On the performance of sparse process structures in partial postponement production systems

Chou, Mabel C.; Chua, Geoffrey A.; Zheng, Huan

2014

Chou, M. C., Chua, G. A., & Zheng, H. (2014). On the performance of sparse process structures in partial postponement production systems. Operations research, 62(2), 348-365.

https://hdl.handle.net/10356/106715

https://doi.org/10.1287/opre.2013.1255

Submitted to *Operations Research*

manuscript OPRE-2011-06-304.R2

On the Performance of Sparse Process Structures in Partial Postponement Production Systems

Mabel C. Chou

NUS Business School, National University of Singapore, Singapore. Email: bizchoum@nus.edu.sg

Geoffrey A. Chua

Nanyang Business School, Nanyang Technological University, Singapore. Email: gbachua@ntu.edu.sg

Huan Zheng

Antai College of Economics and Management, Shanghai Jiaotong University, China. Email: zhenghuan@sjtu.edu.cn

Production postponement, the strategy to hold reserved production capacity that can be deployed based on actual demand signals, is often used to mitigate supply-demand mismatch risk. The effectiveness of this strategy depends crucially on the ease, or flexibility, in deploying the reserved capacity to meet product demands. Existing literature assumes that the reserved capacity is "fully flexible," i.e. capable of being deployed to meet the demand of any item in a multi-product system. Little is known if reserved capacity is held at many different locations, with each location having only a limited range of flexibility on production options. This paper examines how effective the production postponement strategy is in this environment.

When the amount of reserved capacity is small (i.e. postponement level near 0%), no amount of flexibility can reap significant benefits. When the reserved capacity is high (i.e. postponement level near 100%), it is well known that a sparse structure such as a 2-chain can perform nearly as well as a fully flexible structure. Hence, process flexibility beyond 2-chain has little impact on the effectiveness of production postponement strategy in these two extreme environments. Interestingly, in a symmetric system, we prove that the performance of 2-chain, vis-a-viz the full flexibility structure, has a wider gap when postponement level (i.e. amount of reserved capacity) is moderate, and thus process flexibility beyond 2-chain matters and affects appreciably the performance of the production postponement strategy. Fortunately, adding a little more flexibility, say turning a 2-chain into a 3-chain, the system can perform almost as well as a full flexibility structure for all postponement levels. This is important as first stage production capacity can be allocated "as if" the reserve capacity is "fully flexible." Our analysis hinges on an exact analytical expression for the performance of d-chain, obtained from solving a related class of random walk problems. To the best of our knowledge, this is the first paper with analytical results on the performance of d-chain for $d > 2$.

Key words: process flexibility, production postponement, chaining strategy, multi-item newsvendor, stochastic programming

1. Introduction

Since the 1980s, we have witnessed the advent of globalization and the tremendous effects it has on world consumption and production patterns. A quick look at Interbrand's[1] 2011 rankings of the 100 best global brands reveals that these brands already hail from fifteen different countries, up from thirteen in 2009. According to the report, each of these brands derives at least a third of its earnings outside its home country. This tells us that the world is increasingly moving towards a phenomenon of borderless consumption. With the internationalization of market competition, firms nowadays need to build up the capacity to stay competitive as a world-class company. The

most common solution has been to turn to outsourcing and offshoring, essentially tapping into the production capabilities of factories, big and small, all over the world. For example, many American and European brands outsource their sourcing function to Hong Kong-based Li & Fung, one of the world's leading supply chain companies, which controls a network of over 10,000 production facilities scattered everywhere in places like China, Brazil, the Czech Republic, Honduras, Mauritius, Mexico, Poland, South Africa, Zimbabwe, and countries in Southeast Asia (Feng 2007). On this phenomenon of borderless manufacturing, Fung et al. (2007, 2008) believe the trend is "to rip the roof off the factory. In contrast to Henry Ford's assembly line, where all the manufacturing processes were under one roof, the entire world is our factory." Other than granting firms the ability to increase capacity through global aggregation, this strategy also allows the firms to control and reduce operating expenses as well as focus on improving their core businesses, such as product design and marketing.

Another important trend is the fragmentation of consumer demand. Instead of catering to one big market with more or less homogeneous demand, companies are beginning to see more niche markets with diverse tastes as well as the emergence of variety-seeking consumer behavior. As this trend becomes more prevalent, we see an increasing proliferation of product lines as companies struggle to stay competitive. In the automobile industry, the number of car models offered in the United States market has experienced an upward trend since 1984 (Van Biesebroeck 2007). The same phenomenon can be observed in other industries such as electronics, clothing, food products, and even services like entertainment/media and education. As a result, demand uncertainty on a per product basis increases and forecasting becomes more difficult.

Cast in the overall setting of globalization, the increased demand uncertainty confronting manufacturers is further heightened by the complexity of the global production and consumption network. Facing a growing number of facilities and products, firms now need to contend with not only uncertainty, but multiple sources of uncertainty. The challenges here are two-fold; namely, forecasting and production planning. Forecasting becomes more difficult due to disaggregation and accounting for correlations while production planning in a multi-plant, multi-product system entails both network design and production allocation.

To deal with multiple sources of demand uncertainty, the literature suggests two approaches that come hand in hand; namely, (1) process flexibility and (2) production postponement. "Process flexibility" refers to "a firm's ability to provide varying goods or services, using different facilities or resources". The more products each plant is capable of producing, the more flexible is the production system. On the other hand, we interpret "production postponement" as "the proportion of a firm's capacity that can be used to satisfy demand immediately". Since the firm can convert

this postponed capacity into fulfilled demand quickly, the allocation of this capacity can therefore be chosen on a make-to-order basis (after receiving demand information). However, the remaining proportion that cannot be used for immediate fulfillment of demand must be deployed in a make-to-stock fashion (prior to receiving demand information). The more capacity falls under make-to-order, the more production postponement the firm is said to possess. Process flexibility generates value through risk pooling (Eppen 1979), whereas production postponement creates value from the option not to produce and when coupled with the former, the option of what to produce. When there is no postponement, the benefits from risk pooling are lost. On the other hand, when there is no flexibility, postponement only eliminates the cost of overage and nothing else. Hence, it is important to carefully choose the mix of process flexibility and production postponement.

To illustrate, consider a firm that owns a network of several plants whose capacities can be used to meet the (expected) demands for a range of products (i.e. a balanced system). The firm must choose at what levels to deploy the twin approaches of flexibility and postponement. Clearly, with sufficiently long delivery leadtime, the firm can opt for the **first-best solution – full flexibility and full postponement** strategy. However, this strategy is costly because full flexibility requires all plants to be capable of producing all products (i.e., effectively pooling the plants' capacity together) while full postponement is possible only if the delivery leadtime is long - the firm can obtain complete demand information prior to any production activity. In this way, the firm can essentially use a central plant to produce all products (or a network of plants, all of which can produce all products). When production leadtime is moderately long, firms can opt instead to produce a portion of the demand by forecast first, before reverting to make-to-order mode to fully utilize the production capacity during the leadtime window. We call this **Option A – full flexibility and partial postponement**. On the other hand, firms like Li and Fung can contract a network of small manufacturers, each specializing in only a limited number of products. These small manufacturers are typically on standby and can respond to production requests very quickly after receiving firm orders. Unlike a centralized facility, these plants have only limited range of production flexibility. We can think of this as **Option B – partial flexibility and full postponement**. In practice, however, firms often adopt a hybrid of the above - a portion of the capacity from the contractors are used as reactive capacity, but due to short delivery leadtime, a chunk of the contractors' capacity are used in a make-to-stock fashion to produce the products in advance. We call this **Option C – partial flexibility and partial postponement**.

Jordan and Graves (1995) show that **Option B**, configured the right way using a "chaining" strategy, can already accrue most (almost 95%) of the benefits of the first-best solution at a small fraction of the cost. They model the problem as a two-stage stochastic program where the strategic

decision of process flexibility design is carried out in the first stage while the production allocation is chosen in the second stage after demand is realized[2]. This chaining concept has been extended in various other directions (Graves and Tomlin 2003, Gurumurthi and Benjaafar 2004, Hopp et al. 2004, Bish et al. 2005, Iravani et al. 2005, Muriel et al. 2006, Deng and Shen 2013). Likewise, efforts were also expended to strengthen its analytical aspect (Akşin and Karaesmen 2007, Chou et al. 2010a, 2010b, 2011, Bassamboo et al. 2010, 2012, Simchi-Levi and Wei 2012). For a review of process flexibility and discussion on how the concept has been deployed in several manufacturing and service systems, please refer to Chou et al. (2008). However, to our best knowledge, none of these papers consider the impact of partial production postponement.

The analytical papers in the literature focus mainly on the 2-chain, where each plant can produce exactly two products and each product can be produced by two plants. These papers find that the 2-chain performs extremely well. For example, Chou et al. (2010b) use a random walk approach to characterize the asymptotic performance of the 2-chain while Simchi-Levi and Wei (2012) use a supermodularity property to characterize the performance of the 2-chain in finite systems. As will be unveiled in this paper, there are situations where higher chains (e.g. 3-chain) are necessary to offset performance losses due to partial postponement. To the best of our knowledge, there are no existing results on the d-chain for $d > 2$. Moreover, the supermodulariy technique used in Simchi-Levi and Wei (2012) no longer works for d-chains when $d > 2$. We generalize the random walk argument used in Chou et al. (2010b) to higher chains and to arbitrary levels of production postponement. More importantly, this new approach allows us to examine the performance of systems such as Option A and Option C.

In Option A or Option C, we need to address the issue of first-stage production allocation - what is the best way to utilize production capacity in the first stage when the second-stage production capacity is limited by partial flexibility? While a number of papers discuss or study the postponement decision (Signorelli and Heskett 1984, Lee et al. 1993, Lee and Tang 1997, Swaminathan and Lee 2003), most of them consider postponement in terms of deferring certain steps in the manufacturing process to a point when demand information becomes available. Our interest, however, is in production postponement, which we have defined as the proportion of capacity that can be allocated after demand information is known. Van Mieghem and Dada (1999) examine the trade-off between production postponement and price postponement, but they do not consider the issue of process flexibility.

The work that most closely relates to ours is that of Fisher and Raman (1996), who demonstrate that **Option A**– where a single production facility acts equivalently as a fully flexible production network – can lead to significant savings. For Sport Obermeyer, a major fashion skiwear company,

they report that after observing only 20% of initial demand, the company can increase its profits by as much as 60%. While much of this increase is probably attributed to margin arithmetic (Cachon and Terwiesch 2009) because net profit is relatively low to begin with, the study nonetheless demonstrates the substantial impact of production postponement – a small portion of reactive capacity (fully flexible) can have tremendous value in matching production capacity with demand in the supply chain. However, their model assumes full flexibility in the second stage and hence is not able to handle systems with partial flexibility, that is the more general **Option C**. In this paper, we essentially combine the key insights in these areas to arrive at the following observation - that a small amount of flexibility (3-chain, instead of 2-chain) and a small amount of reserved capacity, can add tremendous value in matching supply and demand. Furthermore, we quantify the performance gap of a 3-chain vis-a-viz the first-best solution under different postponement levels.

The rest of the paper is organized as follows. In Section 2, we introduce the basic production allocation model and define the performance measures. Section 3 presents our analysis of the first-stage make-to-stock production decision, given that the second stage make-to-order production network has limited range of production flexibility. This resulted in a complex two stage stochastic programming model. We derived structural results for the optimal production plan in the first stage, when the production system is symmetric but not necessarily balanced. In Section 4, we analyze the overall performance of different productions systems with partial postponement strategy. We show analytically in Section 5 that the 3-chain can recover most of the flexibility loss caused by partial postponement. In Section 5.1, we present the random walk approach for asymptotic performance of long chains with degree greater than two and any arbitrary level of postponement. In Section 6, we examine the postponement and flexibility trade-off under asymmetric systems where plant capacities and product demand distributions are no longer identical. Finally, Section 7 concludes the paper.

2. The Model

In this section, we generalize the process flexibility model under full postponement to the case where the postponement level can range anywhere between the extremes of make-to-stock and make-to-order. To this end, we develop a model to capture partial levels of both process flexibility and production postponement. The setting is as follows. We consider a system with n plants and n products. As in the literature, we let $\mathcal{A}(n)$ and $\mathcal{B}(n)$ represent the set of product nodes and the set of plant nodes, respectively. The product demands are $D_1, D_2, \ldots, D_n$ which are independent and identically distributed continuous random variables with density function ϕ that is symmetrical about the mean μ and distribution function Φ. This family of distributions includes the uniform

and normal distributions. The plants, on the other hand, have fixed capacities of C units each. This setting is known as the symmetric but unbalanced case. In some instances, we shall consider the balanced and symmetric case where $C = \mu$ for ease of exposition.

Early on, the firm carries out two strategic decisions; namely, the level of flexibility and the level of postponement. For flexibility, the firm chooses a flexibility configuration $\mathcal{G}(n) \subset \mathcal{A}(n) \times \mathcal{B}(n)$. Due to their well-established efficiency, we focus on a class of symmetric flexibility structures known as d-chains. Doing so reduces the decision to a scalar d, denoting the common node degree. Although there exist many structures with all nodes having degree d, d-chains are the ones that form the longest possible chain.

DEFINITION 1. For $d = 1, 2, \ldots, n$, the *d-chain* is

$$\mathcal{C}_d(n) \triangleq \left\{ \bigcup_{i=1}^{n-d+1} \{(i,i), (i,i+1), \ldots, (i,i+d-1)\} \right\} \cup \left\{ \bigcup_{i=n-d+2}^{n} \{(i,i), (i,i+1), \ldots, (i,n), (i,1), (i,2), \ldots, (i,i-n+d-1)\} \right\}$$

The extremes of $d = 1$ and $d = n$ correspond to no flexibility (also known as the dedicated system) and full flexibility, respectively. All other values of d in between account for varying levels of partial flexibility, thus generalizing the 2-chain (or chaining) defined earlier. Whenever the context allows, we also return to the following previous notations in the literature.

$$\mathcal{D}(n) = \mathcal{C}_1(n), \qquad \mathcal{C}(n) = \mathcal{C}_2(n), \qquad \mathcal{F}(n) = \mathcal{C}_n(n)$$

For production postponement, we model a two-stage production process and define α as the proportion of capacity postponed to the second stage while $1 - \alpha$ is for first-stage consumption. When $\alpha = 0$, we have a make-to-stock setting and all production must be decided in the first stage. When $\alpha = 1$, our model reduces to the make-to-order, full-postponement setting in the literature. We allow the firm to choose its desired postponement level α over the range $[0, 1]$.

