[
\text{dist}[p, S] \equiv \inf_{s \in S} |p - s|
\]

denotes the Euclidean distance between a point \( p \) and a set \( S \).

We devote most of the remainder of this section to proving Theorem 5.1. After its proof, we discuss how to compute \( \mathcal{H}(\theta) \) for a given \( \theta \in \mathbb{N} \). To begin, we establish the following intermediary result that shows prices and component production capacities must converge to the static planning problem solution.

**Proposition 5.1.** Under any admissible policy \( \alpha \) that is asymptotically optimal,

\[
p^*_n \to p^* \quad \text{and} \quad \gamma^*_n \to \gamma^*,
\]

as \( n \to \infty \).

**Proof.** An asymptotically optimal policy \( u \) has

\[
\lim_{n \to \infty} E[\Pi^*_n] \geq \lim_{n \to \infty} E[\Pi_{(p^*, \gamma^*, A^*_n)}]
\]

\[
= -E \left[ \int_0^\infty e^{-\delta t} \left( \sum_{k=1}^K p^*_n \delta q'_n(B(t)) + \sum_{j=1}^J h_j t'_j(B(t)) \right) dt \right] > -\infty,
\]

where the equality follows from Proposition 4.2(b). Dividing both sides of the above inequality by \( \sqrt{n} \) implies

\[
\lim_{n \to \infty} \inf E \left[ \frac{\Pi^*_n}{\sqrt{n}} \right] = \lim_{n \to \infty} \inf E \left[ \frac{\Pi_{(p^*_n, \gamma^*_n, A^*_n)}}{n} - \delta^{-1} \bar{\pi} \right] \geq 0.
\]

The representation for infinite-horizon discounted profit in (3) implies

\[
\frac{\Pi^*_n}{n} - \delta^{-1} \bar{\pi} \leq \int_0^\infty \delta e^{-\delta t} \left( \sum_{k=1}^K p^*_n \frac{\tilde{Q}_{u,k}(t)}{\sqrt{n}} - \sum_{j=1}^J c_j \gamma^*_n - \bar{\pi} \right) dt
\]

\[
+ \delta^{-1} \left( \sum_{k=1}^K p^*_n \lambda_k (p^*_n) - \sum_{j=1}^J c_j \gamma^*_n - \bar{\pi} \right)
\]

\[
- \int_0^\infty \delta e^{-\delta t} \left( \sum_{k=1}^K p^*_n \frac{Q_{u,k}(t)}{n} \right) dt.
\]
Define

\[ f(S, p) \equiv \min_Q \sum_{k=1}^{K} p_k Q_k \]

subject to:

\[ \sum_{k=1}^{K} a_{kj} Q_k \geq [S_j]^+, \quad j = 1, \ldots, J, \]

\[ Q_k \geq 0, \quad k = 1, \ldots, K. \]

The representation for the shortage process in (11) and the fact that queues are nonnegative imply

\[ \sum_{k=1}^{K} a_{kj} Q^n_k(t) / n \geq [S^n_j(t) / n]^+, \quad j = 1, \ldots, J, \]

for all \( t > 0 \). Therefore, because the functional strong law of large numbers (FSLLN; see, for example, Theorem 5.10 in Chen and Yao [11]) implies

\[ \frac{\hat{\sigma}_{n,k}}{\sqrt{n}} \rightarrow 0, \quad k = 1, \ldots, K \quad \text{and} \quad \frac{\tilde{C}_{n,j}}{\sqrt{n}} \rightarrow 0, \quad j = 1, \ldots, J, \quad \text{a.s.,} \]

uniformly on compact sets, as \( n \rightarrow \infty \), by (40),

\[ \liminf_{n \rightarrow \infty} \left( \frac{\Pi^n(p^*, \gamma^*, \lambda^*)}{n} - \delta^{-1} \pi \right) \leq \liminf_{n \rightarrow \infty} \left[ S^{-1} \left( \sum_{k=1}^{K} p^n_k \lambda_k(p^n_k - \gamma^n_{a, j}) - \int_0^\infty \delta e^{-\delta t} f \left( \frac{S^n(t)}{n}, p^n_u \right) dt \right) \right] \]

By the FSLLN, for any \( t > 0 \),

\[ \sup_{0 \leq s \leq t} \max_{j=1, \ldots, J} \left| n^{-1} S^n_{u,j}(s) - \left( \sum_{k=1}^{K} a_{kj} \lambda_k^* (p^*_u - \gamma^*_{a,j}) \right) s \right| \rightarrow 0, \quad \text{a.s.,} \]

and so, also a.s.,

\[ \liminf_{n \rightarrow \infty} \left( \frac{\Pi^n(p^*, \gamma^*, \lambda^*)}{n} - \delta^{-1} \pi \right) \leq \liminf_{n \rightarrow \infty} \left\{ \begin{array}{l}
\min_{q} \delta^{-1} \left( \sum_{k=1}^{K} p^n_k \left( \lambda(p^n_k) - q_k \right) - \sum_{j=1}^{J} c_j \gamma^n_{a, j} \right)
\end{array} \right\} \]

subject to:

\[ \sum_{k=1}^{K} a_{kj} q_k \geq [\sum_{k=1}^{K} a_{kj} \lambda_k^* (p^n_u - \gamma^n_{a,j})]^+, \quad j = 1, \ldots, J, \]

\[ q_k \geq 0, \quad k = 1, \ldots, K, \]

By comparing the linear program in (41) with the static planning problem (5)–(6), we conclude that (41) is nonpositive. Furthermore, because \((p^*, \gamma^*)\) is the unique optimal solution of the static planning problem, (41) equals zero if and only if

\[ p^n_u \rightarrow p^* \quad \text{and} \quad \gamma^n_{a,u} \rightarrow \gamma^* \quad \text{as} \quad n \rightarrow \infty. \]

(42)

