Commissioned Paper

Capacity Management, Investment, and Hedging: Review and Recent Developments

Jan A. Van Mieghem
Kellogg School of Management, Northwestern University,
Evanston, Illinois 60208-2009
vanmieghem@kellogg.northwestern.edu

This paper reviews the literature on strategic capacity management concerned with determining the sizes, types, and timing of capacity investments and adjustments under uncertainty. Specific attention is given to recent developments to incorporate multiple decision makers, multiple capacity types, hedging, and risk aversion. Capacity is a measure of processing abilities and limitations and is represented as a vector of stocks of various processing resources, while investment is the change of capacity and includes expansion and contraction. After discussing general issues in capacity investment problems, the paper reviews models of capacity investment under uncertainty in three settings:

The first reviews optimal capacity investment by single and multiple risk-neutral decision makers in a stationary environment where capacity remains constant. Allowing for multiple capacity types, the associated optimal capacity portfolio specifies the amounts and locations of safety capacity in a processing network. Its key feature is that it is unbalanced; i.e., regardless of how uncertainties are realized, one typically will never fully utilize all capacities. The second setting reviews the adjustment of capacity over time and the structure of optimal investment dynamics. The paper ends by reviewing how to incorporate risk aversion in capacity investment and contrasts hedging strategies involving financial versus operational means.

(*Capacity; Investment; Expansion; Planning; Real Options; Hedging; Risk; Mean-Variance*)

1. Introduction

This paper reviews the literature on strategic capacity management concerned with determining the sizes, types, and timing of capacity adjustments under uncertainty. Specific attention is given to recent developments to incorporate multiple decision makers, multiple capacity types, hedging, and risk aversion. Given the diverse interpretations and application domains of capacity,[1] any review must be selective. *Capacity* typically describes abilities and limitations. In operations management, it is natural to

[1] There are more than 15,000 peer-reviewed articles in the ProQuest Business Databases with "capacity" in the title or key words, 2,632 of them published during 1999–2002 alone. Restricting the search by adding "planning," "expansion," or "investment" to "capacity" retained more than 5,000 articles.

1523-4614/03/0504/0269
1526-5498 electronic ISSN

consider a network of various processing resources, also called a *processing network*. The types and amounts of these resources, which are the prime economic factors of production, are important determinants of the network's production abilities and limitations. Deciding on the types of resources relates to processing-network design in operations research, or characterizing the production and operating-profit functions in economics. Deciding on the amounts involves optimization of a given network or operating-profit function. While many other factors, such as inventory supply and energy shortages, quality and yield losses, scheduling, lead times, and reliability, may also limit production, this paper will turn up the brightness on the direct effects of resource scarcity and uncertainty, and turn it down on other factors. In this paper then, capacity is a measure of processing abilities and limitations that stem from the scarcity of various processing resources and is represented as a vector of stocks of various processing resources. The operational consequence is that capacities in this paper almost always can be interpreted as some upper bounds on processing quantities, but a more general, higher-level economics interpretation that links capacity directly to operating profits will also be discussed.

While *capacity* refers to stocks of various resources, *investment* refers to the change of that stock over time. Investment[2] thus involves the monetary flow stemming from capacity expansion and contraction in the expectation of future rewards. According to Dixit and Pindyck (1994), most investment decisions share three important characteristics in varying degrees. First, the investment is partially or completely irreversible in that one cannot recover its full cost should one have a change of mind. Second, there is uncertainty over the future rewards from the investment. Third, there is some leeway about the timing or dynamics of the investment. In addition to these three, this paper adds a fourth characteristic: multidimensionality. Typically, a firm invests in multiple types of resources that have different financial and operational properties. Decisions about the types and levels of investment are interdependent and the firm's productive capabilities depend on the complete vector of capacity levels, which we will call its *capacity portfolio*.[3]

Our outline and objectives are as follows. Section 2 reviews various literatures that deal with capacity investment. Section 3 reviews important issues in the formulation of any capacity problem. The remainder of the paper reviews optimal capacity investment under uncertainty that is partially irreversible.

Section 4 reviews optimal capacity investment by risk-neutral decision maker(s) in a stationary environment where capacity remains constant over time. The emphasis of this section is on determining the optimal levels of various types of capacities and the trade-offs among them. The optimal portfolio specifies the amounts and locations of safety capacity in the network. Section 4.1 reviews canonical capacity models that adopt queuing or newsvendor network formulations, an example of which is studied in §4.2. Section 4.3 reviews game-theoretic capacity investment by multiple agents who each control a subset of the processing network.

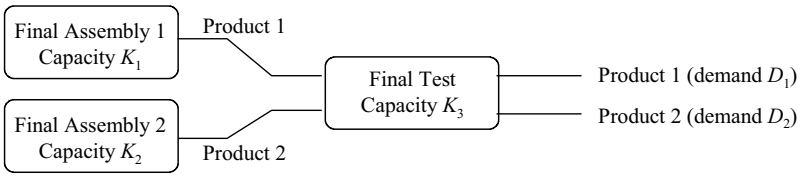
Section 5 reviews dynamic capacity investment models whose emphasis is on characterizing the timing of capacity adjustments. Often, optimal investment dynamics follow a so-called "ISD policy," which is characterized by a continuation region: When the capacity vector falls in this region, it is optimal not to adjust capacity; otherwise, capacity should be adjusted to an appropriate point on the region's boundary. Managerially, this implies that capacity is not adjusted continuously because of several "frictions," such as irreversibility, nonconvex costs, or lumpiness, that will be discussed in §5.3. In addition, the optimal investment sequence defines an endogenous relative flexibility of resource capacity, meaning that some resource capacities will be more frequently adjusted than others.

Section 6 incorporates risk aversion in capacity investment and reviews hedging strategies involv-

[2] Originally, investment was "the act of putting clothes or vestments on." (Shakespeare 1597, 2 Hendrick IV, iv. i. 45)

[3] For the remainder, the term "capacity portfolio" is assumed to imply interdependent resource capacity decisions. Interdependence may stem from, for example, some resource sharing among products or dynamic routing in the processing network. Otherwise, the processing network may be separable and the portfolio problem may decompose into independent, lower-dimensional problems.

Figure 1 An Example of a Capacity Portfolio Investment Problem



ing financial versus operational means. For managers, uncertainty, quite naturally, leads to considerations of notions of risk and methods to manage and mitigate that risk. Section 6.1 reviews the economic theory whether firms should care about risk, and if so, how to incorporate it in capacity decisions. We review expected utility formulations and the conditions under which they lead to mean-variance formulations. Section 6.2 reviews financial hedging and discusses how capacity investment is similar to selling a call option. Hedging means "to protect oneself from losing or failing by a counterbalancing action" or "to protect oneself financially as: (a) To buy or sell commodity futures as a protection against loss due to price fluctuation; (b) to minimize the risk of a bet" (*Merriam-Webster's Collegiate Dictionary* 1998). We refer to hedging that uses counterbalancing positions in futures and financial derivative instruments as *financial hedging*. Section 6.3 reviews *operational hedging*, by which we mean mitigating risk by counterbalancing actions in the processing network that do not involve financial instruments. Operational hedging, thus, may include various types of processing flexibility, such as dual-sourcing, component commonality, having the option to run overtime, dynamic substitution, routing, transshipping, or shifting processing among different types of capital, locations, or subcontractors, holding safety stocks, having warranty guarantees, etc. Some of these actions (e.g., multisourcing, component commonality, product-flexible resources, and having alternate uses of resources) aim to pool the safety capacity of multiple resources. "Counterbalancing their capacities" to mitigate risk is exactly the form of operational hedging that is studied in this paper. Finally, §7 concludes.