Once a combination of $\mathcal{G}(n)$ (equivalently, d) and α is chosen, we have to look beyond just minimizing lost sales because overage cost is no longer zero in this general case. The performance measure to use is expected mismatch cost which can be determined by solving the following two-stage problem. In the first stage, $(1 - \alpha)C$ units are made available at each plant to produce whatever allowed combination of products $1, 2, \ldots, n$ to stock, i.e. without information on actual final demand. In the second stage, the remaining αC units in each plant become available to meet whatever actual demand the firm cannot fill from first-stage stock. Our problem here is essentially a multi-item newsvendor model with second-stage supply and partial capacity sharing, whereby the expected mismatch cost is minimized. For ease of analysis, we let the unit overage cost and

unit underage cost for all products be identical, denoted by c_o and c_u, respectively. Further denote by $G^*_{\mathcal{G}(n)}(\alpha)$ the optimal expected mismatch cost. The production allocation decisions are x_{ij} and y_{ij}, which denote the amounts of product i produced by plant j in the first and second stages, respectively.

$$\begin{aligned}(P1): \quad G^*_{\mathcal{G}(n)}(\alpha, C) = \min_{\boldsymbol{x}} \; & G_{\mathcal{G}(n)}(\boldsymbol{x}, \alpha, C) \\ \text{s.t.} \quad & \sum_{i=1}^{n} x_{ij} \le (1-\alpha)C \quad \forall j = 1, 2, \ldots n \\ & x_{ij} \ge 0 \quad \forall i, j = 1, 2, \ldots n \\ & x_{ij} = 0 \quad \forall (i,j) \notin \mathcal{G}(n)\end{aligned}$$

where

$$G_{\mathcal{G}(n)}(\boldsymbol{x}, \alpha, C) = c_o g_1(\boldsymbol{x}) + c_u g_2(\boldsymbol{x}) - c_u \mathrm{E}[h_{\mathcal{G}(n)}(\boldsymbol{x}, \alpha, \boldsymbol{D}, C)]$$

$$g_1(\boldsymbol{x}) = \sum_{i=1}^{n} \mathrm{E}\left[\left(\sum_{j=1}^{n} x_{ij} - D_i\right)^+\right]$$

$$g_2(\boldsymbol{x}) = \sum_{i=1}^{n} \mathrm{E}\left[\left(D_i - \sum_{j=1}^{n} x_{ij}\right)^+\right]$$

and

$$\begin{aligned}h_{\mathcal{G}(n)}(\boldsymbol{x}, \alpha, \boldsymbol{D}, C) = \max_{\boldsymbol{y}} \; & \sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij} \\ \text{s.t.} \quad & \sum_{j=1}^{n} y_{ij} \le \left(D_i - \sum_{j=1}^{n} x_{ij}\right)^+ \quad \forall i = 1, 2, \ldots n \\ & \sum_{i=1}^{n} y_{ij} \le \alpha C \quad \forall j = 1, 2, \ldots n \\ & y_{ij} \ge 0 \quad \forall i, j = 1, 2, \ldots n \\ & y_{ij} = 0 \quad \forall (i,j) \notin \mathcal{G}(n)\end{aligned}$$

Before proceeding further, we summarize the sequence of events.

1. The firm decides flexibility structure $\mathcal{G}(n)$ (equivalently, the value of d for d-chaining) and the level of postponement α.
2. The first-stage production decisions x_{ij} are made.
3. Product demands D_i are observed.
4. The second-stage production decisions y_{ij} are made.
5. Mismatch costs are incurred.

Because our interest is to compare the performance of any flexibility-postponement combination vis-à-vis the first-best solution, we introduce the following quantities which will help us understand the effects of having only partial flexibility, partial postponement, or both.

DEFINITION 2. Given any combination of $\mathcal{G}(n)$ and α, and capacity C, the *optimality loss* relative to the first-best solution is the difference in optimal expected mismatch costs.

$$L_T(\mathcal{G}(n),\alpha,C) \triangleq G^*_{\mathcal{G}(n)}(\alpha,C) - G^*_{\mathcal{F}(n)}(1,C)$$

Furthermore, this quantity is made up of two components. The *postponement loss* is the loss due to partial postponement

$$L_P(\alpha,C) \triangleq G^*_{\mathcal{F}(n)}(\alpha,C) - G^*_{\mathcal{F}(n)}(1,C)$$

while the *flexibility loss* is the loss due to partial flexibility

$$L_F(\mathcal{G}(n),\alpha,C) \triangleq G^*_{\mathcal{G}(n)}(\alpha,C) - G^*_{\mathcal{F}(n)}(\alpha,C)$$

such that $L_T(\mathcal{G}(n),\alpha,C) = L_P(\alpha,C) + L_F(\mathcal{G}(n),\alpha,C)$.

For the class of flexibility structures in Definition 1, we can also gauge the percentage of flexibility loss as system size grows very large. To do so, we define the following performance measure.

DEFINITION 3. The *asymptotic chaining efficiency (ACE)* of the d-chain at postponement level α and capacity C is the expected improvement (over dedicated structure) ratio of the d-chain relative to full flexibility both at postponement level α as system size approaches infinity.

$$ACE(d,\alpha,C) \triangleq \lim_{n\to\infty} \frac{G^*_{\mathcal{D}(n)}(\alpha,C) - G^*_{\mathcal{C}_d(n)}(\alpha,C)}{G^*_{\mathcal{D}(n)}(\alpha,C) - G^*_{\mathcal{F}(n)}(\alpha,C)}$$

3. Make-to-Stock: The First-Stage Decision

To gain insights that can be useful for the general case where plant capacities and demand distributions are not identical, we first focus on the symmetric but unbalanced case. Interestingly, in this setting, we can characterize the first-stage decision analytically under certain conditions - we show that the first-stage production decision does not depend on the process flexibility structure. This allows us to simplify the entire optimization problem.

To this end, we define Problem 2 by relaxing first-stage production to be fully flexible while still holding second-stage production to $\mathcal{G}(n)$-flexibility. Notice that under full flexibility, there will be multiple optimal solutions in the first stage. Hence, the n^2-dimensional decision vector $\boldsymbol{x}$ in (P1) can be reduced to the n-dimensional decision vector $\boldsymbol{z}$ by letting $z_i = \sum_{j=1}^n x_{ij}, \forall i = 1,2,\ldots,n$.

$$\begin{aligned}(P2): \quad \overline{G}^*_{\mathcal{G}(n)}(\alpha,C) = \min_{\boldsymbol{x}} \; & G_{\mathcal{G}(n)}(\boldsymbol{x},\alpha,C) \\ \text{s.t.} \quad & \sum_{i=1}^n x_{ij} \le (1-\alpha)C \quad \forall j = 1,2,\ldots n\end{aligned}$$

$$
\begin{aligned}
& x_{ij} \geq 0 \quad \forall i,j = 1,2,\ldots n \\
= \min_{\boldsymbol{z}} \ & \overline{G}_{\mathcal{G}(n)}(\boldsymbol{z},\alpha,C) \\
\text{s.t.} \ & \sum_{i=1}^{n} z_i \leq (1-\alpha)nC \\
& z_i \geq 0 \quad \forall i = 1,2,\ldots n
\end{aligned}
$$

where

$$\overline{G}_{\mathcal{G}(n)}(\boldsymbol{z},\alpha,C) = c_o\overline{g}_1(\boldsymbol{z}) + c_u\overline{g}_2(\boldsymbol{z}) - c_u\mathrm{E}[\overline{h}_{\mathcal{G}(n)}(\boldsymbol{z},\alpha,\boldsymbol{D},C)]$$

$$\overline{g}_1(\boldsymbol{z}) = \sum_{i=1}^{n} \mathrm{E}[(z_i - D_i)^+]$$

$$\overline{g}_2(\boldsymbol{z}) = \sum_{i=1}^{n} \mathrm{E}[(D_i - z_i)^+]$$

and

$$
\begin{aligned}
\overline{h}_{\mathcal{G}(n)}(\boldsymbol{z},\alpha,\boldsymbol{D},C) = \max_{\boldsymbol{y}} \ & \sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij} \\
\text{s.t.} \ & \sum_{j=1}^{n} y_{ij} \leq (D_i - z_i)^+ \quad \forall i = 1,2,\ldots n \\
& \sum_{i=1}^{n} y_{ij} \leq \alpha C \quad \forall j = 1,2,\ldots n \\
& y_{ij} \geq 0 \quad \forall i,j = 1,2,\ldots n \\
& y_{ij} = 0 \quad \forall (i,j) \notin \mathcal{G}(n)
\end{aligned}
$$

To characterize the first-stage decision, we present the following results.

LEMMA 1. *Suppose $f : \mathcal{R}^n \to \mathcal{R}$ and dom $f = \mathcal{R}^n_+$. Define $g : \mathcal{R}^n \to \mathcal{R}$ by $g(\boldsymbol{x}) = f(\boldsymbol{x}^+)$, where $\boldsymbol{x}^+$ is the component-wise positive part of $\boldsymbol{x}$. If f is convex in $\boldsymbol{x}$ and nondecreasing in each argument x_i over $[0,\infty)$, then g is convex in $\boldsymbol{x}$.*

Proof. This lemma is a special case of the vector composition result in Section 3.2.4 of Boyd and Vandenberghe (2009, p86). □

This allows us to show that the function $\overline{h}_{\mathcal{G}(n)}(\boldsymbol{z},\alpha,\boldsymbol{D},C)$ is convex in $\boldsymbol{z}$, leading to our first result.

PROPOSITION 1. *$\overline{G}_{\mathcal{G}(n)}(\boldsymbol{z},\alpha,C)$ is convex in $\boldsymbol{z}$ for any structure $\mathcal{G}(n)$.*

Next, we present a short lemma to help us prove our first main result.

LEMMA 2. *Suppose $f : \mathcal{R} \to \mathcal{R}$ is an increasing convex function while $g, \hat{g} : \mathcal{R} \to \mathcal{R}$ are decreasing convex functions such that $\hat{g}'(x) \leq g'(x) \leq 0$. If x^* minimizes $f(x) + g(x)$ and $\hat{x}^*$ minimizes $f(x) + \hat{g}(x)$, then $x^* \leq \hat{x}^*$.*

Proof. It follows from optimality that $f'(x^*) = -g'(x^*)$ and $f'(\hat{x}^*) = -\hat{g}'(\hat{x}^*)$. Since f is convex while $-g, -\hat{g}$ are concave, f' is nondecreasing and $-g', -\hat{g}'$ are nonincreasing. Because $-\hat{g}'(x) \geq -g'(x)$, $x^* \leq \hat{x}^*$. $\square$

We are now ready to present our first main result. Let $\hat{\Phi}_n$ denote the n-fold convolution of Φ, and v^* the unique solution to the following equation:

$$\left[c_o\Phi(v) + c_u\hat{\Phi}_n(nv + \alpha nC)\right] = c_u \tag{1}$$

PROPOSITION 2. *$x^*_{ii} = x^*_{jj}, \forall i \neq j$ and $x^*_{ij} = 0, \forall i \neq j$ is a solution to both (P1) and (P2), when $\mathcal{G}(n)$ is symmetric. Furthermore, if $(1-\alpha)C \leq v^*$, then $x^*_{ii} = (1-\alpha)C$.*

Proof. From Proposition 1, the function $\overline{G}_{\mathcal{G}(n)}(\boldsymbol{z}, \alpha, C)$ is convex. Let $\boldsymbol{z}^{(k)}$ denote the vector obtained by shifting the vector $\boldsymbol{z}$ by k positions in the clockwise direction. Since the flexible structure is symmetric, and D's are i.i.d., $\overline{G}_{\mathcal{G}(n)}(\boldsymbol{z}, \alpha, C) = \overline{G}_{\mathcal{G}(n)}(\boldsymbol{z}^{(k)}, \alpha, C)$ for all k. Furthermore,

$$\overline{G}_{\mathcal{G}(n)}\left(\frac{1}{n}\sum_{k=0}^{n-1}\boldsymbol{z}^{(k)}, \alpha, C\right) \leq \frac{1}{n}\sum_{k=0}^{n-1}\overline{G}_{\mathcal{G}(n)}(\boldsymbol{z}^{(k)}, \alpha, C) = \overline{G}_{\mathcal{G}(n)}(\boldsymbol{z}, \alpha, C)$$

Observe that the components of vector $\frac{1}{n}\sum_{k=0}^{n-1}\boldsymbol{z}^{(k)}$ are all equal to $\frac{1}{n}\sum_{i=1}^{n} z_i$. Hence, the objective function is minimized at $z_i^* = z_0, \forall i = 1, 2, \ldots, n$. Note that only the dedicated arcs need to be utilized for this case. It follows that the solution obtained is also optimal for $(P1)$. Hence, the flexibility structure does not affect the first stage production, as long as the structure is symmetric.

We want to find z_0 that minimizes

$$\begin{aligned}
\overline{G}_{\mathcal{G}(n)}(z_0\mathbf{1}, \alpha, C) &= c_o\overline{g}_1(z_0\mathbf{1}) + c_u\overline{h}_1(z_0\mathbf{1}, \alpha, C) \\
&\geq c_o\overline{g}_1(z_0\mathbf{1}) + c_u\overline{h}_2(z_0\mathbf{1}, \alpha, C) \\
&\geq c_o\overline{g}_1(z_0\mathbf{1}) + c_u\overline{h}_3(z_0\mathbf{1}, \alpha, C)
\end{aligned}$$

where

$$\begin{aligned}
\overline{h}_1(z_0\mathbf{1}, \alpha, C) &= \mathrm{E}\left[\sum_{i=1}^{n}(D_i - z_0)^+ - \overline{h}_{\mathcal{G}(n)}(z_0\mathbf{1}, \alpha, \boldsymbol{D}, C)\right] \\
\overline{h}_2(z_0\mathbf{1}, \alpha, C) &= \mathrm{E}\left[\sum_{i=1}^{n}(D_i - z_0)^+ - \overline{h}_{\mathcal{F}(n)}(z_0\mathbf{1}, \alpha, \boldsymbol{D}, C)\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{n}(D_i - z_0)^+ - \min\left(\sum_{i=1}^{n}(D_i - z_0)^+, \alpha nC\right)\right] \\
&= \mathrm{E}\left[\max\left(0, \sum_{i=1}^{n}(D_i - z_0)^+ - \alpha nC\right)\right] \\
\overline{h}_3(z_0\mathbf{1}, \alpha, C) &= \mathrm{E}\left[\max\left(0, \sum_{i=1}^{n}(D_i - z_0) - \alpha nC\right)\right]
\end{aligned}$$

Let

$$\begin{aligned}\hat{G}(z_0,\alpha,C) &= c_o\overline{g}_1(z_0\mathbf{1}) + c_u\overline{h}_3(z_0\mathbf{1},\alpha,C)\\ &= c_o\sum_{i=1}^{n}\int_0^{z_0}(z_0-\xi_i)d\Phi(\xi_i) + c_u\int_{nz_0+\alpha nC}^{\infty}(\xi-nz_0-\alpha nC)d\hat{\Phi}_n(\xi)\end{aligned}$$

where $\xi=\sum_{i=1}^{n}\xi_i$ and $\hat{\Phi}_n$ is the n-fold convolution of Φ.