Taking expected values of both sides of (41) and observing that (39) lower bounds

\[ \liminf_{n \rightarrow \infty} E \left[ \frac{\Pi^n(p^*, \gamma^*, \lambda^*)}{n} - \delta^{-1} \pi \right] \]

by 0, we conclude any asymptotically optimal policy must satisfy (42). \( \square \)

Proof of Theorem 5.1. We first obtain an upper bound on limiting expected infinite-horizon discounted profit, centered and scaled, under any admissible policy \( u \). We then show that the policy \( u^* \) achieves this upper...
Proposition 5.1 implies that we can assume the policy \( u \) has \( p_n^u \to p^* \) and \( \gamma_n^u \to \gamma^* \) as \( n \to \infty \). Observe from the representation for infinite-horizon discounted profit in (3) that

\[
\tilde{\Pi}_n = \int_0^\infty e^{-\delta t} \left( \sum_{k=1}^K p_{n,k}^u \bar{C}_{u,k}(t) - \sum_{j=1}^J c_j \bar{C}_{u,j}(t) \right) dt + \sqrt{n} \delta^{-1} \left( \sum_{k=1}^K \lambda_k (p_{n,k}^u) p_{n,k}^u - \sum_{j=1}^J c_j \gamma_{u,j}^n - \bar{\pi} \right) - \int_0^\infty e^{-\delta t} \left( \sum_{k=1}^K \delta p_{n,k}^u \frac{Q_{u,k}(t)}{\sqrt{n}} dt + \sum_{j=1}^J h_j \frac{f_{u,j}(t)}{\sqrt{n}} \right) dt.
\]

Because the modified static planning problem upper bounds the maximum revenue of a system having capacity imbalance \( \theta_n^u \),

\[
\sum_{k=1}^K \lambda_k (p_{n,k}^u) p_{n,k}^u - \sum_{j=1}^J c_j \gamma_{u,j}^n - \bar{\pi} \leq \bar{\pi}(\theta_n^u) - \bar{\pi} = \sum_{j=1}^J c_j \theta_{n,j}^u,
\]

where the equality follows from Lemma 3.1. Combining (43) and (44), and applying Lemma 4.2, shows

\[
\limsup_{n \to \infty} E[\tilde{\Pi}_n^u] \leq \limsup_{n \to \infty} \left( \delta^{-1} \sqrt{n} \sum_{j=1}^J c_j \theta_{n,j}^u - E \left[ \int_0^\infty e^{-\delta t} \left( \delta \sum_{k=1}^K p_{n,k}^u \frac{Q_{u,k}(t)}{\sqrt{n}} + \sum_{j=1}^J h_j \frac{f_{u,j}(t)}{\sqrt{n}} \right) dt \right] \right). \tag{45}
\]

Consider any subsequence \( n_i \) having \( \sqrt{n_i} \theta_{n_i,j}^u \to \theta \) in \( \mathcal{Y} \), as \( n_i \to \infty \). On this subsequence, as in Equation (34),

\[
\lim_{n_i \to \infty} \left( \delta^{-1} \sqrt{n_i} \sum_{j=1}^J c_j \theta_{n_i,j}^u - E \left[ \int_0^\infty e^{-\delta t} \left( \delta \sum_{k=1}^K p_{n_i,k}^u \frac{Q_{u,k}(t)}{\sqrt{n_i}} + \sum_{j=1}^J h_j \frac{f_{u,j}(t)}{\sqrt{n_i}} \right) dt \right] \right) = \delta^{-1} \sum_{j=1}^J c_j \theta_j - \mathcal{Y}(\theta). \tag{46}
\]

Next, consider any subsequence \( n_i \) having \( \sqrt{n_i} \theta_{n_i,j}^u \to -\infty \) for at least one \( j \in \{1, \ldots, J \} \) and on which

\[
\limsup_{n_i \to \infty} \sqrt{n_i} \theta_{n_i,m}^u < \infty \quad \text{for all } m \neq j.
\]

Then, as \( n_i \to \infty \),

\[
\delta^{-1} \sqrt{n_i} \sum_{j=1}^J c_j \theta_{n_i,j}^u - E \left[ \int_0^\infty e^{-\delta t} \left( \delta \sum_{k=1}^K p_{n_i,k}^u \frac{Q_{u,k}(t)}{\sqrt{n_i}} + \sum_{j=1}^J h_j \frac{f_{u,j}(t)}{\sqrt{n_i}} \right) dt \right] \to -\infty. \tag{47}
\]

Finally, consider any subsequence \( n_i \) having, as \( n_i \to \infty \), \( \sqrt{n_i} \theta_{n_i,j}^u \to \infty \) for all \( j \in T_u \subset \{1, \ldots, J\} \), \( \sqrt{n_i} \theta_{n_i,j}^u \to -\infty \) for all \( j \in T_u \subset \{1, \ldots, J\} \) and \( \sqrt{n_i} \theta_{n_i,j}^u \to d_j \), a constant for all other \( j \). Let \( \epsilon \) be such that \( p_{n_i,k}^u > \sum_{j \in T_u} (c_j + \epsilon) a_{kj} \) for each \( k = 1, \ldots, K \). (Lemma 3.1 guarantees the existence of such an \( \epsilon \).) Then, observe that