To motivate the study of stochastic capacity portfolio investment and hedging, consider the following example.

Example. Consider a stylized representation of the capacity-planning problem that disk-drive manufacturers routinely face, as described by Van Mieghem (1998b) and illustrated in Figure 1. To support their growth and frequent new product and technology introductions, such companies repeatedly make investments in property and equipment.[4] Assume two high-end disk-drive product families are scheduled to go into volume production in the first calendar quarter of next year. Product 1 boasts faster seek time and higher data-throughput rate, while Product 2 is more energy-efficient and reliable. Given their different product designs, each family's "head-disk assembly" (HDA) and printed-circuit board final assembly requires its own product-specific equipment. Both families, however, can be tested in a single, shared facility. Disk-drive testing involves connecting the drive to intelligent drive testers (IDTs), fast computers that perform a set of read-write tests. IDTs can quickly switch over between testing different products. Product life cycles of disk drives are short. The two new products are planned to be in volume production for only four quarters. There are significant fixed costs associated with commissioning and starting up the three new facilities. In addition to the fixed costs, the facility costs are also driven by the volume capacity, reflecting higher labor, space requirements, and tooling costs.

Capital investments at such a disk drive company are typically the result of a capacity requirements planning (CRP) process, which links with the production-planning and materials requirements planning (MRP) process. The monthly "demand-planning" cycle begins with individual marketing and

[4] For example, Seagate Technologies invested $920 million in fiscal 1997. This amount included $301 million for manufacturing facilities and equipment related to subassembly and disc-drive final assembly and test (FA&T) facilities.

sales managers using their knowledge of planned promotions and local markets to estimate sales potential for the following 12, or even 24 months. These estimates capture both planned purchases by major OEMs, as well as possible orders by distributors, resellers, dealers, and retailers. Obviously, some of these estimates are more reliable than others and the accuracy of these forecasts degrades sharply beyond the immediate quarter so that significant uncertainty in total demand remains. The demand forecast represents the combined estimates of the monthly, worldwide demand for each individual product. Given that the two products are imperfect substitutes, their aggregate demand is much better known than the specific mix. Indeed, there is significant uncertainty regarding the adoption of a particular product. Conceptually, the demand forecast captures demand uncertainty by a probability distribution. The capacity-investment problem is to decide on the size and timing of changes in the capacity of three resources (FA1, FA2, and Test) given a joint probability distribution of quarterly demand for the two product families. The capacity portfolio is denoted by the vector $\mathbf{K}$, where K_1 is the capacity of FA1, K_2 of FA2, and K_3 of Test. □

In practice, dealing with uncertainty in a consistent manner throughout a global corporation is a nontrivial task. Managers know the academic dictum that "point forecasts" in the form of a single number are typically wrong. Yet, aggregating demand forecasts that feature not only means but, at a minimum, also variances and some measure of covariances is not easy. Incorporating such uncertainty into the complex procedure of capacity planning is even harder. In addition, demand is not completely exogenous: Sales-force incentives and compensation are typically set to enhance the likelihood of "meeting the numbers." (Section 4.3 will elaborate on the endogeneity of demand uncertainty.) Therefore, it is not uncommon for current commercial capacity-planning software to consider only a single scenario in the forecast. This is sometimes called the "sales plan" and is, typically, the input to aggregate planning, MRP, and CRP systems. While such approach, which we will call *deterministic* or *sales-plan driven capacity planning*, virtually ignores uncertainty, it "works," even under decentralized decision making. In other words, it is easily explained and understood, and provides a first-order estimate of capacity requirements. We shall see that it typically leads to a balanced capacity configuration, which is attractive from a cost-perspective, as it allows for the possibility to fully utilize all resources simultaneously, resulting in nice accounting efficiency metrics. Another capacity plan that may show up in practice is a plan that minimizes lost sales. In some settings, marketing managers may state that "a customer lost once is lost forever" and advocate ample capacity to prevent that. We refer to such a plan that is attractive from a revenue-perspective as a "total coverage" capacity plan.

Incorporating uncertainty, however, typically changes the capacity investment and can improve performance and mitigate incentive conflicts, which is what this paper aims to illustrate. For example, as observed in Harrison and Van Mieghem (1999), the optimal investment strategy in a stochastic model typically involves some degree of capacity imbalance, which can never be optimal in the deterministic version of the model. Also, it suggests an expected-profit-maximizing compromise between the two conflicting incentives of cost-efficiency of production and revenue maximization of sales. Finally, a stochastic capacity model can capture the risk aversion of decision makers and can show how capacity can be used to mitigate risk and improve performance. These conceptual distinctions between deterministic and stochastic capacity investment, as well as their ramifications for capacity planning practice, will be discussed in detail in the context of the example in §§4.2 and 6.3.

Finally, a caveat: The models in this paper are intended to assist decision making, but seasoned researchers know that practice is more complex. Before a firm can consider capacity decisions, it must articulate its business strategy, decide on its competitive positioning, and on which markets to enter or exit, etc. Here, we restrict attention to mathematical models that provide insights into the nature and financial value of smart investments and that complement other political and strategic considerations that influence capacity decisions.

2. Capacity Research and Related Literatures

This section surveys several major fields that study capacity and explains their commonalities and differences in emphasis. Specific references will be discussed throughout this paper.

The *capacity expansion* literature in operations research is concerned with determining the size, timing, and location of buying additional capacity, according to Luss (1982), which is the latest comprehensive survey to our knowledge. In the beginning, the field's basic concern was how to meet the growing demand in growing economies. The seminal paper by Manne (1961) studies the fundamental trade-off between the economies-of-scale savings of large expansion sizes versus the opportunity cost of installing capacity before it is needed. Often, this literature is deterministic and focuses on minimizing the (discounted) cost of all expansions. Some authors allow for uncertainty: Examples of "post-Luss (1982)," which is the focus of this paper, include Davis et al. (1987), Paraskevopoulos et al. (1991), and Bean et al. (1992). This stream of research is very much related to our topic, but is typically restricted to the expansion of capacity of one resource and cost minimization.

While the capacity-expansion literature assumes that capacity is infinitely durable and is never replaced (let alone, reduced), the *equipment replacement* literature focuses on replacement, while typically ignoring demand changes or scale economies. Rajagopalan (1998) gives a recent review of that literature and presents a unified approach to capacity expansion and equipment replacement in a deterministic setting.

It appears that over the course of the last twenty years, the study of *plant location* has focused on transportation issues and somewhat divorced itself from the study of type, size, and timing of capacity investment.

The *technology management, new product development, and operations strategy* literature deals with deciding on the choice of technology, among many other things. For example, should we invest in flexible or specialized technology? When to switch to a new technology? When should new product designs share common components? How to allocate investment among a set of new product projects? When technology is defined in terms of the capabilities of a network of different resources, then such questions can be addressed with stochastic capacity portfolio investment models. Indeed, a multiresource framework can be used to select the technology (defined by resources with optimal positive capacity levels), along with its capacity plan. For example, capacity papers that study investment in flexible technology include Fine and Freund (1990), He and Pindyck (1992), Jordan and Graves (1995), Van Mieghem (1998a), Netessine et al. (2002), and Bish and Wang (2002). Capacity papers that study timing of new technology adoption include Li and Tirupati (1994) and Rajagopalan et al. (1998). Component commonality is studied with a capacity portfolio investment model in Van Mieghem (2003a).