Next, we define the following unconstrained minimizers.

$$\overline{z}_0^* \triangleq \arg\min_{z_0}\{\overline{G}_{\mathcal{G}(n)}(z_0\mathbf{1},\alpha,C)\}$$

$$\hat{z}_0^* \triangleq \arg\min_{z_0}\{\hat{G}_{\mathcal{G}(n)}(z_0\mathbf{1},\alpha,C)\}$$

To find $\hat{z}_0^*$, we take the following first-order condition.

$$\begin{aligned}\frac{\partial\hat{G}(z_0,\alpha,C)}{\partial z_0} &= c_o n\Phi(z_0) - c_u n[1-\hat{\Phi}_n(nz_0+\alpha nC)]\\ 0 &= n\left[c_o\Phi(\hat{z}_o) - c_u + c_u\hat{\Phi}_n(n\hat{z}_o+\alpha nC)\right]\end{aligned}$$

It follows that $\hat{z}_0^*=v^*$. Note further that $\frac{\partial\overline{h}_1}{\partial z_0}\le\frac{\partial\overline{h}_2}{\partial z_0}\le\frac{\partial\overline{h}_3}{\partial z_0}\le 0$. By Lemma 2, $\overline{z}_0^*\ge\hat{z}_0^*$. Because $\overline{G}_{\mathcal{G}(n)}(z_0\mathbf{1},\alpha,C)$ is convex and $(1-\alpha)C\le v^*\le\overline{z}_0^*$, we have

$$z_0^* = \arg\min_{z_0\le(1-\alpha)C}\{\overline{G}_{\mathcal{G}(n)}(z_0\mathbf{1},\alpha,C)\} = (1-\alpha)C$$

Hence, the optimal solution to (P2) is $z_i^*=z_0^*=(1-\alpha)C,\ \forall i=1,2,\ldots,n$. This is equivalent to $x_{ii}^*=(1-\alpha)C,\ \forall i=1,2,\ldots,n$ and $x_{ij}^*=0,\ \forall i\ne j$. Since this solution is feasible for (P1), it is also optimal for (P1). □

Remark 1. If $(1-\alpha)C>v^*$, then the following results remain true: $z_i^*=z_0,\forall i=1,2,\ldots,n$; $v^*\le\overline{z}_0^*$. However, we may encounter one of two cases: $(1-\alpha)C\le\overline{z}_0^*$ and $(1-\alpha)C>\overline{z}_0^*$. In the first case, the optimal primary production remains at $z_0^*=(1-\alpha)C$. In the second case, the optimal primary production is less definitive at $z_0^*\in[v^*,(1-\alpha)C)$. Here, we again have $x_{ij}=0,\forall i\ne j$, and the optimal solution also optimal for (P1).

Remark 2. It is also interesting to note that our problem extends the single-item newsvendor in two directions. In (1), setting $\alpha=0$ and $n=1$ will result in the classical single-item newsvendor critical fractile formula. If $\alpha>0$, then we have some second-stage backup supply (albeit limited). If we also set $n>1$, then we have multiple items with i.i.d. demands and fully flexible second-stage supply. Our problem of interest is more complicated because we only have partially flexible second-stage supply.

Proposition 2 tells us that for the symmetric but unbalanced case, the optimal first-stage production is to exhaust all first-stage capacity for primary production regardless of the flexibility structure. In other words, whatever the flexibility structure, it must act like a dedicated structure in the first stage. This result is important in two ways. First, it confirms our intuition that flexible capacity is useless if there is no postponement. Second, it allows us to solve the first-stage problem, which in turn simplifies our succeeding analysis of the effect of partial postponement.

For the general asymmetric case where plant capacities and product demands are not necessarily identical, the first-stage production decision becomes unwieldy. How to allocate the first-stage capacity among the different products becomes a very difficult and complicated problem. As a result, we may have to settle for heuristic approaches such as the "Mean Rule" and the "Variance Rule". Essentially, the Mean Rule suggests that total first-stage capacity be allocated among the products proportionally according to the mean values of their demands. On the other hand, the Variance Rule follows the principle that products with higher coefficient of variation should utilize less of the (speculative) first-stage capacity. These are more sophisticated policies and we shall relegate their discussion to a latter part in Section 6. We notice that these rules are often employed in production systems that have second-stage full flexibility structure. We address the performance of these production rules when each plant has only limited range of production flexibility in the rest of this paper.

4. Effect of Partial Postponement

In this section, we use results from the previous section to examine systems with partial postponement. We first study Option A (with full flexibility), followed by Option C (with partial flexibility).

4.1. Full Flexibility

We examine how full flexibility with partial postponement performs relative to the first-best solution. Since we have full flexibility in both systems, it boils down to solving $G^*_{\mathcal{F}(n)}(\alpha, C)$ for different postponement levels $\alpha \in [0,1]$.

Using Proposition 2, we can obtain a closed-form expression for the optimal expected mismatch cost. For C not too large, i.e., $C \leq \frac{v^*}{1-\alpha}, \forall \alpha \in [0,1]$ where v^* is obtained from equation (1),

$$G^*_{\mathcal{F}(n)}(\alpha, C) = c_o \sum_{i=1}^{n} \mathrm{E}\left[\left((1-\alpha)C - D_i\right)^+\right] + c_u \mathrm{E}\left[\max\left(0, \sum_{i=1}^{n}\left(D_i - (1-\alpha)C\right)^+ - \alpha n C\right)\right]$$

There is no flexibility loss (i.e., $L_F(\mathcal{F}(n), \alpha, C) = 0$) because we have a fully flexible system. Hence, we focus on postponement loss $L_P(\alpha, C) \triangleq G^*_{\mathcal{F}(n)}(\alpha, C) - G^*_{\mathcal{F}(n)}(1, C)$. Since $G^*_{\mathcal{F}(n)}(\alpha, C) \sim O(n)$ for $\alpha << 1$, whereas $G^*_{\mathcal{F}(n)}(1, C) \sim O(\sqrt{n}) \approx G^*_{\mathcal{C}(n)}(1, C)$, we conclude that full flexibility with

low postponement (i.e., partial postponement with small α) could not attain the same order of performance as full postponement with partial flexibility. This leads us to the following insight: **If we are to gainfully employ full flexibility with partial postponement, we must be able to postpone a substantial amount of the capacity (say $\alpha \geq 0.5$).**

4.2. Partial Flexibility

Next, what happens when we have partial levels of both flexibility and postponement? In particular, how much is the flexibility loss under partial postponement? Under full postponement, it has already been established that flexibility loss (equivalently, optimality loss) of the 2-chain is negligible. Likewise, Proposition 2 tells us that under no postponement, any form of flexibility brings no additional benefits. Hence, full flexibility is not any better than the 2-chain in the same way that the 2-chain is not any better than the dedicated structure. This implies that the flexibility loss of the 2-chain under no postponement is zero. The next question becomes whether we can also say that the flexibility loss of the 2-chain is negligible under partial postponement.

To answer this question, we characterize the flexibility loss $L_F(\mathcal{C}_2(n), \alpha, C)$ of the 2-chain as postponement level α changes from 0 to 1. Specifically, for $C \leq \frac{v^*}{1-\alpha}, \forall \alpha \in [0,1]$,

$$\begin{aligned} L_F(\mathcal{C}_2(n), \alpha, C) &\triangleq G^*_{\mathcal{C}_2(n)}(\alpha, C) - G^*_{\mathcal{F}(n)}(\alpha, C) \\ &= c_u \mathrm{E}[\overline{h}_{\mathcal{F}(n)}((1-\alpha)C\mathbf{1}, \alpha, \boldsymbol{D}, C)] - c_u \mathrm{E}[\overline{h}_{\mathcal{C}_2(n)}((1-\alpha)C\mathbf{1}, \alpha, \boldsymbol{D}, C)] \end{aligned}$$

where $\overline{h}_{\mathcal{G}(n)}(\cdot)$ is as defined in (P2). Moreover, the second equation is due to Proposition 2, which allows $c_o g_1(\boldsymbol{x}^*) + c_u g_2(\boldsymbol{x}^*)$ to cancel out. The next result shows that the flexibility loss of the 2-chain is largest at some postponement level strictly between zero postponement and full postponement.

PROPOSITION 3. *If $(1-\alpha)C \leq v^*, \forall \alpha \in [0,1]$, then $\exists \alpha \in (0,1)$ such that $L_F(\mathcal{C}_2(n), \alpha, C)$ is largest.*

Proof. Please refer to Appendix A.

Proposition 3 suggests that for capacity not too large and for certain levels of partial postponement, the performance gap between full flexibility and the 2-chain may be much more sizable than it is under the full postponement case or the no postponement case. In order to see how large this gap can grow, we use the Sample Average Approximation (SAA) method for stochastic programming. We sample a large number K of demand scenarios, with which we reformulate (P1) into the following large linear program. For purpose of illustration, we consider the symmetric case where demand and capacity are balanced (i.e., $C = \mu$).

$$(P3): \quad G^*_{\mathcal{G}(n)}(\alpha) = \tfrac{1}{K} \min_{\boldsymbol{x}} \sum_{i=1}^{n} \sum_{k=1}^{K} (c_0 v_i^k + c_u w_i^k) - c_u \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{K} y_{ij}^k$$

$$
\begin{aligned}
\text{s.t.} \quad & v_i^k \geq \sum_{j=1}^{n} x_{ij} - D_i^k \quad \forall i = 1,2,\ldots n, \forall k = 1,2,\ldots K \\
& w_i^k \geq D_i^k - \sum_{j=1}^{n} x_{ij} \quad \forall i = 1,2,\ldots n, \forall k = 1,2,\ldots K \\
& \sum_{i=1}^{n} x_{ij} \leq (1-\alpha)\mu \quad \forall j = 1,2,\ldots n \\
& \sum_{j=1}^{n} y_{ij}^k \leq w_i^k \quad \forall i = 1,2,\ldots n, \forall k = 1,2,\ldots K \\
& \sum_{i=1}^{n} y_{ij}^k \leq \alpha\mu \quad \forall j = 1,2,\ldots n, \forall k = 1,2,\ldots K \\
& x_{ij} \geq 0 \quad \forall i,j = 1,2,\ldots n \\
& x_{ij} = 0 \quad \forall (i,j) \notin \mathcal{G}(n) \\
& y_{ij}^k \geq 0 \quad \forall i,j = 1,2,\ldots n, \forall k = 1,2,\ldots K \\
& y_{ij}^k = 0 \quad \forall (i,j) \notin \mathcal{G}(n), \forall k = 1,2,\ldots K \\
& v_i^k, w_i^k \geq 0 \quad \forall i = 1,2,\ldots n, \forall k = 1,2,\ldots K
\end{aligned}
$$

$\boldsymbol{D}^k$ is the k^{th} demand scenario, while v_i^k and w_i^k are auxiliary variables introduced to linearize the formulation. We can also interpret v_i^k and w_i^k as the overage and underage quantities, respectively.

As in Jordan and Graves (1995), we simulate a 10-plant, 10-product system whereby each product has demand that follows a normal distribution with mean 100 units and standard deviation 30 units. We assume each plant has a capacity of 100 units, and for this illustration, $c_o = c_u = 1$. For each postponement level $\alpha \in \{0.00, 0.05, \ldots, 0.95, 1.00\}$ and each degree of flexibility $d \in \{1, 2, \ldots, 9, 10\}$, we solve (P3) over a fixed set of $K = 1000$ demand scenarios. Figure 1 plots the expected mismatch cost against the postponement level for different levels of flexibility. As expected, the gaps between the 2-chain line and the full flexibility line are negligible and zero at $\alpha = 1$ and $\alpha = 0$, respectively. However, for $\alpha \in [0.1, 0.7]$, the gap becomes quite sizable, especially between 0.2 and 0.5 where the gap ranges from 23% to 33%.

These findings, together with Proposition 3, lead us to the following insight: **the 2-chain which is known to be extremely effective is no longer as effective under partial postponement**. To gain further insights, we take the following example. Consider a 4 × 4 system with identical product demands following a two-point distribution with values 1 and 3, with equal probabilities of 0.5. The plant capacities are all 2 units each. We further suppose that $c_o = c_u = 1$ and $\alpha = 0.5$. Numerical tests show that for all 16 possible demand scenarios, the 3-chain performs as well as full flexibility. The 2-chain is also as good as full flexibility when total demand is strictly lower or strictly greater than total capacity. However, when total demand equals total capacity and

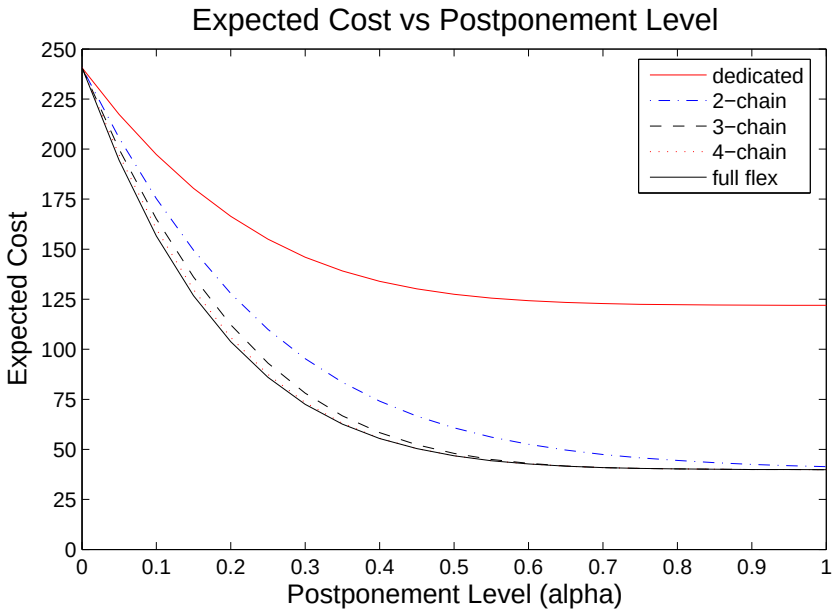


Figure 1 Expected Mismatch Cost vs. Postponement Level

two consecutive products have high demand while the other two have low demand, the 3-chain outperforms the 2-chain. An example is when demands for products 1 and 2 are 3 units each, while demands for products 3 and 4 are 1 unit each. More interestingly, this observation does not hold for $\alpha = 1$, wherein 2-chain is as good as full flexibility.