\[
\delta^{-1} \sqrt{n_i} \sum_{j=1}^J c_j \theta_{n_i,j}^u - E \left[ \int_0^\infty e^{-\delta t} \left( \delta \sum_{k=1}^K p_{n_i,k}^u \frac{Q_{u,k}(t)}{\sqrt{n_i}} + \sum_{j=1}^J h_j \frac{f_{u,j}(t)}{\sqrt{n_i}} \right) dt \right]
\leq \delta^{-1} \sqrt{n_i} \sum_{j=1}^J c_j \theta_{n_i,j}^u - E \left[ \int_0^\infty e^{-\delta t} \delta \sum_{k=1}^K \sum_{j \in T_u} (c_j + \epsilon) a_{kj} E \left[ \frac{Q_{u,k}(t)}{\sqrt{n_i}} \right] dt \right]
\leq \delta^{-1} \sqrt{n_i} \sum_{j=1}^J c_j \theta_{n_i,j}^u - E \left[ \int_0^\infty e^{-\delta t} \delta \sum_{j \in T_u} (c_j + \epsilon) E \left[ \frac{Q_{u,j}(t)}{\sqrt{n_i}} \right] dt \right]
= -\delta^{-1} \sqrt{n_i} \epsilon \sum_{j \in T_u} \theta_{n_i,j}^u + \delta^{-1} \sqrt{n_i} \sum_{j \in T_u} c_j \theta_{n_i,j}^u + \delta^{-1} \sqrt{n_i} \sum_{j \notin (T_u \cup T_{\infty})} c_j \theta_{n_i,j}^u
- \int_0^\infty e^{-\delta t} \sum_{j \in T_u} (c_j + \epsilon) E \left[ \sum_{k=1}^K a_{kj} \bar{C}_{u,k}(t) - \bar{C}_{u,j}(t) \right] dt
\to -\infty. \tag{51}
\]
as \( n_i \to \infty \). The bound in (48) follows from our choice of \( \epsilon \), and the fact that inventory levels \( I_{n_i}^\alpha \) are positive for every \( n_i \). To see (49), observe from the shortage process representation in (11) that

\[
- \sum_{k=1}^{K} a_k Q_{n_i,k}^\alpha (t) = -S_{n_i,j}^\alpha (t) - I_{n_i,j}^\alpha (t) \leq -S_{n_i,j}^\alpha (t)
\]

for all \( t > 0 \), every \( n_i \), and any policy \( u \). Substituting the expression for \( n_i^{-1/2} S_{n_i,j}^\alpha (t) \) in (32) yields the equality (50). Finally, Lemma 4.2 and our assumption on the limiting behavior of \( \sqrt{n_i} \theta_n^\alpha \) shows the limiting behavior in (51). We conclude from (43), (46), (47), and the sequence of inequalities in (48)–(51) that

\[
\limsup_{n \to \infty} E[\Pi_n^\alpha] \leq \sup_{\theta \in \Theta} \left( \delta^{-1} \sum_{j=1}^{J} c_j \theta_j - \mathcal{H} (\theta) \right)
\]

for some \( \theta^* \in \Theta^* \), which is nonempty by Lemma 5.1. By Corollary 4.1, the policy \( \star \) achieves the upper bound in (52).

To conclude the proof, because by Proposition 5.1 any admissible policy \( \alpha \) that is asymptotically optimal has \( p_n^\alpha \to p^* \) and \( \gamma_n^\alpha \to \gamma^* \), it is enough to establish dist\( [\sqrt{n_i} \theta_n^\alpha, \Theta] \to 0 \) as \( n \to \infty \), which we do by contradiction. If the policy \( \alpha \) contains a subsequence on which \( \sqrt{n_i} \theta_n^\alpha \to \infty \), from (47) and (48)–(51), \( \liminf_{n \to \infty} E[\Pi_n^\alpha] \to -\infty \), and so the policy \( \alpha \) cannot be asymptotically optimal. Otherwise, on any subsequence \( n_i \) having \( \sqrt{n_i} \theta_n^\alpha \to \theta \not\in \Theta^* \), from (43), (44), and (46),

\[
\lim_{n_i \to \infty} E[\Pi_n^\alpha] \leq \delta^{-1} \sum_{j=1}^{J} c_j \theta_j^* - \mathcal{H} (\theta^*)
\]

which implies

\[
\limsup_{n \to \infty} E[\Pi_n^\alpha] < \delta^{-1} \sum_{j=1}^{J} c_j \theta_j^* - \mathcal{H} (\theta^*) = \lim_{n \to \infty} E[\Pi_n^\alpha],
\]

and so again the policy \( \alpha \) cannot be asymptotically optimal.

If policy \( \alpha \) is optimal in the \( n \)th system for every \( n \), then policy \( \alpha \) is admissible and \( E[\Pi_n^\alpha] \geq E[\Pi_n^\star] \) for any other admissible policy \( u \), for every \( n \). The asymptotically optimal policy \( \star \) is admissible, so

\[
\liminf_{n \to \infty} E[\Pi_n^\alpha] \geq \liminf_{n \to \infty} E[\Pi_n^\star] \geq \limsup_{n \to \infty} E[\Pi_n^\star]
\]

for any admissible policy \( u \), which establishes that the optimal policy \( \alpha \) is asymptotically optimal. \( \square \)

The key to implementing our proposed policy is to find a \( \theta^* \in \Theta^* \), which requires computing \( E[\mathcal{H} (\theta)] \). This expected value can be computed by numerically solving a partial differential equation, as the following lemma (whose proof is given in the appendix) establishes.

**Lemma 5.2.** Let \( L_\theta \) be the generator of the diffusion \( B_\theta \) so that

\[
L_\theta \equiv \frac{1}{2} \sum_{j=1}^{J} \sum_{m=1}^{M} \Gamma_{m} \frac{\partial^2}{\partial x_j \partial x_m} + \sum_{j=1}^{J} \theta_j \frac{\partial}{\partial x_j}.
\]

Suppose the function \( f \) is twice continuously differentiable on \( \mathbb{R}^J \), has bounded \( f, f' \), and \( f'' \), and solves

\[
L_\theta f_\theta - \delta f_\theta + \delta \sum_{k=1}^{K} p_k^\star q_k + \sum_{j=1}^{J} h_j^\star = 0.
\]

Then,

\[
f_\theta (0) = \mathcal{H} (\theta).
\]

**6. Conclusions and future research.** For a high-volume assemble-to-order system, optimal pricing and capacity investment decisions result in utilization of component production capacity near 100%. In this “heavy-traffic” regime, the inventory position for each component evolves according to a Brownian motion. Optimal dynamic allocation of scarce components to outstanding orders (equivalently, optimal sequencing of orders for
assembly) forces order queues and component inventory levels into a configuration that minimizes instantaneous physical and financial holding cost. In other words, the system exhibits state-space collapse.