The *production* or *aggregate planning* literature studies the problem of the "acquisition and allocation of limited resources to production activities so as to satisfy customer demand over a specified time horizon" (Graves 2002, p. 726). The answer typically is derived via an optimization problem, often a linear programming model, and yields a mixed strategy of "chase demand" by having excess capacity or time flexibility, and "level production" by having inventories. Clearly, aggregate planning is concerned with the determination of the level of processing resources over time, but there are two significant differences with stochastic capacity investment. First, the resources under consideration are often restricted to workforce size, inventory planning, subcontracting, and overtime scheduling. Second, virtually all aggregate planning considers a deterministic future, very similar to the deterministic planning described in the Introduction.

While nothing precludes the inclusion of capital equipment adjustments in aggregate planning models, the planning horizon typically is short-to-medium term, such that capital equipment is fixed, but its utilization and allocation to products over time is variable. As such, aggregate planning takes a first-order approach to endogenizing the fundamental trade-offs in production planning among capacity utilization (regular time, overtime, subcontracted), inventory, and service and responsiveness (e.g., backlogging or lost sales). Exceptions that consider capacity expansion and inventory management jointly in

an aggregate planning model include Bradley and Arntzen (1999), Atamtürk and Hochbaum (2001), and Rajagopalan and Swaminathan (2001). Bradley and Arntzen (1999) present a mixed-integer program to maximize return on assets and apply it to two firms to illustrate the capacity-inventory trade-off. Atamtürk and Hochbaum (2001) investigate the trade-offs between acquiring capacity, subcontracting, production, and holding inventory to satisfy nonstationary deterministic demand for a single period over a finite horizon. Rajagopalan and Swaminathan (2001) consider a multiproduct environment where demand for items is known and growing gradually while capacity additions are discrete. Therefore, periods immediately following a capacity increase are characterized by excess capacity. Their model studies the following trade-off: Should excess capacity be used to do more equipment changeovers, and thus, reduce inventories, or should more inventory be built in order to delay future capacity expansions?

The input to aggregate planning includes a deterministic demand forecast for each period in the planning horizon. To consider the impact of uncertainty, aggregate planning typically makes two suggestions: Perform sensitivity analysis on the inputs of the aggregate plan and use safety inventory or safety capacity to satisfy demand higher than forecasted. While such approaches may work well in some environments, it may be desirable to incorporate the effects of uncertainty directly in the model to ascertain how the optimal capacity plan is affected. Stochastic capacity investment turns up the brightness on the direct effect of uncertainty and turns it down on some tactical activities. As such, it will automatically and optimally (as opposed to manually and heuristically) incorporate sensitivity analysis of input uncertainty and endogenously define safety factors.

From a hierarchical[5] perspective, aggregate planning operates at a lower level, and with shorter planning horizons, than *resource planning*, which is closely related to stochastic capacity-portfolio investment. Resource planning often neglects changes in inventory and overtime scheduling. Such a view is similar to the view in economics that "current output flow depends on installed capital stock, and perhaps on flows of instantaneously variable inputs like labor and raw materials, through a production function . . . . We can regard profit flow as the outcome of an instantaneous optimization problem where variable inputs such as labor or raw materials are chosen holding the level of capital fixed" (Dixit and Pindyck 1994, pp. 357–359). CRP often shows up in the context of production planning and planning software like MRP and MRPII (Hopp and Spearman 1996, Nahmias 1993). Despite its name, however, CRP typically verifies whether the MRP-generated production plans are feasible given the capacity in place; that is, it checks feasibility of resource *allocation* rather than plan resource investment or adjustments.

The *inventory* and *supply chain management* literature is concerned with the flow of material through a multiechelon inventory system. In contrast to the aggregate-planning literature, this field often explicitly considers uncertainty, but rarely capacity. (In single-period models, however, there is no essential difference between capacity or inventory, as shall be illustrated in §4.) Kapuscinski and Tayur (1998) provide a recent review of articles that consider capacitated supply chains. Most of those deal with given, fixed capacities. Capacity, here, is the upper bound on production quantities, which is typically deterministic. An exception is Hu et al. (2002), who consider stochastic upper-bounds, i.e., capacity uncertainty, in addition to demand uncertainty. Inventory papers that also consider capacity investment decisions include Angelus and Porteus (2002), Bradley and Glynn (2002), Van Mieghem and Rudi (2002), and will be discussed in the remaining sections. Networks where multiple agents control their own productive capacity require game-theoretic models as §4.3 will discuss.

Investment in capital and labor has been at the core of the *economics* literature since its inception. Investment contributes to future output, economic growth, current demand, and employment. Given that investment empirically is lumpy (versus continuous over time), linear theories had to be expanded. When it is costly to reverse investment in capital or labor, a firm's investment decisions exhibit the empirically

[5] Cf. Sethi et al. (2002), for a survey on hierarchical control, including applications on capacity expansion and equipment replacement problems.

observed dynamics. This occurs, for example, when a firm faces labor firing costs or when it cannot recoup the acquisition price of capital when it is resold. While the possibility that investment may be costly to reverse has been recognized in the literature at least as far back as Arrow (1968), irreversible investment started receiving more attention in the late 1980s and early 1990s in a stochastic framework. It exploits an analogy with the theory of options in financial markets and, therefore, came to be known as the "real options approach to investment," according to pioneers Dixit and Pindyck (1994). The resources that we consider are real assets and our discussion may very well be labeled as "a real options approach." Section 5 will review why capacity is typically not adjusted continuously over time.

The specialization of economics called *corporate finance* considers methods to value and finance investments in projects and property, plant, and equipment. The traditional valuation methods for projects include net present value, payback years, and various types of hurdles on financial ratios. A central question in this field is: What is the objective of the firm, and how do we value uncertain future cash flows? Clearly, this relates to risk sensitivity, which will be summarized in §6.1. Corporate finance also focuses on the means of financing the investment and the impact of debt versus equity on the capital structure of the firm, which is beyond the scope of this paper. As we shall discuss later, stochastic capacity models assume (often implicitly) either perfect capital markets, so that frictionless borrowing is possible, or that the investment size is relatively small, so that it can be internally financed without material impact on the overall valuation of the firm.

3. General Issues in Capacity Investment Problems

This section discusses important issues in typical capacity investment problems. In the course of that discussion, some definitions and notations will be introduced. Consider a firm that has n different "means of processing," which we will call *resources*. Its *capacity portfolio* at time t is denoted by the nonnegative capacity vector $\mathbf{K}_t \in \mathbb{R}^n_+$ whose ith component represents the level of resource i that is available for processing at time t. The capacity problem is to characterize the desired capacity portfolio over time.

3.1. Capacity Constraint Formulation

General higher-level models, often used in economics, capture the impact of capacity by a direct functional dependence of operating profits on the capacity stock. The operating-profit functions $\pi_t(\mathbf{K}_t, \omega)$ denote the operating profit (excluding capacity investment costs) as a function of time t, the capacity vector $\mathbf{K}_t$ available at that time, and uncertainty represented by ω, which refers to "state of the world," or in more general models, a sample path. (Incorporating ω emphasizes that the entity is random.) The typical assumption is that operating profits are concave in the capacity vector $\mathbf{K}$, which captures decreasing marginal returns from investment. This higher-level formulation is remarkably general and flexible. No further specific assumptions on exactly how capacity constrains processing quantities are needed to characterize general capacity dynamics, as will be reviewed in §5. In general, $\mathbf{K}$ need not even be interpretable in terms of maximal product quantities. Stochastic dependence captures broad capacity formulations, including "random capacity," where profit is a stochastic function of capacity.