Looking at Figure 1, one may wonder why the gap between 2-chain mismatch cost and full flexibility mismatch cost is largest at intermediate levels of postponement. Intuitively, we can think of two forces that affect the size of this gap. The first force is caused by the unbalanced nature of the second-stage allocation problem. That is, the more unequal the expected remaining demand is to the remaining capacity, the larger the mismatch cost gap. It is easy to see that the second-stage problem becomes more unbalanced as α decreases because of demand truncation at 0. The second force is caused by the relative magnitude of the second-stage cost viz-a-viz the first-stage cost. Clearly, the smaller the α, the smaller the effect of the second-stage decision on the total mismatch cost. The trade-off between these two forces explains why the gap is largest at intermediate postponement levels.

That said, if one wants to approximate the benefits of full flexibility and full postponement using only partial levels of both these dimensions, care has to be taken in choosing the proper levels of flexibility and postponement that can give the desired result. In the event of partial postponement, more flexibility (a third or fourth layer of flexible links) is necessary to make up for not only the postponement loss, but more importantly, the increased flexibility loss. We explore in the next section just how much additional flexibility is necessary.

5. Value of the 3-Chain

In this section, we demonstrate that the 3-chain can recover most of the flexibility loss in the 2-chain created by partial postponement. Table 1 shows a partial listing of the cost results generated in Figure 1. As expected, the flexibility loss is smallest at the two extremes of $\alpha = 0$ and $\alpha = 1$. For $\alpha \in [0.1, 0.5]$, the gap between 2-chain and full flexibility increases to over 30% and beyond. However, by employing the 3-chain, the flexibility loss can be reduced to below 8.2% for all values of α.

Table 1 Expected Cost and Flexibility Loss for 2-Chain and 3-Chain with Partial Postponement

Postponement	Expected Cost			Flexibility Loss (%age Full Flex)	
Level α	2-Chain	3-Chain	Full Flex	2-Chain	3-Chain
0.00	4812.90	4812.90	4812.90	0.00 (0.0%)	0.00 (0.0%)
0.10	3505.60	3303.30	3131.90	373.70 (11.9%)	171.40 (5.5%)
0.20	2557.40	2245.40	2075.20	482.20 (23.2%)	170.20 (8.2%)
0.30	1906.40	1562.20	1450.10	456.30 (31.5%)	112.10 (7.7%)
0.40	1481.20	1167.80	1108.80	372.40 (33.6%)	59.00 (5.3%)
0.50	1214.60	959.63	935.77	278.83 (29.8%)	23.86 (2.5%)
0.60	1050.90	860.31	854.37	196.53 (23.0%)	5.94 (0.7%)
0.70	949.47	818.51	817.51	131.96 (16.1%)	1.00 (0.1%)
0.80	889.27	804.55	804.55	84.72 (10.5%)	0.00 (0.0%)
0.90	850.25	799.14	799.14	51.11 (6.4%)	0.00 (0.0%)
1.00	827.05	797.84	797.84	29.21 (3.7%)	0.00 (0.0%)

Remark 3. Even though the focus of this paper is on flexibility loss and how it is affected by changes in postponement level, our study also provides interesting insights that can guide companies in building their postponement capabilities. To be more specific, Table 1 can also be used to obtain the postponement loss percentages at various postponement levels. For example, when $\alpha = 0.4$, the postponement loss percentage is $(1,108.80 - 797.84)/797.84 = 39.0\%$. By observing the postponement loss percentages at various postponement levels, we can see that the lower the postponement level, the higher both the postponement loss and the marginal postponement loss (i.e., the additional postponement loss the system would incur with every unit of postponement level reduced). We also observe that this marginal postponement loss can be quite substantial when the postponement level is low. In other words, when the postponement level is low, adding a little postponement capability into the system can greatly improve the system efficiency, and the

benefit of further increasing the postponement level can be quite marginal once the postponement level reaches a certain degree. While it is often hard to achieve full postponement in practice, this observation provides useful insights for companies by pointing out the importance and sufficiency of partial postponement. Interestingly, we also find that unlike postponement loss percentage which decreases with postponement level, flexibility loss percentage is largest at intermediate postponement levels.

While system size $n = 10$ is used in Jordan and Graves' (1995) initial example on the effectiveness of 2-chain, Chou et al.'s (2010b) asymptotic analysis shows that as n increases to say 100, the 2-chain still performs very well. We examine next if the 3-chain can still recover the flexibility loss of the 2-chain when n grows very large. To this end, we perform asymptotic analysis similar to Chou et al.'s (2010b) analysis of the 2-chain under full postponement. Unlike their paper, our method works for any d-chain, $d \geq 2$, under any postponement level α.

5.1. Asymptotic Performance of the 3-Chain

To study the phenomenon of increasing system size, we extend the method of asymptotic analysis introduced in Chou et al. (2010b). In Definition 3, we propose that the relative flexibility loss of every d-chain with postponement level α and capacity C can be measured by its asymptotic chaining efficiency (ACE). While the next result also holds for the unbalanced case, we shall consider the balanced case where $C = \mu$ for ease of exposition. This allows us to remove C from the function arguments and notations, and simplify ACE as follows.

$$ACE(d,\alpha) \triangleq \lim_{n\to\infty} \frac{G^*_{\mathcal{D}(n)}(\alpha) - G^*_{\mathcal{C}_d(n)}(\alpha)}{G^*_{\mathcal{D}(n)}(\alpha) - G^*_{\mathcal{F}(n)}(\alpha)} = \lim_{n\to\infty} \frac{\hat{h}_{\mathcal{C}_d(n)}(\alpha) - \hat{h}_{\mathcal{D}(n)}(\alpha)}{\hat{h}_{\mathcal{F}(n)}(\alpha) - \hat{h}_{\mathcal{D}(n)}(\alpha)}$$

where $\hat{h}_{\mathcal{G}(n)}(\alpha) = \mathrm{E}[\overline{h}_{\mathcal{G}(n)}((1-\alpha)\mu\mathbf{1}, \alpha, \boldsymbol{D}, \mu)]$ and $\overline{h}_{\mathcal{G}(n)}(\cdot)$ is the maximum flow problem defined in (P2). When $d = 2$ and $\alpha = 1$, the system boils down to 2-chain with full postponement, precisely the system studied by Chou et al. (2010b). In what follows, we develop a generalized method to analyze $ACE(d,\alpha)$ for $2 \leq d \leq n$ and $\alpha \in [0,1]$ with the following new features.

- The entire analysis involves a different structure called a "d-path".
- The resulting random walk has a different random step size $X_i \triangleq \tilde{D}_i - \tilde{C}$.
- The resulting random walk has a different upper absorbing boundary at $(d-1)\tilde{C}$.
- The transformed random walk is an alternating regenerative process whose odd cycles are not identical to its even cycles.

We are now ready to present our method. First, we note that our problem of interest is an expected maximum flow problem whose demands and capacities we denote by $\tilde{D}_i = (D_i - (1-\alpha)\mu)^+$ and $\tilde{C} = \alpha\mu$. This is an unbalanced problem because $\mathrm{E}[\tilde{D}_i] \geq \tilde{C}$ for $\alpha \in [0,1]$. For the dedicated

and the fully flexible systems, it is easy to see that $\hat{h}_{\mathcal{D}(n)}(\alpha) = \mathrm{E}[\sum_{i=1}^{n} \min(\tilde{C}, \tilde{D}_i)]$ and $\hat{h}_{\mathcal{F}(n)}(\alpha) = \mathrm{E}[\min(n\tilde{C}, \sum_{i=1}^{n} \tilde{D}_i)]$. It follows that

$$ACE(d,\alpha) = \lim_{n\to\infty} \frac{\hat{h}_{\mathcal{C}_d(n)}(\alpha) - n\tilde{C} + n\mathrm{E}[(\tilde{C}-\tilde{D}_i)^+]}{n\mathrm{E}[(\tilde{C}-\tilde{D}_i)^+] - O(\sqrt{n})} = \frac{\mathrm{E}[(\tilde{C}-\tilde{D}_i)^+] - \tilde{C} + \lim_{n\to\infty} \frac{1}{n}\hat{h}_{\mathcal{C}_d(n)}(\alpha)}{\mathrm{E}[(\tilde{C}-\tilde{D}_i)^+]} \quad (2)$$

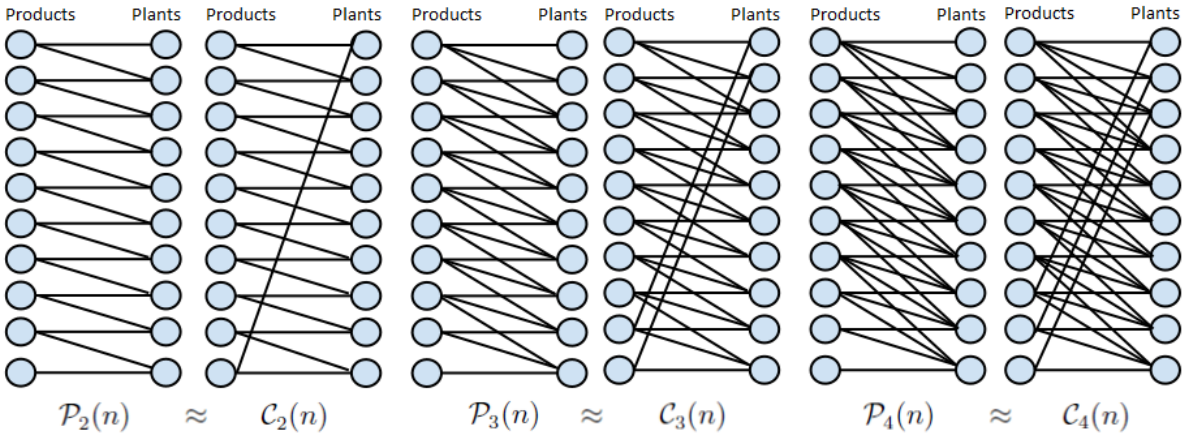
Hence, our problem reduces to finding $\lim_{n\to\infty} \frac{1}{n}\hat{h}_{\mathcal{C}_d(n)}(\alpha)$. Unlike Chou et al. (2010b), we delete from the d-chain the links connecting the first $d-1$ facility nodes with product nodes numbered higher than the facility node. Note that facilities d, $d+1$, and so on can still produce the same d products as before in the d-chain. The result is the following "d-path".

$$\mathcal{P}_d(n) = \mathcal{C}_d(n) \backslash \{(i,j) : j = 1,\ldots,d-1; i = n-d+1+j,\ldots,n\}$$

Figure 2 shows some d-paths and their corresponding d-chains. Moreover, it is easy to see that

$$0 \le \hat{h}_{\mathcal{C}_d(n)}(\alpha) - \hat{h}_{\mathcal{P}_d(n)}(\alpha) \le \frac{d(d-1)}{2} \cdot \tilde{C}$$

Figure 2 Examples of d-paths and d-chains



Hence, we have the following lemma, which allows us to focus on $\mathcal{P}_d(n)$.

LEMMA 3. *For finite d,*

$$\lim_{n\to\infty} \frac{\hat{h}_{\mathcal{P}_d(n)}(\alpha)}{n} = \lim_{n\to\infty} \frac{\hat{h}_{\mathcal{C}_d(n)}(\alpha)}{n}$$

We let the arc linking demand node i to supply node i denote the "primary" arc, and the arcs linking demand node i to supply node $j \neq i$ the "secondary" arcs. As can be seen in Figure 2, every plant i in $\mathcal{P}_d(n)$ can only serve products $i, i-1, \ldots, \max(1, i-d+1)$. This implies that the maximum flow on $\mathcal{P}_d(n)$ can be determined in a greedy fashion. First, satisfy demand $\tilde{D}_1$ of product

1 using the primary capacity in supply node 1, then if necessary, use supply node 2, and so on, in that order up to a maximum total capacity of $d\tilde{C}$ units. Next, move on to the next product, and based on the capacity left over from the previous product, add $\tilde{C}$ units more from a new supply node, and consume again according to lowest supply node number. The amount of maximum flow obtained in this greedy fashion is a random variable, depending on the values of $\tilde{D}_i$.

To present this greedy approach formally and to facilitate our analysis, we need to keep track of T_i, which denotes the amount of leftover capacity for product $i+1$ prior to adding $\tilde{C}$ units from the new supply node. At the beginning, $T_0 = (d-1)\tilde{C}$. As we move to the next product, T_i is updated as follows: $T_i := \min[(d-1)\tilde{C}, (T_{i-1} + \tilde{C} - \tilde{D}_i)^+]$. Alternatively, we can keep track of $S_i = (d-1)\tilde{C} - T_i$ which begins at $S_0 = 0$ and updates accordingly: $S_i := \min[(S_{i-1} + \tilde{D}_i - \tilde{C})^+, (d-1)\tilde{C}]$. Next, we let TF denote total maximum flow. Similarly, let $\text{TE} = \sum_{i=1}^{n} \tilde{D}_i - TF$ denote the difference between total demand and total flow, i.e., total excess or unmet demand. This implies that

$$\overline{h}_{\mathcal{G}(n)}((1-\alpha)\mu\mathbf{1}, \alpha, \boldsymbol{D}, \mu) = TF = \sum_{i=1}^{n} \tilde{D}_i - TE. \tag{3}$$

We account for TF by keeping track of TE as we assign capacity to demand. Consider step i of the greedy approach, wherein S_{i-1} is known before $\tilde{D}_i$ is observed. The greedy allocation implies $\text{TE} := \text{TE} + [(S_{i-1} + \tilde{D}_i - \tilde{C})^+ - (d-1)\tilde{C}]^+$. We summarize the greedy approach as follows.