In reality, customers are impatient, and want to know their waiting time before placing an order. In response, most ATO manufacturers guarantee that orders will be filled within some specified maximum delay. Our analysis suggests that delays are short compared to order arrival rates under the policy \( \star = (p^*, \gamma^* - n^{-1/2}\theta^*, A^n) \). To see this, observe that when order arrival rates \( \lambda_1, \ldots, \lambda_K \) are large, a sample-path version of Little’s law, analogous to Reiman’s “snapshot principle” [33], shows that for all \( t \geq 0 \),

\[
D^*_k(t) \approx \frac{Q^*_k(t)}{\lambda^*_k},
\]

where \( D^*_k(t) \) is the delay in the \( n \)th system associated with filling a product \( k \) order that arrives at time \( t \), and \( \approx^D \) denotes approximately equal in distribution. Because Corollary 4.1 shows that under the policy \( \star \) queue lengths are of order \( \sqrt{n} \), the approximation (55) implies that delays are of order \( n^{-1/2} \) when order arrival rates \( \lambda_1, \ldots, \lambda_K \) are of order \( n \). Hence, it will be economical to quote a maximum delay of order \( n^{-1/2} \).

Our follow-up paper (Plambeck and Ward [31]) considers an ATO system having delay constraints. To ensure that the system manager can fill all product orders quickly enough, we allow him to expedite components at a high per-unit cost. We provide an asymptotically optimal policy for setting prices, component production capacity, sequencing orders for assembly, and expediting components. Additionally, we allow the system manager to sell excess components for a low “salvage” value when inventories grow large, then heuristically derive and numerically solve an approximating diffusion control problem that motivates a policy for expediting and salvaging components. The state-space collapse demonstrated in this paper greatly simplifies the inventory policy: One need not track the number of outstanding orders for every product, only the component inventory positions.

We have assumed (both in this paper and Plambeck and Ward [31]) that when the system manager invests in component production capacity, he knows the mean arrival rate for customer orders as a function of prices. In practice, capacity investment occurs under considerable uncertainty regarding future demand rates. Further research is needed to address dynamic pricing in response to shocks in the order arrival rate. Although closed-form solutions are rare in dynamic pricing models (the only papers we know of that obtain closed-form dynamic pricing strategies are Xu and Hopp [40] and Gallego and van Ryzin [18]), an asymptotic analysis in a high-volume regime may provide some insight. We have also taken a monopolist’s perspective. Research is needed on capacity, pricing, and inventory management in an assemble-to-order system with oligopolistic competition.

An essential assumption for our analysis is that the assembly operation is not capacitated. This is a reasonable assumption in consumer electronics and PC manufacturing. However, automobile assembly is capacitated and yet automobile manufacturers are straining to assemble to order (Economist [15], Forbes [17]). The potential increase in expected profit has been estimated at $80 billion per year for the auto industry as a whole (Economist [14], Agrawal et al. [3]). Further research is needed to determine how to optimally control systems with capacitated assembly.

**Appendix A. Lemma proofs.**

**Proof of Lemma 2.1.** Because \( \lambda(p) \) is continuously differentiable and its Jacobian matrix \( J = [\frac{\partial \lambda_k}{\partial p_m}]_{k,m=1,\ldots,K} \) is invertible everywhere, the inverse function theorem guarantees that the inverse demand function \( p(\lambda) \) is unique, continuous, and differentiable, and has Jacobian matrix

\[
J^{-1} = \left[ \frac{\partial p_k}{\partial \lambda_m} \right]_{k,m=1,\ldots,K}.
\]

Define the \( K \times K \) matrices:

\[
T = \begin{bmatrix}
0 & \frac{\partial \lambda_1}{\partial p_1} & \frac{\partial \lambda_1}{\partial p_2} & \cdots & \frac{\partial \lambda_1}{\partial p_K} \\
-\frac{\partial \lambda_2}{\partial p_1} & 0 & \frac{\partial \lambda_2}{\partial p_2} & \cdots & \frac{\partial \lambda_2}{\partial p_K} \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
-\frac{\partial \lambda_N}{\partial p_1} & \frac{\partial \lambda_N}{\partial p_2} & \cdots & \frac{\partial \lambda_N}{\partial p_K}
\end{bmatrix},
\]

\[
\Lambda = \text{diag} \left( \frac{\partial \lambda_k}{\partial p_k} \right)_{k=1,\ldots,K},
\]

and the diagonal matrix:

\[
\Lambda = \text{diag} \left( \frac{\partial \lambda_k}{\partial p_k} \right)_{k=1,\ldots,K}.
\]
and observe that
\[ J = \Lambda(I - T). \]

Therefore, \( T \) is a substochastic matrix because (4) guarantees that all of its elements are nonnegative and its row sums are less than one. Therefore,
\[ (I - T)^{-1} = \sum_{i=0}^{\infty} T^i \]
and \( (I - T)^{-1} \) has all nonnegative elements. Furthermore,
\[ J^{-1} = (I - T)^{-1} \Lambda^{-1}, \]
where \( \Lambda^{-1} \) is a diagonal matrix with strictly negative elements. Specifically, \( \Lambda^{-1} = \text{diag}((\partial \lambda_k / \partial p_k)^{-1})_{k=1}^{K} \) where \( \partial \lambda_k / \partial p_k < 0 \) by assumption. We can therefore conclude that all elements of \( J^{-1} \) are negative, and all diagonal elements of \( J^{-1} \) are strictly negative, which completes the proof.