In operations management, one often extends and details the formulation so that the operating-profit function $\pi_t(\mathbf{K}_t, \omega)$ becomes endogenous to the model. Typically, the model explicitly specifies the input-output relationship and its dependence on capacity, process structure, and management. The operating-profit functions $\pi_t(\mathbf{K}_t, \omega)$ then become the outcome of tactical optimization problems that depend on the state of the process and its capacity vector. For example, this formulation could capture a complicated setting with product-sequence-dependent setup times. The outputs and operating profit during a period, then, depend not only on the available capacity vector $\mathbf{K}$, but also on the ex-post optimal production schedule, which can be a function of the actual output demand and input supply during that period. This captures a setting where maximal output quantities of various products are state-dependent and nonseparable among products.

A simpler and fairly typical capacity constraint formulation is via a "recourse" linear-programming problem, as illustrated by the following example: A firm sells m products in a competitive market where prices are uncertain. Let $\mathbf{p}_t$ represent the unit price vector for period t, which is observed at the beginning of the period *before* $\mathbf{K}_t$ is chosen. According to the philosophy of continuous improvement, the firm is improving its manufacturing technologies, but not in a deterministic fashion. The firm's capacity consumption matrix $\mathbf{A}_t$ and marginal processing costs $\mathbf{c}_t$ for period t are also observed at the beginning of the period, before $\mathbf{K}_t$ is chosen. Assuming that the firm's processing quantities, $\mathbf{x}_t$, during that period are linearly constrained, the firm will set period t processing according to the linear program, to maximize the operating profit:

$$\pi_t(\mathbf{K}_t, \omega) = \max_{\mathbf{x}_t \in \mathbb{R}_+^m} \quad (\mathbf{p}_t(\omega) - \mathbf{c}_t(\omega))' \mathbf{x}_t \tag{1}$$

$$\text{s.t.} \quad \mathbf{A}_t(\omega)\mathbf{x}_t \le \mathbf{K}_t, \tag{2}$$

where $'$ denotes transpose and vector inequalities should be interpreted componentwise. The resulting operating-profit function $\pi_t(\cdot, \omega)$ is concave for each ω and t.

The operational consequence is that capacities in operations-management models (and in this paper) can almost always be interpreted as some upper bounds on processing rates. Often, the capacity vector is the right-hand side of a linear constraint on processing quantities during a period (i.e., processing rates) similar to (2). Clearly, appropriately defining or adding a "period" corresponds to determining or increasing the total number of hours (or shifts) worked, which can modify the total capacity over a given horizon. Newsvendor networks, which will be discussed in the next section, fall into this formulation where the impact of capacity is modeled explicitly via "hard," linear constraints. Another typical capacity constraint formulation is via queuing model, which highlights the uncertainty of processing and the impact of tactical scheduling and resource starvation on realized capacity, as will be discussed in §4.1.

In reality, capacity constraints may be "soft," in the sense that the output of resource i is not rigidly bounded, but can be increased, albeit at an increasing cost. The increasing marginal costs may reflect extraordinary charges including expediting, overtime, etc. Such capacity constraints can be captured implicitly by a concave operating-profit function, or explicitly by processing costs that are piecewise linear or convex increasing when quantity exceeds a critical number, say K_i.

3.2. Capacity Adjustment Costs

Capacity adjustment cost are the investment costs incurred when changing capacity. When adjusting capacity vector $\mathbf{K}_{t-1}$ to $\mathbf{K}_t$ at time t, the associated cost is, in general, a bivariate function of $(\mathbf{K}_{t-1}, \mathbf{K}_t)$. In addition, when evaluated at an earlier time, it may be uncertain. Typically, however, the adjustment cost at time t is assumed to only depend on the capacity change and is denoted by $C_t(\mathbf{K}_t - \mathbf{K}_{t-1})$. In addition, C_t is assumed to be convex to guarantee a well-behaved concave capacity investment optimization problem. Let $\mathbf{x}^+$ and $\mathbf{x}^-$ denote the vectors with components $\max(0, x_i)$ and $\max(0, -x_i)$, respectively. The typical economic assumption is that C_t is a kinked piecewise linear convex function C_t:

$$C_t(\mathbf{x}) = \mathbf{c}'_{K,t}\mathbf{x}^+ - \mathbf{r}'_{K,t}\mathbf{x}^-, \tag{3}$$

where the marginal investment costs $\mathbf{c}_{K,t}$ and disinvestment revenues $\mathbf{r}_{K,t}$ are usually, but not necessarily, positive. Very seldom can capital investment be reversed at no cost; that unique setting where $\mathbf{c}_{K,t} = \mathbf{r}_{K,t}$ is called *reversible* or *frictionless investment*. Typically, the focus is on resources that are costly to reverse,[6] meaning that $\mathbf{c}_{K,t} > \mathbf{r}_{K,t}$, so that only a fraction of the investment cost is recovered when selling real assets. If $\mathbf{r}_{K,t} = 0$, nothing is recovered, which is called *irreversible investment*.

The nature of the adjustment cost function depends on the type of adjustment that is considered and, hence, the convexity assumption of adjustment costs is not always appropriate. It may be appropriate when capacity is divisible or when considering

[6] Optimal capacities of costlessly reversible resources are found by optimizing the (extended) operating-profit function. Those resources are assumed to have been "maximized out" and, therefore, don't appear in the operating-profit function.

investment at the macro-economic level.[7] Also, ongoing or gradual adjustments, such as maintenance, training, etc., might be labeled "frictionless" investments, and hence, modeled by a linear adjustment cost. In contrast, however, infrequent and major capacity adjustments typically stem from some friction or nonlinearity in the adjustment costs and often enjoy economies of scale, implying that $C_t(\cdot)$ is concave. For example, at the firm level, there is often a fixed cost associated with making any capacity adjustment, creating a discontinuity in C_t at 0: $C_t(0) = 0$, while $\lim_{\|\mathbf{K}_t - \mathbf{K}_{t-1}\| \to 0} C_t(\mathbf{K}_t - \mathbf{K}_{t-1}) > 0$. Luss (1982) gives two often-used adjustment cost functions that exhibit economies of scale. The first one, called the "fixed charge" cost function, is the affine version of (3), but with a discontinuity at 0. Its single-resource version for capacity additions is $C(x) = c_0 + c_1 x$ for $x > 0$ with $C(0) = 0$. Such affine capacity adjustment costs can be handled fairly well as discussed in Van Mieghem and Rudi (2002); a pure fixed cost only affects the boundary invest-or-not decisions, while the size of the adjustment remains given by interior optimality conditions that are independent of the fixed cost component. The second often-used adjustment cost function with economies of scale is the "power" cost function, which is strictly concave. Its single resource version is: $C(x) = kx^\alpha$, where $k > 0$ and $0 < \alpha < 1$. Frictions from irreversibility and/or economies of scale in adjustment cost lead to a firm's optimal capacity dynamics involving occasional large changes, as will be discussed in §5.