Algorithm 1. (Greedy Approach)

1. Set $i := 1, S_0 := 0, T_0 := (d-1)\tilde{C}$, and $\text{TE} := 0$.
2. Observe $\tilde{D}_i$.

If $\tilde{D}_i > \tilde{C}$, then $S_i := \min[S_{i-1} + \tilde{D}_i - \tilde{C}, (d-1)\tilde{C}]$, and $\text{TE} := \text{TE} + \max[S_{i-1} + \tilde{D}_i - \tilde{C} - (d-1)\tilde{C}, 0]$.
If $\tilde{D}_i < \tilde{C}$, then $S_i := \max[S_{i-1} + \tilde{D}_i - \tilde{C}, 0]$, and $\text{TE} := \text{TE}$.

3. If $i = n-1$, then STOP. $\text{TE} := \text{TE} + \max(S_{n-1} + \tilde{D}_n - (d-1)\tilde{C}, 0)$. Return TE as the minimum excess. Otherwise, $i := i+1$ and go to Step 2.

Observe that $\{S_i : i = 0, 1, 2, \ldots\}$ behaves like a generalized random walk, with random step size $X_i \triangleq \tilde{D}_i - \tilde{C}$ and boundaries 0 and $(d-1)\tilde{C}$. The value TE grows in Step 2 only when $\tilde{D}_i - T_{i-1} > \tilde{C}$; that is, when $S_i = \min[S_{i-1} + \tilde{D}_i - \tilde{C}, (d-1)\tilde{C}] = (d-1)\tilde{C}$. We call this quantity $(X_i - T_{i-1})$ the level of overshoot at the upper boundary. Conversely, when $\tilde{D}_i < \tilde{C}$, it is possible that $S_{i-1} + \tilde{D}_i - \tilde{C} < 0$. We call this amount $(-S_{i-1} - X_i)$ the level of overshoot at the lower boundary. Note that this overshoot at the lower boundary does not add to TE in the greedy algorithm.

The random walk starts at $S_0 = 0$, the lower boundary. It gets trapped at the lower boundary whenever $X_i \leq 0$, and escapes only when $X_i > 0$. An interesting phenomenon happens when the random walk hits the upper boundary - it gets trapped at the upper boundary whenever $X_i \geq 0$, and escapes only when $X_i < 0$, the exact opposite.

Now, observe that $\{T_i : i = 0, 1, 2, \ldots\}$ also behaves like a similar random walk. In fact, it is the reflection of $\{S_i : i = 0, 1, 2, \ldots\}$ across the horizontal axis at $\frac{d-1}{2}\tilde{C}$. That is, its random step size is $X_i' \triangleq -X_i = \tilde{C} - \tilde{D}_i$, and TE grows whenever there is an overshoot in the lower boundary and not when the overshoot is at the upper boundary. Unlike $\{S_i : i = 0, 1, 2, \ldots\}$, $\{T_i : i = 0, 1, 2, \ldots\}$ begins at its upper boundary $T_0 = (d-1)\tilde{C}$. It gets trapped in the upper boundary whenever $X_i' \geq 0$, and escapes only when $X_i' < 0$. When it hits the lower boundary, it gets trapper there whenever $X_i' \leq 0$, and escapes only when $X_i' > 0$.

To stay consistent with the literature (Chou et al. 2010b), we define a new random walk $\{W_i, i = 0, 1, 2, \ldots\}$ that alternates between $\{S_i, i = 0, 1, 2, \ldots\}$ and $\{T_i, i = 0, 1, 2, \ldots\}$. This new random walk begins at $W_0 = S_0 = 0$ and upon hitting its upper boundary, switches to $\{T_i, i = 0, 1, 2, \ldots\}$. At this point, $W_i = T_i = 0$ and upon hitting its upper boundary, switches back to $\{S_i, i = 0, 1, 2, \ldots\}$. To model the switching times, we define the following stopping times.

$$\tau(j) \triangleq \begin{cases} \inf\{n : S_{n+\sum_{k=0}^{j-1}\tau(k)} = (d-1)\tilde{C}\}, & \text{if j is odd} \\ \inf\{n : T_{n+\sum_{k=0}^{j-1}\tau(k)} = (d-1)\tilde{C}\}, & \text{if j is even} \end{cases}$$

where $\tau(0) = 0$. That is,

$$W_i = \begin{cases} S_i, & \text{if } \tau(j-1) < i \leq \tau(j) \text{ and } j \text{ is odd} \\ T_i, & \text{if } \tau(j-1) < i \leq \tau(j) \text{ and } j \text{ is even} \end{cases} \qquad \forall i = 0, 1, 2, \ldots$$

Interestingly, $\{W_i, i = 0, 1, 2, \ldots\}$ turns out to be an alternating regenerative process. Because all alternating cycles are probabilistically identical, it suffices to examine just one pair of odd and even cycles with the following characteristics.

- Cycle Duration τ (resp, $\hat{\tau}$): the length of any odd (resp, even) regenerative cycle.

$$\tau \triangleq \inf\left\{n : S_n = (d-1)\tilde{C},\ n \geq 1, S_0 = 0\right\}, \quad \hat{\tau} \triangleq \inf\left\{n : T_{\tau(1)+n} = (d-1)\tilde{C},\ n \geq 1, T_{\tau(1)} = 0\right\}.$$

- Cycle Overshot ψ (resp, $\hat{\psi}$): the amount of overshoots at both the lower and upper boundaries in any odd (resp, even) cycle.

$$\psi \triangleq \sum_{i=1}^{\tau}\Big((S_i - S_{i-1} - X_i)\chi(X_i < 0) + (S_{i-1} + X_i - S_i)\chi(X_i > 0)\Big),$$

$$\hat{\psi} \triangleq \sum_{i=\tau(1)+1}^{\tau(1)+\hat{\tau}}\Big((T_i - T_{i-1} - X_i')\chi(X_i' < 0) + (T_{i-1} + X_i' - T_i)\chi(X_i' > 0)\Big).$$

where $\chi(\cdot)$ denote the indicator function.

Note that ψ can be decomposed into two components; namely, upper and lower overshoots, such that $\psi = \psi_L + \psi_U$ (resp, $\hat{\psi} = \hat{\psi}_L + \hat{\psi}_U$), where

$$\psi_L \triangleq \sum_{i=1}^{\tau} \Big((S_i - S_{i-1} - X_i)\chi(X_i < 0) \Big), \quad \hat{\psi}_L \triangleq \sum_{i=\tau(1)+1}^{\tau(1)+\hat{\tau}} \Big((T_i - T_{i-1} - X_i')\chi(X_i' < 0) \Big)$$

and

$$\psi_U \triangleq \sum_{i=1}^{\tau} \Big((S_{i-1} + X_i - S_i)\chi(X_i > 0) \Big), \quad \hat{\psi}_U \triangleq \sum_{i=\tau(1)+1}^{\tau(1)+\hat{\tau}} \Big((T_{i-1} + X_i' - T_i)\chi(X_i' > 0) \Big).$$

Consider an alternating renewal process $\{N(t) : t \geq 0\}$, having inter-arrival times $\tau(j)$ such that $\tau(j) \sim \tau$ if j is odd and $\tau(j) \sim \hat{\tau}$ if j is even. The reward R_j obtained at the jth renewal is ψ_U if j is odd, and is $\hat{\psi}_L$ if j is even. Note that from (3),

$$\sum_{i=1}^{n} \tilde{D}_i - \sum_{j=1}^{N(n)+1} R_j \leq \overline{h}_{\mathcal{G}(n)}((1-\alpha)\mu\mathbf{1}, \alpha, \boldsymbol{D}, \mu) \leq \sum_{i=1}^{n} \tilde{D}_i - \sum_{j=1}^{N(n)} R_j. \tag{4}$$

Because W_i toggles alternately between S_i and T_i and by the renewal reward theorem,

$$\lim_{n\to\infty} \frac{\mathrm{E}[\sum_{j=1}^{N(n)} R_j]}{n} = \frac{\mathrm{E}[\psi_U] + \mathrm{E}[\hat{\psi}_L]}{\mathrm{E}[\tau] + \mathrm{E}[\hat{\tau}]}.$$

Hence, taking expectation and limit in (4), we obtain

$$\lim_{n:=\infty} \frac{\hat{h}_{\mathcal{P}_d(n)}(\alpha)}{n} = \mathrm{E}[\tilde{D}_i] - \frac{\mathrm{E}[\psi_U] + \mathrm{E}[\hat{\psi}_L]}{\mathrm{E}[\tau] + \mathrm{E}[\hat{\tau}]}.$$

Substituting into (2), we arrive at the following result.

PROPOSITION 4. *For a d-chain with postponement level α, such that $2 \leq d \leq n$ and $\alpha \in [0,1]$, its asymptotic chaining efficiency can be computed as follows.*

$$ACE(d,\alpha) = \frac{1}{\mathrm{E}[(\tilde{C} - \tilde{D}_i)^+]} \cdot \left(\mathrm{E}[(\tilde{D}_i - \tilde{C})^+] - \frac{\mathrm{E}[\psi_U] + \mathrm{E}[\hat{\psi}_L]}{\mathrm{E}[\tau] + \mathrm{E}[\hat{\tau}]} \right)$$

The next step is to find a method that can efficiently calculate the values for $\mathrm{E}[\tau], \mathrm{E}[\hat{\tau}], \mathrm{E}[\psi_U]$, and $\mathrm{E}[\hat{\psi}_L]$. This can be easily done by solving a set of linear systems of equations. We refer the readers to Appendix B for details.

As mentioned earlier, the supermodularity property used in Simchi-Levi and Wei (2012) to characterize the performance of 2-chain no longer holds for 3-chain. Hence, to our best knowledge, this is the **first analytical result that characterizes the performance of d-chains for $d \geq 3$.** With this result, we can now examine how $ACE(d,\alpha)$ behaves as d and α change. To illustrate, we suppose that demand for each product follows a normal distribution with mean

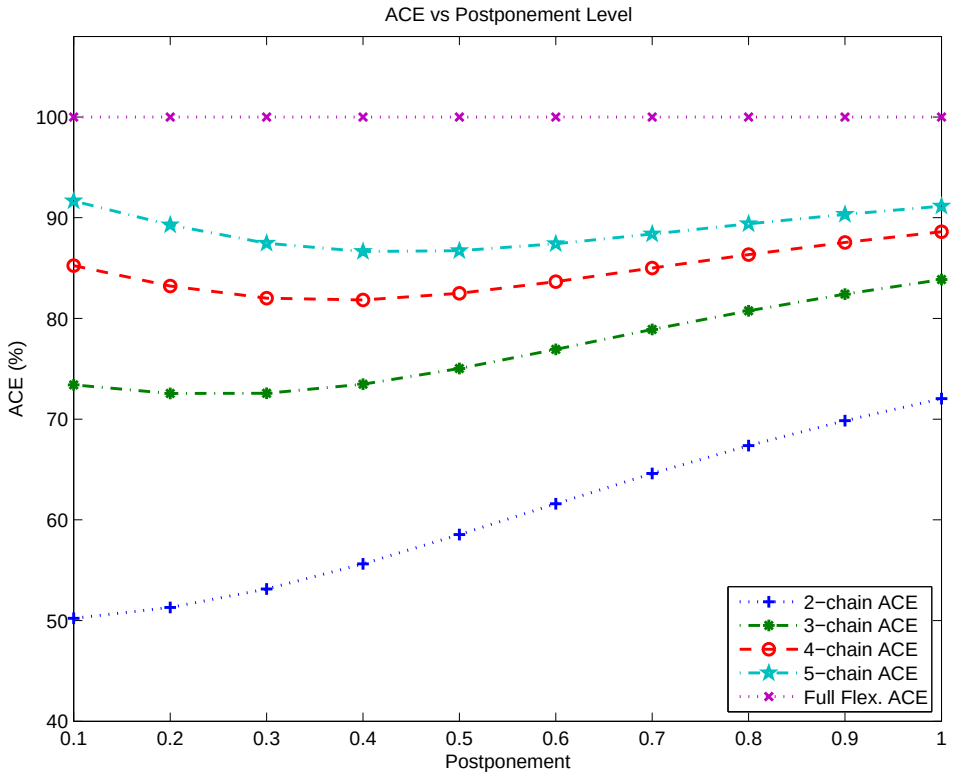


Figure 3 Asymptotic Chaining Efficiency vs Level of Production Postponement

100 units and standard deviation 30 units[3], while capacity for each plant is 100 units. Figure 3 summarizes the asymptotic chaining efficiency for various levels of production postponement ($\alpha = 0.1, 0.2, \ldots, 0.9, 1.0$) and partial flexibility (2-chain, 3-chain, 4-chain, 5-chain, and full flexibility).

Under full postponement, we already expect the 2-chain to perform quite well providing 72% of the benefits of full flexibility even for large production systems. However, under 50% postponement, this number drops to only 58%. This confirms our earlier result that the 2-chain may not be sufficient under partial postponement. Fortunately, adding a third layer of flexible links can restore the performance back to 75%. Adding a fourth layer can bring some benefits (ACE up from 75% to 82%) but significantly less than the gain from 2-chain to 3-chain (from 58% to 75%). We also see that further improvements from the fifth layer (from 82% to 85%) and higher chains are negligible. Such investments are no longer worthwhile, more so in the common scenario where cost of additional flexibility increases in the amount of flexibility already installed. For various other scenarios (normal distribution with other coefficients of variation, and also other demand distributions, say uniform, etc.), we find similar results. That is, **the 2-chain incurs substantial flexibility loss in the case of partial postponement, but the 3-chain recovers most of this flexibility loss.** Lastly, it is important to note that the asymptotic analysis in this section shows that the value of the 3-chain established here is not a mere artifact of the system size.

6. The Asymmetric Case

This section's primary objective is to investigate whether the results obtained thus far carry over to the asymmetric case, where product demands are no longer identically distributed and plant capacities are no longer equal. We consider balanced networks where the number of products equals the number of plants. Moreover, each plant is primarily assigned to produce one product and has capacity equal to that product's mean demand. To achieve some capacity sharing, the 2-chain has been gainfully employed in previous works involving similar settings but with full postponement (see Simchi-Levi 2010, and Simchi-Levi and Wei 2012). Our paper examines the performance of 2-chain and 3-chain under the asymmetric case but with partial postponement. In addition, our second objective is to obtain additional insights that may arise due to system asymmetry. In particular, how does heterogeneity in demand uncertainty affect the flexibility and production decisions?