**Proof of Lemma 3.1.** Because the demand function \( \lambda \) has a unique inverse demand function \( p \), it is equivalent to formulate the perturbed static planning problem as a maximization over order arrival and component production capacities as follows:
\[ \bar{\pi}(\theta) = \max_{\lambda \geq 0, \gamma \geq 0} r(\lambda) - \sum_{j=1}^{J} \gamma_j c_j \tag{56} \]
subject to
\[ \sum_{k=1}^{K} a_{kj} \lambda_k \leq \gamma_j + \theta_j, \quad j = 1, \ldots, J. \tag{57} \]

Our assumption that \( r(\lambda) \) is strictly concave guarantees that (56)–(57) is a convex program, and so has a unique optimal solution for all \( \theta \in \mathbb{R}^{J} \). The Lagrangian associated with (56)–(57), when equivalently viewed as a minimization problem, is
\[ \mathcal{L}(\lambda, \gamma, u) = -r(\lambda) + \sum_{j=1}^{J} c_j \gamma_j + \sum_{j=1}^{J} u_j \left( \sum_{k=1}^{K} a_{kj} \lambda_k - \gamma_j - \theta_j \right), \]
which yields the Karush-Kuhn-Tucker (KKT) conditions
\[ \frac{\partial \mathcal{L}}{\partial \lambda_m} = -\frac{\partial r(\lambda)}{\partial \lambda_m} + \sum_{j=1}^{J} a_{mj} u_j = 0, \quad m = 1, \ldots, K, \tag{58} \]
\[ \frac{\partial \mathcal{L}}{\partial \gamma_j} = c_j - u_j = 0, \tag{59} \]
\[ \sum_{k=1}^{K} a_{kj} \lambda_k - \gamma_j - \theta_j \leq 0, \quad j = 1, \ldots, J, \tag{60} \]
\[ u_j \left( \sum_{k=1}^{K} a_{kj} \lambda_k - \gamma_j - \theta_j \right) = 0, \quad j = 1, \ldots, J, \tag{61} \]
\[ u \geq 0. \tag{62} \]

Because \( u_j = c_j > 0 \) by (59), and our assumption that component production costs are positive, from (58),
\[ 0 < \sum_{j=1}^{J} a_{mj} c_j = \frac{\partial r(\lambda)}{\partial \lambda_m} = \sum_{k=1}^{K} \lambda_k \frac{\partial p_k(\lambda)}{\partial \lambda_m} + p_m(\lambda), \]
and so \( p_m(\lambda) > \sum_{j=1}^{J} a_{mj} c_j \) because, by Lemma 2.1, \( \partial p_k(\lambda) / \partial \lambda_m \leq 0 \). By (60), when \( \theta = 0 \),
\[ \gamma^* = \sum_{k=1}^{K} a_{kj} \lambda_k^* > 0, \]
because \( p(\lambda) > 0 \) implies \( \lambda(p) > 0 \). Next, for \( \theta > 0 \), observe that conditions (58)–(62) are satisfied when \( \lambda^*(\theta) = \lambda^* \) and \( \gamma^*(\theta) = \gamma^* - \theta \), \( j = 1, \ldots, J \). Hence, the KKT theorem ensures that \( (\lambda^*(\theta), \gamma^*(\theta)) \) is a global maximizing point for (56)–(57), which implies that \( (p^*(\theta) = p(\lambda^*), \gamma^*(\theta)) \) is a global maximizing point for the original perturbed problem in (8)–(9), and
\[ \bar{\pi}(\theta) - \bar{\pi} = \sum_{j=1}^{J} c_j \theta_j. \]
Proof of Lemma 4.1. We have assumed that \( \mathbb{E}|x_1(1)|^{2+\varepsilon} < \infty, k = 1, \ldots, K \) and \( \mathbb{E}|y_j(1)|^{2+\varepsilon} < \infty, j = 1, \ldots, J \) for \( \varepsilon = 3 \). Therefore, from Lemma 9 and bound (113) on the residual interarrival time at the start of a review period in Ata and Kumar [4], for any finite constant \( \alpha > 0 \) there exists a constant \( C(\alpha) \) such that for \( nl^n \geq 2/\alpha \),

\[
P \left( \max_{i=0,1,\ldots,n-1} \max_{k=1,\ldots,K} |O^p_k((i+1)l^n) - O^p_k(il^n) - \lambda_k(p^n)nl^n| \geq \alpha nl^n \right) \leq I^n \frac{C(\alpha)}{(nl^n)^{1+\varepsilon/2}} \leq C(\alpha)\alpha^{3/2}n^{-1/6}.
\]

Setting \( \beta = \max(C(\alpha)^{-1/6} |\lambda^*|^{2/5}, (2/\alpha)^{1/2}) \) establishes (19). Similar arguments yield (20).

Proof of Lemma 4.2. For brevity, we suppress the subscript indicating the policy \( u \). We first show uniform integrability, and then argue the convergence to zero. Observe that for any \( k = 1, \ldots, K \) and \( t \geq 0 \),

\[
\frac{O^p_k(t) - nl^n_k t}{\sqrt{n}} = -M_k(O^p_k(t) + 1) + \frac{\mathcal{E}_k(O^p_k(t) + 1) - n\lambda^*_k t}{\sqrt{n}} - \frac{1}{\sqrt{n}}, \quad (63)
\]

where \( M_k(m) = \mathcal{E}_k(m) - m \) is a square-integrable martingale relative to the filtration \( \sigma(x_1(1), x_2(2), \ldots, x_k(m)) \), and \( \lambda^*_k = \lambda_k(p^n) \rightarrow \lambda^*_k \) as \( n \rightarrow \infty \). Furthermore, bounding the residual interarrival time at \( t \) with the full interarrival time encompassing \( t \) yields

\[
\left| \mathcal{E}_k(O^p_k(t) + 1) - n\lambda^*_k (p^n)t \right| \leq x_k(O^p_k(t) + 1) \frac{1}{\sqrt{n}} = M_k(O^p_k(t) + 1) - M_k(O^p_k(t)) + \frac{1}{\sqrt{n}} \leq 2 \sup_{0 \leq s \leq t} \left| M_k(O^p_k(s) + 1) \right| + \frac{1}{\sqrt{n}}. \quad (64)
\]