3.3. Capacity Investment Objective, Including Planning Horizon, Discounting, and Decision Makers

Traditionally, the objective is to maximize expected net present value of the firm. (Section 6 will discuss more recent work that moves beyond maximizing expected present values and incorporates risk aversion.) Net present value calculations require a planning horizon T and discounting, which typically assumes a constant per-period discount factor $\delta > 0$. Obviously, the values of the planning horizon T and discount factor δ influence dynamic capacity decisions. Capacity models typically use longer planning horizons (several years and sometimes decades) than aggregate planning models (several months and sometimes years). This longer planning horizon reduces forecasting accuracy (see below) and increases uncertainty, making stochastic models more desirable. On the other hand, higher discount factors mitigate the impact of longer horizons. (The appropriate choice of δ is a central problem in finance that goes beyond the scope of this paper.) Discounting also puts restrictions on the adjustment cost functions. For example, one typically assumes that the present value of a unit of used capacity cannot be higher than a new unit, i.e., $\mathbf{c}_{K,t} \geq \delta^{\tau - t}\mathbf{r}_{K,\tau}$ for $\tau > t$, to exclude unrealistic capacity dynamics. In addition, finite horizon models ($T < \infty$) require an additional element in their formulation: A salvage function $f(\mathbf{K}, \omega)$, which is the final (salvage) value for capacity portfolio $\mathbf{K}$ given that state ω obtains.

The traditional objective of capacity-investment problems is either processing-network optimization or network design by a single decision maker. Deciding on the amounts of capacity involves optimization of a given processing network or operating-profit function. By necessity, all single-resource capacity models fall into this class. By considering a portfolio of different types of resources, however, the capacity problem can amount to network design, or selecting the most appropriate mix of types of resources and their configuration (cf. the discussion on technology management in §2). Often, however, capacity decisions are strategic and may depend on multiple decision makers, including other firm's decisions, as §4.3 will review.

3.4. Continuous Versus Discrete Capacity

Many models assume that the capacity vector is a nonnegative real variable, so that capacity is divisible and, thus, its investment can be continuous or "incremental." While this is a valid assumption in settings where the number of possible capacity sizes is large, more detailed and precise capacity-investment models may treat capacity as a discrete variable. Such "lumpy" investment is appropriate when capacity is indivisible and can only be installed in a small

[7] In a macro-economic equilibrium setting, the marginal cost of investment should eventually increase, because the shadow price of investment should eventually rise as resources are diverted from other uses, which tends to convexify the cost of investing.

number of possible sizes. Integer restrictions typically make these models less amenable to analysis.

3.5. Leadtimes

Leadtimes refer to the time between the purchase and availability of new capacity. With growing stochastic demand, capacity leadtimes increase the risk of capacity shortage caused by demand uncertainty. The basic approaches to protect against such risk are to either expand capacity earlier or to adopt larger capacity increments. Only a handful of papers, reviewed in Ryan (2002), consider leadtimes in a stochastic setting. Erlenkotter et al. (1989) consider one capacity expansion of a single resource where expansion completion timing is held constant but the leadtime is a decision variable. They show that the presence of uncertainty has the effect of reducing the optimal leadtime compared to its optimal value in a deterministic model. When demand follows a geometric Brownian motion, Ryan (2002) shows that leadtimes influence timing, cost parameters determine expansion size, and demand characteristics affect both timing and size.

3.6. Physical Depreciation and Degradation

Physical depreciation and degradation of capacity refers to the diminishing financial and operating value of a resource over time. Depreciation is typically modeled by a financial loss to the firm's value. (The depreciation on the marginal capacity unit of resource i is proportional to its marginal value and is typically captured by increasing the discount rate.) Similarly, physical degradation can be modeled by a decrease in the capacity vector, typically proportional to the installed capacity $\mathbf{K}$, or as in equipment replacement models as reviewed by Rajagopalan (1998).

3.7. Tactical Activities, Starvation, and Inventory

General capacity-investment models take operating-profit functions $\pi_t(\mathbf{K}, \omega)$ as primitives. While the observed dynamics of inventories and other tactical flows could be incorporated in state- or sample-path ω, many capacity models simply ignore tactical flows, as discussed in §2.

Clearly, modeling must strike a balance between complexity and realism: The appropriateness of ignoring tactical flows may depend on the time-scale and planning horizon under study. For example, in some settings, capacity adjustments are only allowed infrequently. When adjustments of tactical flows occur on a much smaller time-scale than capacity adjustments, tactical flows can be optimized, given the capital stock that only varies on a larger time-scale and thus can be taken as fixed for the tactical flow optimization. Such "time-scale separation" would make inventory changes, scheduling, and other tactical decisions "invisible" at the higher level of capacity planning. The justification in Eppen et al. (1989) for ignoring inventory-carryover between periods can be interpreted as a time-scale separation argument: "[ignoring inventory-carryover is] consistent with the fact that each time period is of sufficient length (one year) so that production levels can be altered within the time period in order to satisfy as closely as possible the demand that is actually experienced." In many settings, however, actual realized output quantities often depend on tactical activities and flows, such as product-sequence-dependent setups and scheduling that determines availability of raw material and work-in-progress or resource starvation. For example, in seasonal environments, it is not unusual to set capacity for constant processing and buffer seasonalities with inventory build-up and depletion. To capture these detailed operational effects and trade-offs, it is often necessary to incorporate tactical activities and inventory carryover, especially in capacity portfolio models. When activities or products are not divisible, scheduling conflicts may lead to resource blocking and starving, which may reduce actual process capacity below individual bottleneck resource capacities. (See §4.1.) Interactivity inventory buffers those problems.

Mathematically, capacity investment is similar to inventory management in that it deals with stochastic optimization of very similar functions. In single-period models, there is no essential difference between inventory and capacity. There are some differences in multiperiod (i.e., dynamic) models. First, capacity is "utilized," but not "consumed:" Unlike inventory, the capacity vector is not depleted by demand. It may, however, degrade faster with higher usages, requiring faster depreciation. In inventory models, different periods are linked via inventory-carryover, which is often neglected in capacity models. On the other hand, inventory models mostly

focus on "moving the goods" from stage to stage, whereas capacity portfolio models focus more on processing and the resulting interproduct couplings. Second, inventory models typically assume that unused inventory and all capacity are not perishable and that they are available for use in the next period. Third, capacity models typically have longer time horizons than inventory models so that discounting and forecasting gain in importance in capacity models. Last, and least, stationary capacity models typically collapse to an essentially static capacity problem: Given that no new information is ever gained, the optimal capacity investment policy typically makes only one initial investment ever (see next section). Stationary inventory models, on the other hand, retain dynamic reordering, while the ordering policy itself also is described by one or two critical numbers which also can be found in an essentially static problem. Dynamic capacity models, where timing becomes nontrivial, require a more sophisticated stochastic formulation: Typically the problem is nonstationary, or capacity degrades over time.