To this end, we use the Sport Obermeyer example in Hammond and Raman (1996) as an illustration. Table 2 shows ten styles of women's parkas and their respective demand forecasts. For each style i, the demand is assumed to be normally distributed with mean μ_i given in the third column and standard deviation σ_i given in the fourth column. Negative values are truncated at zero.

We consider a production network of 10 facilities, each one primarily assigned to manufacture one style. Also, each facility has enough capacity to meet the expected demand of its primary product, i.e., $C_i = \mu_i$. For example, facility 1 mainly produces the Gail style and has a capacity of 1,017 units. Each facility's capacity is further divided into two parts; namely first-stage capacity employed before actual demand is known, and second-stage capacity to be employed after actual demand is known. As in earlier sections, this capacity split is determined by the postponement level α. Although facilities have their primary style assignments, it would serve the firm well if they can also produce other styles. While full flexibility whereby all facilities can make all styles is most desirable, the firm may only afford a limited amount of process flexibility. Hence, we analyze the performance of 2-chain and 3-chain against full flexibility under varying postponement levels.

Note that SAA method can be used to solve the 2-stage stochastic programming problem, but it can be time-consuming and generates first-stage allocations that are highly variable because the method is sample-based. In practice, heuristic rules are commonly used to determine this first-stage production decision. We consider two heuristic rules which we call (1) the Mean Rule, and (2) the Variance Rule. To explain how they work, we denote the first-stage allocation for style i given postponement level α by $X_i(\alpha) = \sum_{j=1}^{n} x_{ij}(\alpha)$. Clearly, total first-stage allocation should satisfy $\sum_{i=1}^{n} X_i(\alpha) = (1-\alpha)\sum_{i=1}^{n} \mu_i$. The Mean Rule says that the way to allocate the total first-stage capacity of $(1-\alpha)\sum_{i=1}^{n} \mu_i$ among the n different styles is proportional to the mean values of their demands. That is, we produce $X_i(\alpha) = (1-\alpha)\mu_i$ of style i. This is equivalent to the allocation

obtained in Proposition 2. This allocation says that even if we have full flexibility, we only use first-stage capacity for primary production. However, this rule ignores the different variability in demand forecast. The Variance Rule tries to exploit this additional information. It follows the common belief in the accurate response literature (Fisher and Raman 1996) that some styles with high coefficient of variation ought never to be made to stock using speculative (first-stage) capacity. This allocation rule is specified by the following formula.

$$X_i(\alpha) = (\mu_i - \gamma(\alpha)\sigma_i)^+, \forall i = 1, \ldots, n,$$

where

$$\gamma(\alpha) = \frac{\sum_{j\in\mathcal{N}} \alpha\mu_j - \sum_{j\in\mathcal{S}} \mu_j}{\sum_{j\in\mathcal{N}\setminus\mathcal{S}} \sigma_j}$$

and

$$\mathcal{N} := \{1, 2, \ldots, n\} \text{ and } \mathcal{S} := \left\{ i \left| \mu_i < \frac{\sum_{j\in\mathcal{N}} \alpha\mu_j}{\sum_{j\in\mathcal{N}} \sigma_j}\sigma_j, i = 1, \ldots, n \right. \right\}.$$

This rule is similar to Theorem 2 in Fisher and Raman (1996) with a modification which requires first-stage allocation $X_i(\alpha)$ to be non-negative. For our 10-style example in Sport Obermeyer, the first-stage allocation values for various α levels are given in columns 6 to 15 in Table 2. Indeed, the numbers show that as first-stage capacity becomes less (α increases), styles with the highest CVs will be the first ones to receive zero first-stage allocation. For example, the Stephanie style is the first to get zero allocation at $\alpha = 0.6$, followed by the Teri style at $\alpha = 0.7$, and so on.

Table 2 Sport Obermeyer: Product Demand Information and First-Stage Allocation Using Variance Rule

					First-stage production using Variance Rule									
	Style	Mean	SD	CV	α=0.1	α=0.2	α=0.3	α=0.4	α=0.5	α=0.6	α=0.7	α=0.8	α=0.9	α=1
1	Gail	1017	194	19.08%	934.69	852.38	770.06	687.75	605.44	512.90	416.63	230.00	7.25	0.00
2	Isis	1042	323	31.00%	904.95	767.91	630.86	493.82	356.77	202.69	42.41	0.00	0.00	0.00
3	Entice	1358	248	18.26%	1252.78	1147.55	1042.33	937.10	831.88	713.58	590.52	351.93	67.19	0.00
4	Assault	2525	340	13.47%	2380.74	2236.48	2092.22	1947.97	1803.71	1641.52	1472.80	1145.72	755.34	0.00
5	Teri	1100	381	34.64%	938.35	776.69	615.04	453.38	291.73	109.98	0.00	0.00	0.00	0.00
6	Electra	2150	404	18.79%	1978.59	1807.17	1635.76	1464.35	1292.94	1100.22	899.74	511.09	47.23	0.00
7	Stephanie	1113	524	47.08%	890.67	668.34	446.02	223.69	1.36	0.00	0.00	0.00	0.00	0.00
8	Seduced	4017	556	13.84%	3781.10	3545.19	3309.29	3073.38	2837.48	2572.25	2296.35	1761.47	1123.09	0.00
9	Anita	3296	1047	31.77%	2851.77	2407.54	1963.31	1519.08	1074.85	575.40	55.85	0.00	0.00	0.00
10	Daphne	2383	697	29.25%	2087.27	1791.54	1495.81	1200.08	904.35	571.86	226.00	0.00	0.00	0.00

For the first part of our numerical study, we apply the Sample Average Approximation (SAA) method used in Section 4.2. We generate $K = 1,000$ demand scenarios[4] denoted by $\mathbf{D}^k = (D_1^k, D_2^k, \ldots, D_{10}^k)$ for $k = 1, 2, \ldots, K$ such that D_i^k is a random draw from $N(\mu_i, \sigma_i)$ with negative values truncated at zero. We then replace μ in the third constraint of the linear program (P3) with μ_i, set $c_o = c_u = 1, n = 10, K = 1,000$, and solve the linear program to obtain the expected mismatch costs for $\mathcal{C}_d(n)$, for $d = 1, 2, 3$ and 10. The results are shown in Figure 4(a). We observe that similar to the symmetric case, there is substantial flexibility loss between the 2-chain and full flexibility

particularly at intermediate postponement levels. Moreover, the 3-chain once again recovers most of this flexibility loss. This implies that our 3-chain theory carries over to the asymmetric case.

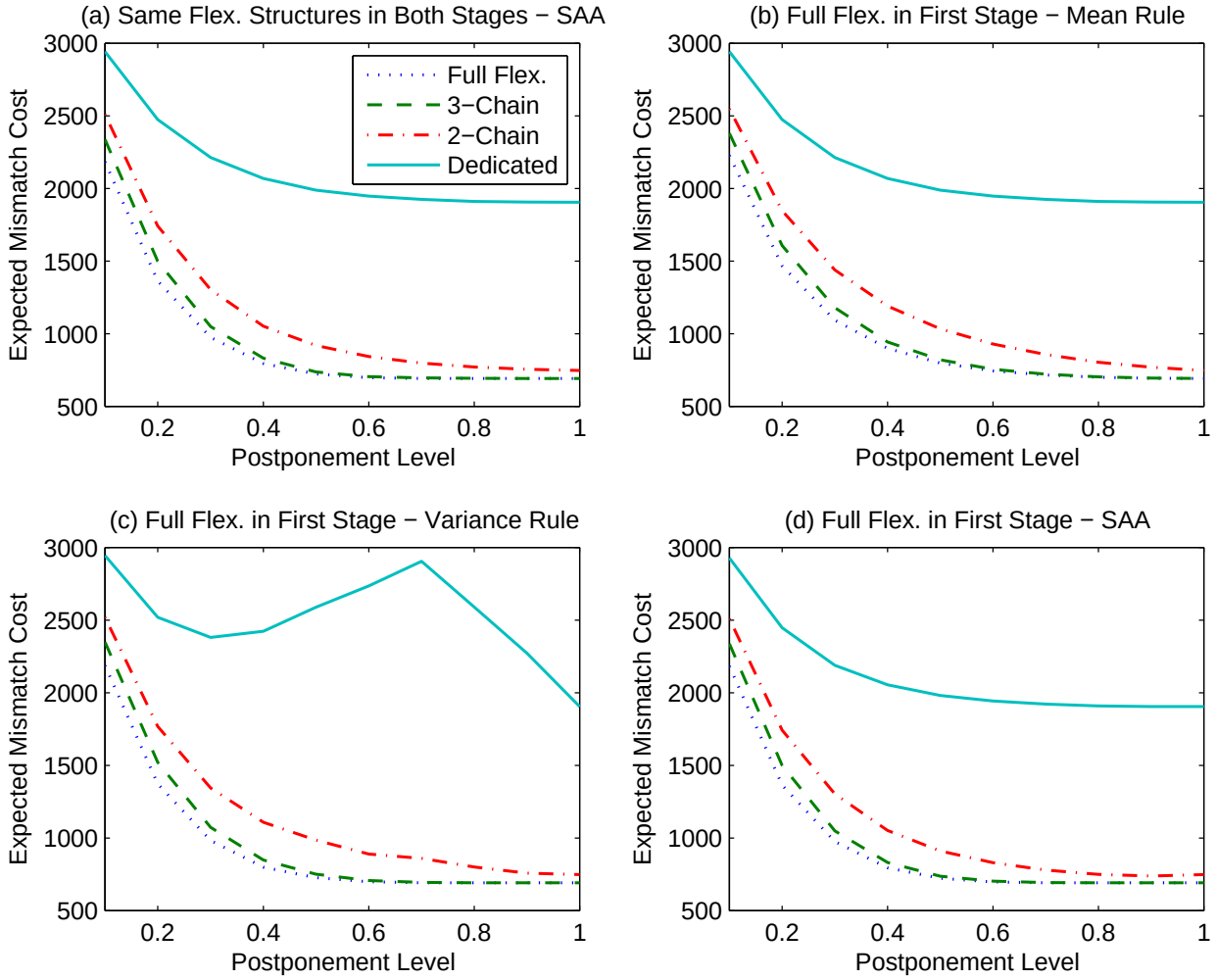


Figure 4 Expected Mismatch Cost vs Postponement Level for Asymmetric Case

For the second part of our numerical study, we relax the assumption that the flexibility structure must be the same in both stages. Specifically, we now consider full flexibility in the first stage regardless of the second-stage structure. This is typically the case for fashion manufacturers like Sport Obermeyer whose first-stage production is carried out in one or two big manufacturing facilities that can produce all styles. However, the second-stage production is usually outsourced to smaller manufacturers, each of which specializes in a limited range of styles. That said, we can again use the SAA method and (P3), but we must remove the constraint $x_{ij} = 0, \forall (i,j) \notin \mathcal{G}(n)$. The results, shown in Figure 4(d), once again corroborate our 3-chain theory.

The Variance Rule used in the literature often assumed that the reserved (second-stage) capacity is fully flexible. As we shall soon see, when there is limited or no flexibility in the second stage, the Variance Rule may no longer perform well. To continue our numerical study, we use the sampled demand scenarios $\mathbf{D}^k$ to obtain the expected mismatch cost values for various second-stage flexibility structures and postponement levels using the two heuristic rules. The results are shown in Figure 4(b) and Figure 4(c). We make two observations here. First, the 3-chain theory is repeatedly supported. Second and more interestingly, the Variance Rule performs poorly when second stage capacity is dedicated. This is because the Variance Rule creates unfair first-stage allocation, leading to imbalance in the second stage which the absence of flexibility cannot handle.

Note that under the symmetric setting, the Mean Rule and the Variance Rule are equivalent. But clearly for the asymmetric case, the two rules prescribe very different first-stage allocations. Hence, we compare the Mean Rule and the Variance Rule viz-a-viz the SAA method and the first-best solution (full flexibility in both stages but under same postponement level) as shown in Figure 5. Interestingly, the Variance Rule does not always help close the gap between the Mean Rule and the SAA method. In fact, the Variance Rule even worsens the performance for the dedicated structure and brings little benefit for the 2-chain. This poor performance will be further magnified when we restore the assumption that flexibility structure in both stages must be the same, because the Mean Rule requires no flexibility in the first stage while the Variance Rule requires full flexibility.

Meanwhile, for 3-chain and full flexibility, the Variance Rule performs extremely well as it almost if not completely closes the gap between the Mean Rule and the SAA method. That this is the case for full flexibility is not surprising because the Variance Rule is optimal when there is full flexibility in both stages (see Theorem 2 in Fisher and Raman 1996). What is more interesting is that the 3-chain can already capture most of this optimal performance, which the 2-chain cannot. This provides us another useful insight: that **the benefit of variance information is largest when the structure is fully flexible and most of this benefit can be achieved by employing the 3-chain**.