Now use (64) to bound (63):

\[
\frac{O^p_k(t) - nl^n_k t}{\sqrt{n}} \leq 3 \sup_{0 \leq s \leq t} \left| M_k(O^p_k(s) + 1) \right| + \frac{2}{\sqrt{n}},
\]

and observe that for \( n > 1 \),

\[
\left( \sup_{0 \leq s \leq t} \frac{O^p_k(s) - nl^n_k (p^n)s}{\sqrt{n}} \right)^2 \leq \left( 3 \sup_{0 \leq s \leq t} \left| M_k(O^p_k(s) + 1) \right| + \frac{2}{\sqrt{n}} \right)^2 \leq 18 \sup_{0 \leq s \leq t} \left| M_k(O^p_k(s) + 1) \right|^2 + \frac{8}{n}. \quad (65)
\]

Because\(^5\)

\[
\mathbb{E} \left( \sup_{0 \leq s \leq t} \left| M_k(O^p_k(s) + 1) \right|^2 \right) \leq 4\mathbb{E}M_k(O^p_k(t) + 1)^2 \quad (L^p \text{ maximum inequality})
\]

\[
\leq 4\mathbb{E}(x_k(1) - 1)^2 E(O^p_k(t) + 1) \quad \text{(Wald’s 2nd moment identity)}
\]

\[
= 4\mathbb{E}(x_k(1) - 1)^2 E[\mathcal{E}_k(O^p_k(t) + 1)] \quad \text{(Wald’s 1st moment identity)}
\]

\[
= 4\mathbb{E}(x_k(1) - 1)^2 \left[ \lambda_k(p^n)nt + (\mathcal{E}_k(O^p_k(t) + 1) - nl^n_k (p^n)t) \right] \leq 4\mathbb{E}(x_k(1) - 1)^2 \left( \lambda_k(p^n)nt + \mathbb{E}x_k(1)^2 \right) \quad \text{(Lorden’s inequality)},
\]

the bound in (65) and the assumption that \( p^n \rightarrow p^* \) as \( n \rightarrow \infty \) implies

\[
\mathbb{E} \left( \sup_{0 \leq s \leq t} \left| O^p_k(s) - nl^n_k (p^n)s \right| \right)^2 \leq \kappa_1 t + \kappa_2
\]

\(^5\)Lorden’s inequality is found in Lorden [22] or pages 99–100 in Gut [20].
for large enough \( n \), where \( \kappa_1 \) and \( \kappa_2 \) are finite constants. From Jensen’s inequality,

\[
\left( \int_0^\infty e^{-\delta t} \sup_{0 \leq s \leq t} \frac{\left| O_k^\delta(s) - n \lambda_k(p^\delta)s \right|}{\sqrt{n}} dt \right)^2 \< \int_0^\infty \left( \sup_{0 \leq s \leq t} \frac{\left| O_k^\delta(s) - n \lambda_k(p^\delta)s \right|}{\delta \sqrt{n}} \right)^2 \delta e^{-\delta t} dt,
\]

and so, taking expectations and using Fubini to justify interchanging the expectation and integral, we find

\[
E \left( \int_0^\infty e^{-\delta t} \sup_{0 \leq s \leq t} \frac{\left| O_k^\delta(s) - n \lambda_k(p^\delta)s \right|}{\sqrt{n}} dt \right)^2 
\leq \int_0^\infty \delta e^{-\delta t} E \left( \sup_{0 \leq s \leq t} \frac{\left| O_k^\delta(s) - n \lambda_k(p^\delta)s \right|}{\delta \sqrt{n}} \right)^2 dt
\leq \delta^{-2} (\kappa_1 t + \kappa_2).
\]

Therefore, for any \( k = 1, \ldots, K \), the family

\[
\left\{ \int_0^\infty e^{-\delta t} \sup_{0 \leq s \leq t} |\tilde{O}_k^\delta(s)| dt \right\}
\]

is uniformly integrable.

For any \( k = 1, \ldots, K \) and \( B \) a standard Brownian motion, the functional central limit theorem and the continuous-mapping theorem show

\[
\int_0^\infty p_k \delta e^{-\delta t} \frac{O_k^\delta(t) - n \lambda_k^* t}{\sqrt{n}} dt \Rightarrow \int_0^\infty p_k \delta e^{-\delta t} \sqrt{\lambda_k^* \sigma^2_{0,k}} B(t) dt,
\]
as \( n \to \infty \). Because we have just established uniform integrability,

\[
E \left[ \int_0^\infty p_k \delta e^{-\delta t} \frac{O_k^\delta(t) - n \lambda_k^* t}{\sqrt{n}} dt \right] \to E \left[ \int_0^\infty p_k \delta e^{-\delta t} \sqrt{\lambda_k^* \sigma^2_{0,k}} B(t) dt \right],
\]
as \( n \to \infty \). Observe that because \( e^{-\delta t}|B(t)| \geq 0 \) for all \( t \geq 0 \) and \( B(t) \) has a normal distribution with mean 0 and variance \( t \),

\[
E \left[ \int_0^\infty e^{-\delta t} |B(t)| dt \right] = \int_0^\infty e^{-\delta t} E|B(t)| dt = \int_0^\infty e^{-\delta t} \sqrt{\frac{2t}{\pi}} dt < \infty.
\]