3.8. Unsatisfied Demand and Capacity Shortages

When demand is uncertain, capacity is costly, or capacity adjustments are not instantaneous, there will be instances of insufficient capacity to meet demand. Depending on the view one takes, these are called capacity shortages or excess demand. In practice, firms employ tactical countermeasures, such as allocation schemes, increased pricing, backlogging, or advance inventory build-up, to manage shortages. When such tactical countermeasures are not incorporated in the model, assumptions must be made on what happens with excess demand. There is a strand in the capacity literature that "assumes that available capacity must meet or exceed demand" (Bean et al. 1992, p. S210), which is also called the "no backlogs in demand" assumption by Manne (1961). Obviously, with uncertain demand, a no-capacity-shortage assumption must be accompanied by a zero-capacity-leadtime assumption. The alternative is to allow for excess demand, which either is backlogged, or lost, or some combination of both. In either case, a demand-shortage penalty is typically included, which may represent the loss of goodwill that will manifest itself in a reduction of future demand and is very hard to quantify, or the cost to fill the demand through an alternate process, as discussed in Manne (1961). The lost-sales case is relatively easily incorporated in single-stage processes, while backlogging requires one to incorporate negative inventory carry-over. In a capacity-expansion setting where no contraction is allowed, backlogging can easily be incorporated, as shown in Manne (1961), for a single resource and growing demand, and in Van Mieghem and Rudi (2002) for a capacity portfolio and stationary demand. The "no backlogging" assumption is typically easier to analyze because it corresponds to the limiting case where the shortage penalty becomes infinite and the peak process, $\sup_{\tau \le t} \mathbf{D}_\tau$, contains all relevant information. In multiproduct systems, as studied in capacity-portfolio problems, an alternative option to backlogging or lost sales is that unsatisfied demand for one product "spills over" to another product. In other words, ex-post substitution provides another mechanism for matching capacity with demand. Substitution has started to be explored in an inventory context, and its effects are likely to be similar in capacity problems. An important issue is whether the substitution is executed by the firm or by customers, as discussed in Bassok et al. (1999) and Netessine and Rudi (2003).

3.9. Uncertainty, Information Structure, Learning, and Forecasting

Arguably, the most important factor in a dynamic capacity-investment model is the description of uncertainty and its resolution over time. For the theorist, the tractability of the model is directly related to the mathematical properties of the stochastic process that represents uncertainty and information resolution, as will be illustrated in §5. In theory, that stochastic process should depend on the employed forecasting procedures. While forecasting is extremely relevant to the practitioner, however, it is rarely discussed in capacity-investment research, Ryan (2003) being a notable exception. (Chand et al. 2002, review the literature on forecast and rolling horizons.) Paraskevopoulos et al. (1991) distinguish three sources of uncertainty: The uncertainty ("forecast error") in econometrically estimated demand equations and exogenous assumptions, uncertainty in structural shifts in demand over time, and uncertainty in system parameters, such as

the rate of learning. (See Hiller and Shapiro 1986, for a capacity-investment model that incorporates learning effects.) Often one can "learn" and update the demand forecast as information is revealed and observed over time. In such setting, Burnetas and Gilbert (2001) analyze the trade-off between better demand information by waiting to procure capacity versus the increased capacity cost when capacity acquisition cost increases over time.

General capacity models assume an operating-profit function and a stochastic environment that does not specify the origin of uncertainty. They can capture uncertainty in supply, internal processing, demand, costs, prices, and environmental factors. In operations-management models, demand is often the generator of uncertainty and typically is assumed to be exogenous. The other extreme would be to assume that demand is endogenous, reflecting practices where marketing and sales would agree on a sales plan and then each function would be responsible for making it happen. Manufacturing produces according to the plan and marketing and sales do whatever is necessary to realize sales according to the plan. With an endogeneity assumption, deterministic or sales-plan driven capacity planning, as described in the Introduction, may be appropriate to assess how to deploy capacity to meet the sales plan, as well as the cost of doing so. In that case, however, the capacity problem has just been translated into the problem of determining the best sales plan. Of course, reality is somewhere in between: Demand is neither completely exogenous nor completely endogenous; a property that has not received much attention in the operations literature. A notable exception is Cachon and Lariviere (1999) where demand is influenced by the scarcity of capacity via capacity allocation schemes. Then, demand can be decreasing in capacity if the firm's customers become convinced that their requirements will surely be met.

4. Optimal Capacity Investment: Types and Amounts

This section reviews optimal capacity levels for various types under risk-neutral investment. (Timing and risk aversion will be discussed in §§5 and 6, respectively.) Here, we consider a stationary environment where it is optimal to keep capacity constant over time. The simplest such environment is an i.i.d. structure, which Eberly and Van Mieghem (1997) define as capacity portfolio models with (1) stationary operating profit and kinked, piecewise linear adjustment cost functions (i.e., $\pi_t = \pi$, $\mathbf{r}_t = \mathbf{r}$, $\mathbf{c}_t = \mathbf{c}$), (2) stationary stochastic structure (i.e., probability measures for ω_t and ω_1, where $\omega = \Pi_t \omega_t$, are identical), and (3) independent periods (i.e., probability measures for any pair, $\omega_i \omega_j$, is equal to the product of their measures). They show that an i.i.d. setting in infinite horizon[8] reduces the general dynamic capacity problem to a single, initial capacity investment, effectively collapsing the problem to a single-period problem. While losing dynamics, these essentially static capacity-investment models are able to capture rich modeling detail in the processing network and the nature of uncertainty.

Section 4.1 reviews canonical stochastic capacity models that adopt either queuing or newsvendor network formulations. Section 4.2 illustrates newsvendor network analysis in the context of the example in the Introduction. Section 4.3 reviews game-theoretic capacity investment by multiple agents, who each control a subset of the processing network.

4.1. Canonical Capacity Models in an i.i.d. Setting with Multivariate Uncertainty: Queuing Versus Newsvendor Models

Queuing Models. As discussed above, operations-management models often use detailed formulations where the operating-profit function $\pi(\mathbf{K}, \omega)$ becomes endogenous. One formulation involves queuing models, which are typically set in continuous-time and focus on flow times and responsiveness in the presence of stochastic processing and stochastic demand. Multiple resources lead to queuing networks in the obvious way and product-dependent scheduling, processing, and routing leads to multiclass queuing networks. Such systems allow for multivariate (often product-dependent) uncertainty, although different classes typically are assumed to be independent

[8] That property retains in finite horizon settings *if* the final value function f is identical to the disinvestment cost: $f(\mathbf{K}, \omega) = \mathbf{r}'\mathbf{K}$. Otherwise, end-of-horizon investment effects may appear.

(queuing theory is not very successful in handling correlated arrivals). Queuing models can study tactical activities, such as dynamic sequencing and routing in a given network, but are vastly underutilized in the study of capacity investment (perhaps because of their increased complexity and tactical detail).

In queuing models, the "maximal average processing rate" at a node (location) in the processing network represents the capacity of that node. In the simplest setting, a node contains a single resource, or "server," that processes a single product, which may represent a good, a customer, or both. Traditionally, the maximal average processing rate at resource i is denoted by μ_i, which is the reciprocal of the average time m_i to process the product at that resource. Capacity then constrains the average processing rate at resource i, which is typically denoted by λ_i, as $\lambda_i \leq \mu_i$. This is often scaled into an equivalent capacity utilization constraint, which is traditionally denoted by $\rho_i = m_i\lambda_i \leq 1$. The latter representation directly allows extension to a multiproduct setting, or a multiclass queuing network. Letting m_{ij} and λ_{ij} denote the average processing time and rate of product j at resource i, the capacity constraint for resource i becomes $\sum_j m_{ij}\lambda_{ij} \leq 1$.