Finally, we discuss how expected mismatch cost is influenced by the interaction of all three dimensions: (1) flexibility, (2) postponement, and (3) variance information. To this end, we illustrate using Figure 6, together with parts of Figure 5. In Figure 5(a), if there is no flexibility at all (dedicated system), then even full postponement with the use of Variance Rule performs very poorly (expected cost close to 2,000). In Figure 5(d) and Figure 6, if there is close to no postponement ($\alpha = 0.1$), then even full flexibility with the use of Variance Rule also performs very poorly (expected cost exceeding 2,000). In Figure 6, if variance information is not used (Mean Rule is used), then limited flexibility (2-chain) with sufficient postponement ($\alpha = 0.5$) can provide reasonable but not

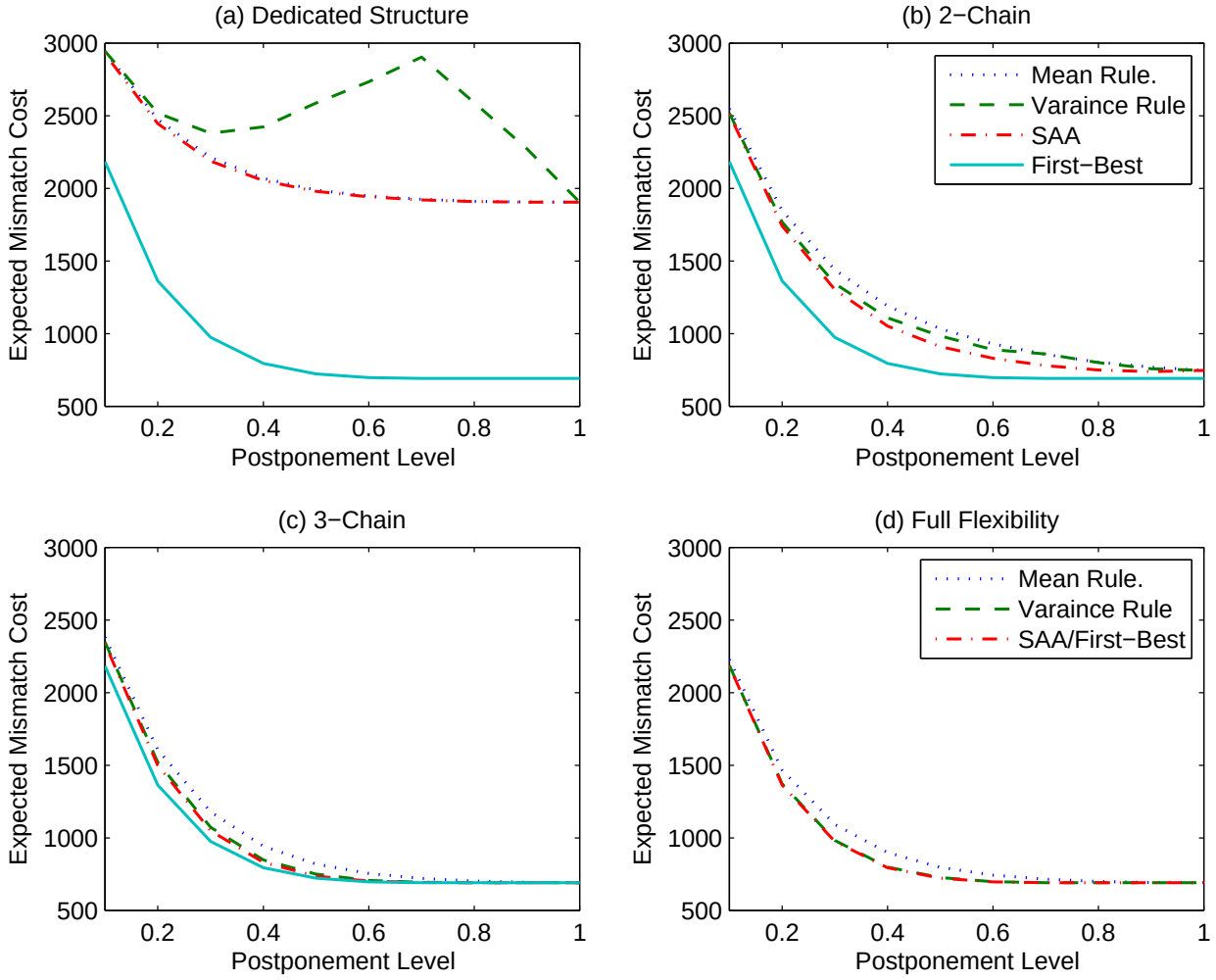


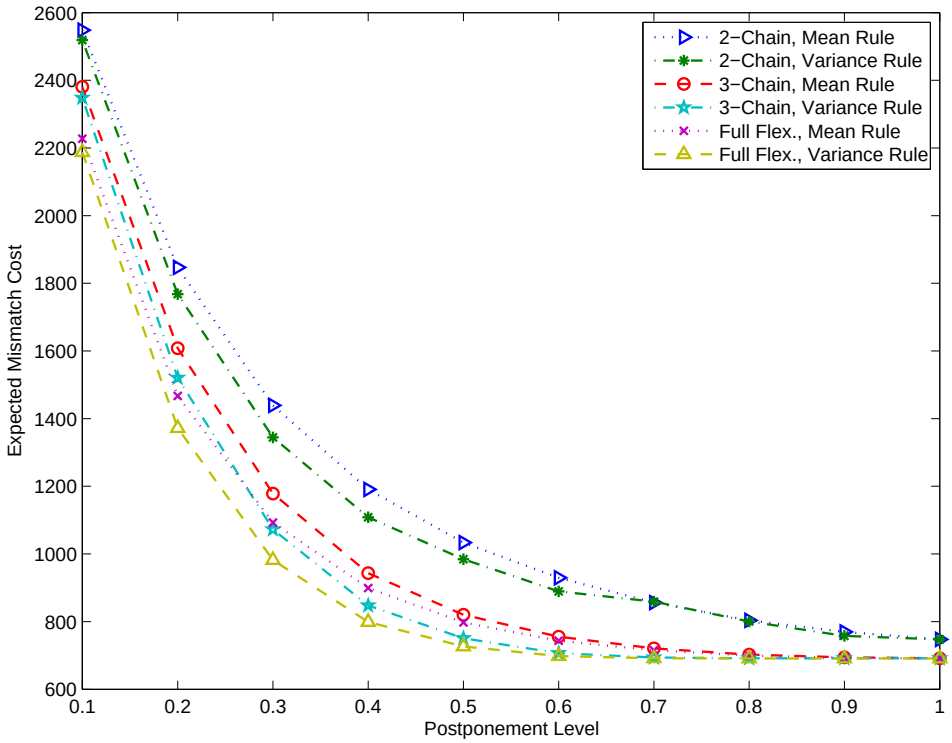
Figure 5 Value of Variance Information for Various Flexibility and Postponement Levels

near-optimal performance (expected cost close to 1,000). Moreover, 3-chain with Variance Rule outperforms full flexibility with Mean Rule for $\alpha > 0.3$, but the same cannot be said for 2-chain with Variance Rule compared to 3-chain with Mean Rule. Hence, all three dimensions are critical and must be properly planned to achieve low expected mismatch cost without overinvesting.

7. Conclusion

In this paper, we study both process flexibility and production postponement, and their effectiveness in mitigating the uncertainty and complexity prevalent in global production and consumption networks. Because the first-best solution of full flexibility and full postponement is very expensive, one approximate solution suggested in the literature is full postponement with partial flexibility. The performance of this solution is almost as good as the first-best solution but incurs only a small

Figure 6 Expected Mismatch Cost as Function of Postponement, Flexibility, and Variance Information



fraction of the cost. In this study, we examine if other solutions can do the same. In particular, we consider the effect of partial postponement on the benefits and design of process flexibility.

To this end, we develop a multi-item newsvendor model with second-stage supply and partial capacity sharing in order to minimize the expected mismatch cost. Subsequently, we define optimality loss of any solution as the gap between that solution and the first-best solution. This loss is further broken down into postponement loss and flexibility loss which are defined as the losses due to partial postponement and partial flexibility, respectively. Our first result shows that full flexibility with low partial postponement ($\alpha << 1$) could not attain the same order of performance as full postponement with partial flexibility. However, if a substantial amount (say $\alpha \geq 0.5$) of capacity can be postponed, then full flexibility with partial postponement can also approximate the first-best solution, albeit at a much higher installation cost.

Having established that the flexibility losses of partial flexibility (2-chain) at full postponement and no postponement are both negligible or zero, we discover that these results no longer hold when postponement is partial. For example, in a 10×10 system, we report that for postponement levels between 10% and 50%, the flexibility loss is quite sizable, ranging from 20% to 30%. In these scenarios, we find that the 3-chain not only recovers most of the flexibility loss, but sometimes, even parts of the postponement loss. In the 10×10 example, the 3-chain with 50% postponement

already restores the flexibility loss to the same level as 2-chain with full postponement. Furthermore, we extend the random walk approach in Chou et al. (2010b) to obtain the asymptotic chaining efficiency of any d-chain at arbitrary postponement levels. Using this method, we demonstrate that the value of the 3-chain we established for small systems is valid even for extremely large systems. Moreover, further flexibility upgrades (e.g. fourth or fifth chain) can no longer produce as much benefit and usually incurs even higher flexibility installation costs.

In conclusion, all our results strongly suggest that the 3-chain brings substantial value in the face of partial postponement. As is well known in the community, the 2-chain in flexible production systems proves effective because it takes care of baseline uncertainty in the product demand. We have, in this paper, extended that theory by providing evidence of the value of the 3-chain, that it can be used to compensate for the flexibility loss brought about by lost postponement. Finally, that chains higher than the 3-chain are unnecessary also supports the belief that even for large systems with partial postponement, one still only needs a sparse structure (3-chain) to achieve most of the benefits of the first-best solution.

Endnotes

1. www.interbrand.com
2. It is important to note that while the expected shortfall minimization (equivalently, expected flow maximization) in the second stage of this model may appear to be a single period decision, the model can in fact be used for the expected performance over multiple independent periods.
3. Note that Corollary 1 in Chou et al. (2010b) still holds for the d-chain where $d > 2$ and any postponement level $\alpha \in (0, 1)$, implying the invariance of $ACE(d, \alpha)$ over the scale of demand. Hence, the results presented here is valid for any demand distribution that is normal with coefficient of variation equal to 0.30.
4. Based on our numerical tests, we obtain similar patterns for number of demand scenarios larger than 1,000.

Acknowledgments

We would like to thank the area editor, the associate editor, and two anonymous referees for their valuable comments and suggestions that helped improve this paper. We would also like to give special thanks to Prof. Chung-Piaw Teo for his helpful suggestions on an earlier draft, continuing interest and encouragement. We also thank him for introducing a simpler proof for Proposition 2 and pointing out a gap in our original proof for Proposition 3. This research was supported in part by A*STAR SERC Grant 1122904020, R-314-000-091-305, NUS Academic Research Fund R-314-000-085-112, National Science Foundation of China No. 71371119, and Shanghai Chen Guang Program.

References

[1] Akşin, O., F. Karaesmen. 2007. Characterizing the performance of process flexibility structures. *Operations Research Letters* 35(4) 477-484.

[2] Bassamboo A., R. Randhawa, J. A. Van Mieghem. 2010. Optimal flexibility configurations in newsvendor networks: Going beyond chaining and pairing. *Management Science* 56(8) 1285-1303.

[3] Bassamboo A., R. Randhawa, J. A. Van Mieghem. 2012. A little flexibility is all you need: On the asymptotic value of flexible capacity in parallel queuing systems. *Operations Research* 60(6) 1423-1435.

[4] Bish, E., A. Muriel, S. Biller. 2005. Managing flexible capacity in a make-to-order environment. *Management Science* 51(2) 167-180.

[5] Boyd, S., L. Vandenberghe. 2009. *Convex optimization.* Cambridge University Press, UK.

[6] Cachon, G., C. Terwiesch. 2009. *Matching Supply with Demand: An Introduction to Operations Management.* McGraw-Hill.

[7] Chou, M.C., G.A. Chua, C.P. Teo. 2010a. On range and response: Dimensions of process flexibility. *European Journal of Operational Research* 207(2) 711-724.

[8] Chou, M.C., G.A. Chua, C.P. Teo, H. Zheng. 2010b. Design for process flexibility: Efficiency of the long chain and sparse structure. *Operations Research* 58(1) 43-58.

[9] Chou, M.C., G.A. Chua, C.P. Teo, H. Zheng. 2011. Process flexibility revisited: The graph expander and its applications. *Operations Research* 59(5) 1090-1105.

[10] Chou, M.C., C.P. Teo, H. Zheng. 2008. Process flexibility: Design, evaluation, and applications. *Flexible Services and Manufacturing Journal* 20(1-2) 59-94.

[11] Deng, T., Z.J. Shen. 2013. Process flexibility design in unbalanced networks. *Manufacturing & Service Operations Management* 15(1) 24-32.

[12] Eppen, G., 1979. Effects of centralization on expected costs in a multilocation newsboy problem. *Management Science* 25(5) 498-501.

[13] Feng, B.Y. 2007. *100 Years of Li & Fung: Rise from Family Business to Multinational.* Thomson Learning.

[14] Fisher, M., A. Raman. 1996. Reducing the cost of demand uncertainty through accurate response to early sales. *Operations Research* 44(1) 87-99.

[15] Fung, V., W. Fung, Y. Wind. 2007. Competing in a flat world: The perils and promise of global supply chains. *ChangeThis* 40, 1-9. November 14.

[16] Fung, V., W. Fung, Y. Wind. 2008. *Competing in a Flat World: Building Enterprises for a Borderless World.* Wharton School Publishing.

[17] Graves, S., B. Tomlin. 2003. Process flexibility in supply chains. *Management Science* 49(7) 907-919.

[18] Gurumurthi, S., S. Benjaafar. 2004. Modeling and analysis of flexible queueing systems. *Naval Research Logistics* 51(5) 755-782.

[19] Hammond, J.H., A. Raman. 1996. Sport Obermeyer, Ltd. *Harvard Business School Case (9-695-022).* Boston, MA. 1-21.

[20] Hopp, W., E. Tekin, M. Van Oyen. 2004. Benefits of skill chaining in serial production lines with cross-trained workers. *Management Science* 50(1) 83-98.

[21] Iravani, S., M. Van Oyen, K. Sims. 2005. Structural flexibility: A new perspective on the design of manufacturing and service operations. *Management Science* 51(2) 151-166.

[22] Jordan, W., S. Graves. 1995. Principles on the benefits of manufacturing process flexibility. *Management Science* 41(4) 577-594.

[23] Lee, H.L., C.A. Billington and B. Carter. 1993. Hewlett Packard gains control of inventory and service through design for localization. *Interfaces* 23(4) 1-11.

[24] Lee, H.L., C.S. Tang. 1997. Modeling the costs and benefits of delayed product differentiation. *Management Science* 43(1) 40-53.

[25] Muriel, A., A. Somasundaram, Y. Zhang. 2006. Impact of partial manufacturing flexibility on production variability. *Manufacturing & Service Operations Management* 8(2) 192-205.

[26] Signorelli, S., J. L. Heskett. 1984. Benetton (A) and (B). *Harvard Business School Case* (9-685-014), Boston, MA. 1-20.

[27] Simchi-Levi, D. 2010. *Operations Rules: Delivering Customer Value through Flexible Operations.* MIT Press.

[28] Simchi-Levi, D., Y. Wei. 2012. Understanding the Performance of the Long Chain and Sparse Designs in Process Flexibility. *Operations Research*, To appear.

[29] Swaminathan, J. M., H. L. Lee. 2003. Design for Postponement. A. G. de Kok, S. C. Graves, eds. *Handbook of OR/MS in Supply Chain Management.* Chap. 5. Elsevier, Amsterdam, The Netherlands.

[30] Van Biesebroeck, J., 2007. Complementarities in automobile production. *Journal of Applied Econometrics* 22(7) 1315-1345.