Therefore, Fubini justifies the interchange necessary to conclude

\[
E \int_0^\infty p_k \delta e^{-\delta t} \sqrt{\lambda_k^* \sigma^2_{0,k}} B(t) dt = 0.
\]

Identical arguments establish both that

\[
\left\{ \int_0^\infty e^{-\delta t} \sup_{0 \leq s \leq t} |\tilde{C}_j^\delta(s)| dt \right\}
\]
is uniformly integrable for each \( j = 1, \ldots, J \), and that

\[
E \int_0^\infty c_j \delta e^{-\delta t} \sup_{0 \leq s \leq t} \tilde{C}_j^\delta(s) dt \to 0
\]

for each \( j = 1, \ldots, J \). □

**Proof of Lemma 5.1.** We will show that

\[
\left\{ \theta \in \mathcal{R} : \delta^{-1} \sum_{j=1}^J c_j \theta_j - \mathcal{H}(\theta) \geq -\mathcal{H}(0) \right\} \subseteq \bigcup_{j=1}^J \left\{ -c_j^{-1} \delta \mathcal{H}(0) \left( 1 + \max_{l \neq j} \{c_l \} / \Delta \right), \delta \mathcal{H}(0) / \Delta \right\}.
\]

(66)

Hence, in maximizing \( \delta^{-1} \sum_{j=1}^J c_j \theta_j - \mathcal{H}(\theta) \), one may restrict attention to the compact set

\[
\bigcup_{j=1}^J \left\{ -c_j^{-1} \delta \mathcal{H}(0) \left( 1 + \max_{l \neq j} \{c_l \} / \Delta \right), \delta \mathcal{H}(0) / \Delta \right\}.
\]
By Fubini’s theorem,
\[
\mathcal{H}(\theta) = \int_0^\infty e^{-\delta t} \mathbb{E} \left[ \sum_{k=1}^K p_k^* q_k^* (B_k(t)) + \sum_{j=1}^J h_j^* (B_j(t)) \right] dt
\]
\[
= \int_0^\infty e^{-\delta t} \int_{x \in \mathbb{R}^d} (2\pi)^{-d/2} |\Gamma_t|^{-1/2} \exp(-x - \theta t^\top (\Gamma_t)^{-1} (x - \theta t)) \left( \sum_{k=1}^K p_k^* q_k^* (x) + \sum_{j=1}^J h_j^* (x) \right) dx dt
\]
because \(B_k(t)\) is a multivariate normal random variable with mean \(\theta t\) and covariance matrix \(\Gamma_t\). Assumption 3.1 guarantees that \(\sum_{k=1}^K p_k^* q_k^* (x) + \sum_{j=1}^J h_j^* (x)\) is Lipschitz continuous in \(x\), and it follows that \(\delta^{-1} \sum_{j=1}^J c_j \theta_j + \mathcal{H}(\theta)\) is a continuous function of \(\theta\). Therefore, \(\delta^{-1} \sum_{j=1}^J c_j \theta_j + \mathcal{H}(\theta)\) achieves its maximum in the compact set
\[
\sum_{j=1}^J c_j \theta_j + \mathcal{H}(\theta) = \frac{c_j^{-1} \mathcal{H}(0)}{1 + \max_{i \neq j} |c_j|/\Delta}, \quad \delta \mathcal{H}(0)/\Delta.
\]
It remains to prove (66). Define
\[
J^+(\theta) = \{ j : \theta_j \geq 0 \},
\]
\[
\Delta(\theta) = \min_{k=1, \ldots, K} \left\{ \left( p_k^* - \sum_{j \in J^+(\theta)} a_k^{c_j} \right) / \left( \sum_{j \in J^+(\theta)} a_k^{c_j} \right) \right\},
\]
\[
p_k(\theta) = \sum_{j \in J^+(\theta)} a_k^{c_j} (c_j + \Delta(\theta)).
\]
It follows from Lemma 1 that \(\Delta(\theta) \geq \Delta > 0\) and \(p_k(\theta) \leq p_k^*\) for all \(\theta \in \mathbb{R}^J\). Furthermore, define
\[
f(B; \theta) = \min_q \sum_{k=1}^K p_k(\theta) q_k
\]
s.t.
\[
\sum_{k=1}^K a_{kj} q_j \geq B_j \quad \text{for} \quad j \in J^+(\theta)
\]
\[
q_k = 0 \quad \text{if} \quad a_{kj} = 0 \quad \text{for all} \quad j \in J^+(\theta).
\]
Observe that for any \(\theta, B \in \mathbb{R}^J\) and \(t \geq 0\),
\[
f(B; \theta) \leq \sum_{k=1}^K p_k^* q_k^* (B) + \sum_{j=1}^J h_j^* (B),
\]
\[
f(tB; \theta) = tf(B; \theta),
\]
and \(f(B; \theta)\) is a convex function of \(B\). Another useful property is that
\[
\sum_{j=1}^J c_j \theta_j - f(\theta; \theta) = \max_{\theta', q} \sum_{j \in J^+(\theta)} c_j \theta'_j - \sum_{k=1}^K p_k(\theta) q_k
\]
s.t.
\[
\sum_{j \in J^+(\theta)} \theta'_j \geq \sum_{j \in J^+(\theta)} \theta_j
\]
\[
\sum_{k=1}^K a_{kj} q_j \geq \theta_j \quad \text{for} \quad j \in J^+(\theta),
\]
\[
q_k = 0 \quad \text{if} \quad a_{kj} = 0 \quad \text{for all} \quad j \in J^+(\theta)
\]
\[
= -\Delta(\theta) \sum_{j=1}^J [\theta_j]^+.\]
Employing these properties of \( f(B, \theta) \) with Fubini’s theorem and Jensen’s inequality, we find that for any \( \theta \in \mathbb{R}^J \),