When comparing this to the linear programming constraint (2), processing times m and processing rates λ are the obvious counterparts of capacity consumption rates A and activity rates x, but what about the level K_i of investment of resource i? There are several interpretations of capacity investment in queuing. Capacity investment can refer to increasing the service rate μ_i of the server at node i, typically by reducing the mean processing time of a particular product, or of all products, served. Capacity investment can also refer to increasing the number of servers at node i. While capacity can be changed continuously by increasing processing speed, it can only assume discrete values when increasing number of processors in the latter case. Obviously, both approaches can be used simultaneously to adjust capacity. Multiserver nodes quickly become difficult to handle analytically. In some formulation that adopt, for example, fluid or Brownian approximations, one can introduce an additional "capacity-scaling factor" for processing node i, say μ_i^*. The multiproduct constraint, $\sum_j m_{ij}\lambda_{ij} \leq 1$, then becomes $\sum_j m_{ij}\lambda_{ij} \leq \mu_i^*$, which is the counterpart of the linear-programming constraint $Ax \leq K$. Queuing models, however, typically offer a more detailed representation of processing networks than linear programming models in that they incorporate the impact of tactical scheduling conflicts on total network capacity. As discussed earlier, when activities or products are not divisible, processing cannot be interrupted or preempted. Resulting "bang-bang" processing and poor scheduling may then lead to downstream resource starving, which may reduce actual process capacity below individual bottleneck resource capacities. For example, in multiclass queuing networks, naive scheduling rules can induce product-specific starvation at different servers at different times. This reduces the aggregate network's effective capacity in the sense that the network can become instable, even though each server is less than 100% utilized. (Dai and Vande Vate 2000 review the recent surge in the study of stability conditions for multiclass queueing networks.)

Queuing models become capacity-investment models when superposed with optimization and a capacity adjustment cost function. Some examples of queuing capacity models include Mendelson (1985), Loch (1991), Lederer and Li (1997), and Cachon and Harker (2002). The continental divide between inventory and queuing models also applies to the subfield of capacity investment. Production-inventory models attempt to bridge the two worlds with representative examples (Caldentey and Wein 2003, Armony and Plambeck 2002). (Section 4.3 reviews these multiagent queuing models.) Boyaci and Ray (2003) study a firm with two substitutable products that differ only in their prices and delivery times and are produced by dedicated capacities. They analyze how capacity costs impact capacity-investment levels and market positioning of the two products.

Newsvendor Models. Besides queuing formulations, another, more popular capacity-investment formulation is via a "recourse" linear-programming problem. The newsvendor network formulation in Van Mieghem and Rudi (2002) is an example of such i.i.d. recourse-based models. Newsvendor network problems are typically set in discrete-time and focus on the impact of multivariate demand uncertainty,

while assuming deterministic processing. Often a sequence of i.i.d. demand vectors $\{\mathbf{D}_t\colon t \in \mathbb{N}\}$ is the generator of uncertainty in these models, although supply, yields, and other uncertainty also can be captured. At the beginning of the period (called Stage 1), the capacity vector $\mathbf{K}$ and inventory vector $\mathbf{S}$ are chosen. Then demand $\mathbf{D}$ is revealed. At the end of the period (Stage 2), an activity vector $\mathbf{x}$ is chosen to maximize operating profit by transforming $\mathbf{R}_S\mathbf{x}$ of input stock into $\mathbf{R}_D\mathbf{x}$ units of output and sales for given supply and demand routing matrices $\mathbf{R}_S$ and $\mathbf{R}_D$. The kernel of a newsvendor network generalizes (1)–(2) to a network structure (but typically assumes deterministic processing):

$$\pi(\mathbf{K}, \mathbf{S}, \mathbf{D}) = \max_{\mathbf{x}\in\mathbb{R}^m_+} \mathbf{p}'\mathbf{x} - \mathbf{c}'\mathbf{x} - \mathbf{c}'_P(\mathbf{D} - \mathbf{R}_D\mathbf{x}) - \mathbf{c}'_H(\mathbf{S} - \mathbf{R}_S\mathbf{x}) \quad (4)$$

$$\text{s.t.} \quad \mathbf{R}_S\mathbf{x} \le \mathbf{S}, \quad \mathbf{R}_D\mathbf{x} \le \mathbf{D}, \quad \mathbf{A}\mathbf{x} \le \mathbf{K}. \quad (5)$$

In addition to prices $\mathbf{p}$, processing and transportation costs $\mathbf{c}$, routing matrices $\mathbf{R}_S$ and $\mathbf{R}_D$, and capacity consumption matrix $\mathbf{A}$, a newsvendor network's operating-profit function, thus, incorporates unit demand (output) shortage cost $\mathbf{c}_P$, and unit inventory procurement and holding costs $\mathbf{c}_S$ and $\mathbf{c}_H$. The objective is to maximize the expected net present firm value, denoted by $\mu(\mathbf{K}, \mathbf{S})$, by choosing capacity $\mathbf{K}$ and inventory $\mathbf{S}$ before demand is known, and choosing activity $\mathbf{x}$ afterwards:

$$\mu(\mathbf{K}, \mathbf{S}) = \delta\mathbb{E}\pi(\mathbf{K}, \mathbf{S}, \mathbf{D}) - \mathbf{c}'_S\mathbf{S} - \mathbf{c}'_K\mathbf{K}, \quad (6)$$

where $\mathbb{E}$ denotes expectation and δ is the per-period, risk-neutral discount factor. The expected value function μ is jointly concave in $\mathbf{K}$ and $\mathbf{S}$ so that the optimization problem is well behaved.

The single-period optimality equations can be expressed in terms of the expected dual prices of the capacity and inventory constraints. As the example in the next section will illustrate, these sufficient first-order conditions can be interpreted as coupled, generalized critical-fractile conditions, that specify the optimal trade-off between the marginal value of capacity, inventory, and their corresponding marginal costs. In other words, newsvendor networks construct the optimal capacity portfolio by trading off the opportunity cost of capacity-underages with the cost of excess-capacity. With lost sales or with backlogging for certain networks, Van Mieghem and Rudi (2002) show that the single-period solution extends to a dynamic i.i.d. structure: The optimal capacity strategy collapses to a single initial capacity investment and all further dynamics involve inventory, but no capacity adjustments. For small networks, the model can be solved analytically; otherwise, it is also easily solved using stochastic optimization via simulation.

By appropriately structuring the three network matrices, newsvendor networks can feature commonality, flexibility, substitution, or transshipment, in addition to assembly and distribution. Multivariate uncertainty allows the study of the important impact of correlation on capacity investment. Van Mieghem and Rudi (2002) show that, if demand is normally distributed, the optimal expected firm value is increasing in the mean demand vector and decreasing in any variance term. In addition, if $\pi(\mathbf{K}, \mathbf{S}, \mathbf{D})$ is submodular in $\mathbf{D}$, then the expected value is decreasing in any covariance term (and, thus, pairwise demand correlation). (There is, however, no general theory yet on the impact of covariances on the optimal capacity portfolio $\mathbf{K}$.) By allowing for alternate or "nonbasic" activities that can redeploy inputs and resources to best respond to resolved uncertain events, newsvendor network analysis can be used for network design. For example, newsvendor networks that study the choice between investing in product-flexible or dedicated capacity include Netessine et al. (2002), Van Mieghem (1998a), and Bish and Wang (2002), who add ex-post pricing to the capacity decisions. Kulkarni et al. (2002) study the choice between product- or process-focused plant network configuration. Newsvendor networks analyzing when it is worthwhile to allow for input commonality or substitution include Van Mieghem (2003a) and others reviewed in Van Mieghem and Rudi (2002). Similar network design questions have been addressed by Graves and coauthors using different, but related models; e.g., Jordan and Graves (1995) present an original and insightful analysis on the impact of "chaining" capacities in a network, and Graves and Willems (2000) study the strategic placement of safety stock in a supply network.