[31] Van Mieghem, J.A., M. Dada. 1999. Price versus product postponement: Capacity and competition. *Management Science* 45(12) 1631-1649.

Appendix

A. Proof of Proposition 3

Define $\hat{h}_{\mathcal{G}(n)}(\alpha, C) = \mathrm{E}[\tilde{h}_{\mathcal{G}(n)}(\alpha, \boldsymbol{D}, C)]$ and

$$\begin{aligned}\tilde{h}_{\mathcal{G}(n)}(\alpha, \boldsymbol{D}, C) = \max_{\boldsymbol{y}} & \sum_{i=1}^{n}\sum_{j=1}^{n} y_{ij} \\ \text{s.t.} \quad & \sum_{j=1}^{n} y_{ij} \le (D_i - (1-\alpha)C)^+ \quad \forall i = 1, 2, \dots n \\ & \sum_{i=1}^{n} y_{ij} \le \alpha C \quad \forall j = 1, 2, \dots n \\ & y_{ij} \ge 0 \quad \forall i, j = 1, 2, \dots n \\ & y_{ij} = 0 \quad \forall (i,j) \notin \mathcal{G}(n)\end{aligned}$$

so that $L_F(\mathcal{C}_2(n), \alpha, C) = c_u[\hat{h}_{\mathcal{F}(n)}(\alpha, C) - \hat{h}_{\mathcal{C}_2(n)}(\alpha, C)]$.

We want to show that

$$\frac{\partial}{\partial \alpha} L_F(\mathcal{C}_2(n), 1^-, C) = c_u \left[\frac{\partial}{\partial \alpha} \hat{h}_{\mathcal{F}(n)}(1^-, C) - \frac{\partial}{\partial \alpha} \hat{h}_{\mathcal{C}_2(n)}(1^-, C) \right] < 0$$

To this end, we obtain

$$\begin{aligned}\frac{\partial}{\partial \alpha} \hat{h}_{\mathcal{G}(n)}(1^-, C) &= \lim_{\delta \to 0^+} \frac{\mathrm{E}[\tilde{h}_{\mathcal{G}(n)}(1, \boldsymbol{D}, C) - \tilde{h}_{\mathcal{G}(n)}(1-\delta, \boldsymbol{D}, C)]}{\delta} \\ &= \lim_{\delta \to 0^+} \frac{\mathrm{E}[\tilde{h}_{\mathcal{G}(n)}(1, \boldsymbol{D}, C) - \tilde{h}_{\mathcal{G}(n)}(1-\delta, \boldsymbol{D}, C) | D_i \ge \delta C, \forall i]}{\delta} \cdot \mathrm{P}\{D_i \ge \delta C, \forall i\} \\ &\quad + \lim_{\delta \to 0^+} \frac{\mathrm{E}[\tilde{h}_{\mathcal{G}(n)}(1, \boldsymbol{D}, C) - \tilde{h}_{\mathcal{G}(n)}(1-\delta, \boldsymbol{D}, C) | \exists i : D_i < \delta C]}{\delta} \cdot (1 - \mathrm{P}\{D_i \ge \delta C, \forall i\}) \\ &= \lim_{\delta \to 0^+} \frac{\mathrm{E}[\tilde{h}_{\mathcal{G}(n)}(1, \boldsymbol{D}, C) - \tilde{h}_{\mathcal{G}(n)}(1-\delta, \boldsymbol{D}, C) | D_i \ge \delta C, \forall i]}{\delta} \cdot \lim_{\delta \to 0^+} \mathrm{P}\{D_i \ge \delta C, \forall i\} \\ &\quad + \lim_{\delta \to 0^+} \frac{\mathrm{E}[\tilde{h}_{\mathcal{G}(n)}(1, \boldsymbol{D}, C) - \tilde{h}_{\mathcal{G}(n)}(1-\delta, \boldsymbol{D}, C) | \exists i : D_i < \delta C]}{\delta} \cdot \lim_{\delta \to 0^+} (1 - \mathrm{P}\{D_i \ge \delta C, \forall i\}) \\ &= \lim_{\delta \to 0^+} \frac{\mathrm{E}[\tilde{h}_{\mathcal{G}(n)}(1, \boldsymbol{D}, C) - \tilde{h}_{\mathcal{G}(n)}(1-\delta, \boldsymbol{D}, C) | D_i \ge \delta C, \forall i]}{\delta}\end{aligned}$$

The last equation is due to demand non-negativity, hence $\lim_{\delta \to 0^+} \mathrm{P}\{D_i \ge \delta C, \forall i\} = 1$. It follows that for full flexibility, we have

$$\begin{aligned}\frac{\partial}{\partial \alpha} \hat{h}_{\mathcal{F}(n)}(1^-, C) &= \lim_{\delta \to 0^+} \frac{\mathrm{E}[\tilde{h}_{\mathcal{F}(n)}(1, \boldsymbol{D}, C) - \tilde{h}_{\mathcal{F}(n)}(1-\delta, \boldsymbol{D}, C) | D_i \ge \delta C, \forall i]}{\delta} \\ &= \lim_{\delta \to 0^+} \frac{\mathrm{E}[\min(\sum_{i=1}^n D_i, nC) - \min(\sum_{i=1}^n (D_i - \delta C), n(1-\delta)C) | D_i \ge \delta C, \forall i]}{\delta} \\ &= \lim_{\delta \to 0^+} \frac{\mathrm{E}[\min(\sum_{i=1}^n D_i, nC) - \min(\sum_{i=1}^n D_i, nC) + n\delta C | D_i \ge \delta C, \forall i]}{\delta} \\ &= nC\end{aligned}$$

For the 2-chain, we let $\boldsymbol{d}$ be a given demand realization such that $d_i \ge \delta C, \forall i$, and $y_{ij}(\alpha, \boldsymbol{d}, C)$ and $y^*_{ij}(\alpha, \boldsymbol{d}, C)$ be a feasible solution and an optimal solution, respectively, for $\tilde{h}_{\mathcal{C}(n)}(\alpha, \boldsymbol{d}, C)$. Since $y_{ii}(1, \boldsymbol{d}, C) = y^*_{ii}(1-\delta, \boldsymbol{d}, C) + \delta C, \forall i$ and $y_{ij}(1, \boldsymbol{d}, C) = y^*_{ij}(1-\delta, \boldsymbol{d}, C), \forall i \ne j$ is a feasible solution to $\tilde{h}_{\mathcal{C}(n)}(1, \boldsymbol{d}, C)$, it follows that $\tilde{h}_{\mathcal{C}(n)}(1, \boldsymbol{d}, C) - \tilde{h}_{\mathcal{C}(n)}(1-\delta, \boldsymbol{d}, C) \ge n\delta C$.

To rule out $\mathrm{E}[\tilde{h}_{\mathcal{C}(n)}(1, \boldsymbol{D}, C) - \tilde{h}_{\mathcal{C}(n)}(1-\delta, \boldsymbol{D}, C) | D_i \ge \delta C, \forall i] = n\delta C$, it can be verified that there exists a demand realization $\boldsymbol{d} \ge \delta C \boldsymbol{1}$ such that $d_k + d_{k+1} + d_{k+2} + d_{k+3} \le 4C$ but $d_k + d_{k+1} > 3C - \delta C$ for some k. In

this case, an increase in α from $1-\delta$ to 1 will result in an increase in $\tilde{h}_{\mathcal{C}(n)}(.)$ that is strictly greater than $n\delta C$, i.e.,

$$\tilde{h}_{\mathcal{C}(n)}(1,\boldsymbol{d},C)-\tilde{h}_{\mathcal{C}(n)}(1-\delta,\boldsymbol{d},C)>n\delta C.$$

It follows that

$$\begin{aligned}\frac{\partial}{\partial\alpha}\hat{h}_{\mathcal{C}(n)}(1^-,C) &= \lim_{\delta\to 0^+}\frac{\mathrm{E}[\tilde{h}_{\mathcal{C}(n)}(1,\boldsymbol{D},C)-\tilde{h}_{\mathcal{C}(n)}(1-\delta,\boldsymbol{D},C)|D_i\geq\delta C,\forall i]}{\delta}\\ &> \lim_{\delta\to 0^+}\frac{n\delta C}{\delta}=nC\end{aligned}$$

We have shown $\frac{\partial}{\partial\alpha}L_F(\mathcal{C}_2(n),1^-,C)<0$. Since $L_F(\mathcal{C}_2(n),1,C)>0=L_F(\mathcal{C}_2(n),0,C)$, it suffices to check if $L_F(\mathcal{C}_2(n),\alpha,C)>L_F(\mathcal{C}_2(n),1,C)$ for some $\alpha\in(0,1)$. Because $\frac{\partial}{\partial\alpha}L_F(\mathcal{C}_2(n),1^-,C)<0$, the result follows. □

B. Linear Systems for Calculating $ACE(d,\alpha)$

In this section, we present a method that can efficiently calculate the values for $\mathrm{E}[\tau],\mathrm{E}[\hat{\tau}],\mathrm{E}[\psi_U]$, and $\mathrm{E}[\hat{\psi_L}]$. To this end, we assume that for each i, the support of $\tilde{D}_i$ lies in $\{\frac{j}{N}\cdot(1+\alpha)\mu|j=0,1,2,\ldots,N\}$ where $N\geq 1$ denotes the level of discretization on the demand distribution. Moreover, we let $p_j=\mathrm{P}(\tilde{D}_i=\frac{j}{N}\cdot(1+\alpha)\mu),\forall j=0,1,2,\ldots,N-1,p_N=\mathrm{P}(\tilde{D}_i\geq(1+\alpha)\mu)$, and $p_j=0,\forall j=N+1,N+2,\ldots$. On the other hand, to represent capacity by the same discretization level, we rewrite capacity as $\tilde{C}=\alpha\mu=\lfloor\frac{\alpha N}{1+\alpha}\rfloor$. That is, there are $\tilde{C}=\lfloor\frac{\alpha N}{1+\alpha}\rfloor$ units of capacity in each plant just as there are $N+1$ possible demand states for each product.

Next, we define τ_x (resp, $\hat{\tau}_x$) as the stopping time if the random walk $\{\hat{S}_i,i=0,1,2,\ldots\}$ is currently in an odd (resp, even) cycle at a state x. We also define ψ_x (resp, $\hat{\psi}_x$) as the overshoot at the upper (resp, lower) boundary if the random walk is currently in an odd (resp, even) cycle at a state x. The value of the state x can range from 0 to $(d-1)\tilde{C}-1$. Hence, we further define the following $(d-1)\tilde{C}\times 1$ vectors $\mathbf{v},\hat{\mathbf{v}},\mathbf{w},\hat{\mathbf{w}}$ to collect the expected values $\mathrm{E}[\tau_x],\mathrm{E}[\hat{\tau}_x],\mathrm{E}[\psi_x],\mathrm{E}[\hat{\psi}_x]$, respectively, for all x. That is, $v_{x+1}=\mathrm{E}[\tau_x]$,$\hat{v}_{x+1}=\mathrm{E}[\hat{\tau}_x]$,$w_{x+1}=\mathrm{E}[\psi_x]$, and $\hat{w}_{x+1}=\mathrm{E}[\hat{\psi}_x]$, for $x=0,1,\ldots,(d-1)\tilde{C}-1$. Most importantly, for each of these four vectors, we can condition on the next move of the random walk starting at state x and obtain a system of linear equations. Solving these systems gives us the values of the four vectors. Hence, we arrive at the following result.

Proposition 5. *For a d-chain with postponement level α, such that $2\leq d\leq n$ and $\alpha\in[0,1]$, the values of $\mathrm{E}[\tau],\mathrm{E}[\hat{\tau}],\mathrm{E}[\psi_U]$, and $\mathrm{E}[\hat{\psi_L}]$ can be obtained by solving the following systems of linear equations*

$$\mathbf{v}-\mathbf{M}\mathbf{v}=\mathbf{1},\mathbf{w}-\mathbf{M}\mathbf{w}=\mathbf{r},\hat{\mathbf{v}}-\hat{\mathbf{M}}\hat{\mathbf{v}}=\mathbf{1},\hat{\mathbf{w}}-\hat{\mathbf{M}}\hat{\mathbf{w}}=\hat{\mathbf{r}}$$

where $\mathbf{M},\hat{\mathbf{M}}$ are $(d-1)\tilde{C}\times(d-1)\tilde{C}$ matrices, $\mathbf{v},\mathbf{w},\hat{\mathbf{v}},\hat{\mathbf{w}},\mathbf{r},\hat{\mathbf{r}}$ are $(d-1)\tilde{C}\times 1$ vectors, and

$$M_{k,l}=\begin{cases}p_{\tilde{C}+l-k} & \forall k=\max(\tilde{C}-N+l,1),\ldots,\min(\tilde{C}+l,(d-1)\tilde{C}),\forall l=2,\ldots,(d-1)\tilde{C}\\ \sum_{j=0}^{\tilde{C}+1-k}p_j & \forall k=1,2,\ldots,\tilde{C}+1,l=1\\ 0 & \textit{otherwise}\end{cases}\tag{5}$$

$$\hat{M}_{k,l}=\begin{cases}p_{\tilde{C}+k-l} & \forall k=\max(-\tilde{C}+l,1),\ldots,\min(N-\tilde{C}+l,(d-1)\tilde{C}),\forall l=2,\ldots,(d-1)\tilde{C}\\ \sum_{j=\tilde{C}-1+k}^{N}p_j & \forall k=1,2,\ldots,N-\tilde{C}+1,l=1\\ 0 & \textit{otherwise}\end{cases}\tag{6}$$

$$r_k=\begin{cases}\sum_{j=d\tilde{C}+2-k}^{N}(j-d\tilde{C}-1+k)p_j & \forall k=\max(d\tilde{C}+2-N,1),\ldots,(d-1)\tilde{C}\\ 0 & \textit{otherwise}\end{cases}\tag{7}$$

$$\hat{r}_k=\begin{cases}\sum_{j=\tilde{C}+k}^{N}(j-\tilde{C}+1-k)p_j & \forall k=1,\ldots,\min(N-\tilde{C},(d-1)\tilde{C})\\ 0 & \textit{otherwise}\end{cases}\tag{8}$$

and assigning $\mathrm{E}[\tau]=v_1,\mathrm{E}[\hat{\tau}]=\hat{v}_1,\mathrm{E}[\psi_U]=w_1$, and $\mathrm{E}[\hat{\psi_L}]=\hat{w}_1$.