\[
\begin{align*}
\delta^{-1} \sum_{j=1}^J c_j \theta_j - \mathcal{H}(\theta) &= \delta^{-1} \sum_{j=1}^J c_j \theta_j - \int_0^\infty e^{-\delta t} \mathbb{E} \left[ \sum_{k=1}^K p_k^* q_k^*(B_0(t)) + \sum_{j=1}^J h_j^* (B_0(t)) \right] dt \\
&\leq \delta^{-1} \sum_{j=1}^J c_j \theta_j - \int_0^\infty e^{-\delta t} \mathbb{E} [f(B_0(t); \theta)] dt \\
&\leq \delta^{-1} \sum_{j=1}^J c_j \theta_j - \int_0^\infty e^{-\delta t} f(B_0(t); \theta) dt \\
&= \delta^{-1} \sum_{j=1}^J c_j \theta_j - \int_0^\infty e^{-\delta t} f(\theta; \theta) dt \\
&= \delta^{-1} \sum_{j=1}^J c_j \theta_j - \int_0^\infty e^{-\delta t} f(\theta; \theta) dt \\
&\leq -\delta^{-1} \Delta(\theta) \sum_{j=1}^J [\theta_j]^+.
\end{align*}
\]

Therefore, to achieve \( \delta^{-1} \sum_{j=1}^J c_j \theta_j - \mathcal{H}(\theta) \geq -\mathcal{H}(0) \), \( \theta \) must satisfy

\[
\sum_{j=1}^J [\theta_j]^+ \leq \delta \mathcal{H}(0)/\Delta \leq \delta \mathcal{H}(0)/\Delta,
\]

which implies

\( \theta_j \leq \delta \mathcal{H}(0)/\Delta \) for \( j = 1, \ldots, J \).

Also observe that

\[
\delta^{-1} \sum_{j=1}^J c_j \theta_j - \mathcal{H}(\theta) \leq \delta^{-1} \sum_{j=1}^J c_j \theta_j,
\]

so to achieve \( \delta^{-1} \sum_{j=1}^J c_j \theta_j - \mathcal{H}(\theta) \geq -\mathcal{H}(0) \), \( \theta \) must also satisfy

\[
\begin{align*}
c_j \theta_j &\geq - \sum_{l \neq j} c_l \theta_l - \mathcal{H}(0) \\
&\geq -\delta \mathcal{H}(0) \left( 1 + \max_{l \neq j} [c_j]/\Delta \right),
\end{align*}
\]

which completes the proof of (66). \( \square \)

**Proof of Lemma 5.2.** By Ito’s formula, for any \( t > 0 \),

\[
e^{-\delta t} f(B_0(t)) - f(B_0(0)) \\
= \int_0^t e^{-\delta s} L_0 f(B_0(s)) ds - \int_0^t e^{-\delta s} f(B_0(s)) ds + \sum_{j=1}^J \int_0^t e^{-\delta s} \frac{\partial f(B_0(s))}{\partial x_j} dB_j(s).
\]

Because the stochastic integrals

\[
\int_0^t e^{-\delta s} \frac{\partial f(B_0(s))}{\partial x_j} dB_j(s), \quad j = 1, \ldots, J
\]

are martingales,

\[
\mathbb{E} [e^{-\delta t} f(B_0(t)) - f(B_0(0)) - \int_0^t e^{-\delta s} L_0 f(B_0(s)) ds + \int_0^t e^{-\delta s} f(B_0(s)) ds] = 0.
\]
Letting $t \to \infty$, by the bounded convergence theorem,
\[
E\left[ \int_0^\infty e^{-\delta t} \left( L_0f(B_0(t)) - \delta f(B_0(t)) \right) dt \right] = -f(B_0(0)),
\]
and so
\[
E\left[ \int_0^\infty e^{-\delta t} \left( L_0f(B_0(t)) - \delta f(B_0(t)) + \sum_{k=1}^{K} p_k^* \delta q_k^*(B_0(t)) + \sum_{j=1}^{J} h_j^* j_i^*(B_0(t)) \right) dt \right] = \mathcal{H}(\theta) - f(B_0(0)). \quad (67)
\]
Assumption (53) applied to (67) when $B_0(0) = 0$ establishes (54). \square

Appendix B. Table of notation. In the following, $k = 1, \ldots, K$ and $j = 1, \ldots, J$.

**System parameters**
- $p_k$ Price of product $k$.
- $\lambda_k(p)$ Price-dependent product $k$ order arrival rate.
- $\gamma_j$ Component $j$ production capacity.
- $c_j$ Component cost.
- $h_j$ Holding cost per unit time for component $j$.
- $a_{kj}$ Number of type $j$ components required to assemble product $k$.
- $\delta$ Discount rate.

**Notation associated with the proposed policy**
- $(p^*, \gamma^*)$ Solution to the static planning problem in (5)–(6).
- $(p^*(\theta), \gamma^*(\theta))$ Solution to the perturbed static planning problem in (8)–(9).
- $\theta^* = \arg \max (\delta^{-1} \sum_{j=1}^{J} c_j \theta_j - \mathcal{H}(\theta))$ Stochastic imbalance cost, where $\mathcal{H}(\theta)$ is defined in (35).
- $A^*$ Assembly policy defined in (12).
- $(q^*, l^*)$ Profit-maximizing queue lengths and inventory levels, defined as the solution to the linear program in (13)–(14).
- $l \equiv (1/|\lambda|^*)^{2/3}$ Review period length.

**Acknowledgments.** The authors thank Li Chen, Mike Harrison, Anton Kleywegt, Sunil Kumar, Joshua Reed, Alex Shapiro, and Ruth Williams for helpful discussions relating to the issues presented in this paper. They especially acknowledge the anonymous referees for their explicit instructions on how to justify the existence of a Lipschitz continuous function $Q^*$. This research was supported in part by the National Science Foundation under Grant DMI-0239840.

**References**