4.2. Solution, Discussion, and Ramifications of the Example

Solution of the Example (Risk-Neutral). To simplify the exposition, consider the example from the introduction in a single-period setting, thereby dispensing with inventory dynamics and discounting (i.e., $\delta = 1$). For concreteness, assume the following data: Demand (in thousands) for the year is estimated for three scenarios and shows high mix uncertainty: a "pessimistic scenario" $\mathbf{D} = \mathbf{D}^1 = (150, 350)$ with probability 1/4, the "expected scenario" $\mathbf{D} = \mathbf{D}^2 = (300, 300)$ with probability 1/2, and an "optimistic scenario" of $\mathbf{D} = \mathbf{D}^3 = (450, 250)$ with probability 1/4. The capacity adjustment cost function $C(\mathbf{K}) = c_{K,0} + \mathbf{c}_K'\mathbf{K}$ includes the fixed cost $c_{K,0}$ of \$40 million and unit-adjustment costs of $c_{K_1} = \$30$ for final-assembly facility 1 (FA1), whereas the more cost-efficient FA2 facility requires a modest $c_{K_2} = \$20$. The testing facility (resource 3), however, is more expensive with $c_{K_3} = \$80$, reflecting the cost of many expensive IDTs. The two high-end drives are estimated to have unit contribution margins $\mathbf{p} - \mathbf{c} = (\$400, \$300)$, respectively,[9] which will be denoted by net value $\mathbf{v}$.

This is a newsvendor network model with two outputs or products, three resources, and four processing activities: activity 1 (2) = FA of Product 1 (2), 3 = test Product 1, 4 = test Product 2. In the static problem, we do not consider inventories. The general activity vector $\mathbf{x}$ is a nonnegative four-vector for any state $\mathbf{D}$. Given the process structure and the assumption that we do not allow for intraresource buffers, we clearly have that $x_1 = x_3$ and $x_2 = x_4$. Thus, a reduced two-vector, which we also will call $\mathbf{x}$, is a sufficient activity descriptor. Without loss of generality, we assume that each product requires approximately the same amount of tester time. Therefore, the relevant network matrices in (4)–(5) are demand routing matrix $\mathbf{R}_D$ and capacity consumption matrix $\mathbf{A}$:

$$\mathbf{R}_D = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

Table 1 The Optimal Activity Vector and Marginal Values of Capacities in Each Demand Domain for the Example

Demand Domain	$\mathbf{x}(\mathbf{K}, \mathbf{D})$	$\boldsymbol{\lambda}(\mathbf{K}, \mathbf{D}) = \nabla_{\mathbf{K}}\pi(\mathbf{K}, \mathbf{D})$
$\Omega_0(\mathbf{K}) = \{y \in \mathbb{R}^2_+ : y_1 \le K_1,\ y_2 \le K_2,\ y_1 + y_2 \le K_3\}$	$\mathbf{D}$	$\boldsymbol{\lambda}^0 = (0, 0, 0)$
$\Omega_1(\mathbf{K}) = \{y \in \mathbb{R}^2_+ : y_1 < K_3 - K_2,\ K_2 < y_2\}$	(D_1, K_2)	$\boldsymbol{\lambda}^1 = (0, v_2, 0)$
$\Omega_2(\mathbf{K}) = \{y \in \mathbb{R}^2_+ : K_3 - K_2 < y_1 < K_1,\ K_3 < y_1 + y_2\}$	$(D_1, K_3 - D_1)$	$\boldsymbol{\lambda}^2 = (0, 0, v_2)$
$\Omega_3(\mathbf{K}) = \{y \in \mathbb{R}^2_+ : K_1 < y_1,\ K_3 - K_1 < y_2\}$	$(K_1, K_3 - K_1)$	$\boldsymbol{\lambda}^3 = (v_1 - v_2, 0, v_2)$
$\Omega_4(\mathbf{K}) = \{y \in \mathbb{R}^2_+ : K_1 < y_1,\ y_2 < K_3 - K_1\}$	(K_1, D_2)	$\boldsymbol{\lambda}^4 = (v_1, 0, 0)$

To determine the newsvendor-network solution $\mathbf{K}^*$, first observe that it is suboptimal to have $K_3 > K_1 + K_2$ or $\max(K_1, K_2) > K_3$. The solution domain for the optimal capacity vector $\mathbf{K}^*$, thus, becomes $\{\mathbf{K} \in \mathbb{R}^3_+ : \max(K_1, K_2) \le K_3 \le K_1 + K_2\}$. Given that $v_1 = \$400 > v_2 = \$300$ and resource consumption rates A_{31} and A_{32} are equal, the optimal contingent activity vector $\mathbf{x}(\mathbf{K}, \mathbf{D})$ is the greedy solution to (4)–(5):

$$\begin{aligned} x_1(\mathbf{K}, \mathbf{D}) &= \min(D_1, K_1, K_3), \\ x_2(\mathbf{K}, \mathbf{D}) &= \min(D_2, K_2, K_3 - x_1) \\ &= \min(D_2, K_2, K_3 - \min(D_1, K_1)). \end{aligned}$$

The demand space can be partitioned into five domains $\Omega_i(\mathbf{K})$, as defined in Table 1 and displayed in Figure 2, in which $\mathbf{x}(\mathbf{K}, \mathbf{D})$ is linear in $\mathbf{K}$, and its optimal shadow value or dual price $\boldsymbol{\lambda}(\mathbf{K}, \mathbf{D}) \overset{\text{def}}{=} \nabla_{\mathbf{K}}\pi(\mathbf{K}, \mathbf{D})$, thus, constant. The first-stage capacity decision maximizes expected value (6) with sufficient condition:[10]

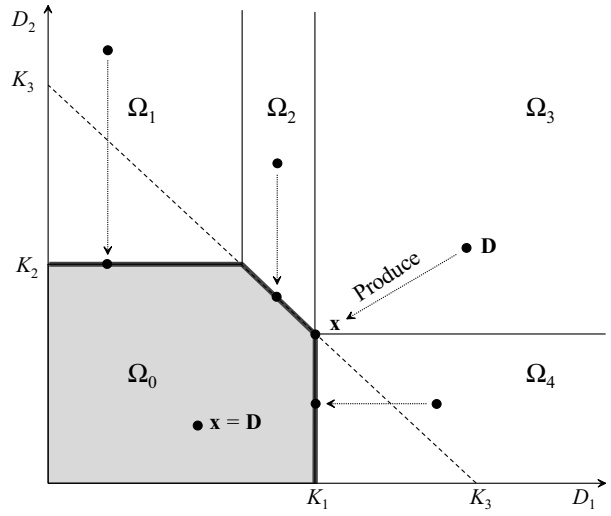
$$\nabla\mu(\mathbf{K}^*) \overset{\text{def}}{=} \nabla\mathbb{E}\pi(\mathbf{K}^*, \mathbf{D}) - \mathbf{c}_K = 0.$$

For newsvendor networks, gradient and expectation interchange: $\nabla\mathbb{E}\pi(\mathbf{K}, \mathbf{D}) = \mathbb{E}\nabla_{\mathbf{K}}\pi(\mathbf{K}, \mathbf{D})$, so that the

[9] Strategic objectives and "competitive intelligence" fixed margins in advance so that uncertainty is manifested mainly through quantities demanded. While our example focuses on demand uncertainty, a complete analysis would also consider uncertainty in margins, or price and variable costs.

[10] With a discrete demand distribution, the expected operating profit is piecewise linear and, hence, not-differentiable at the breakpoints. At those points, $\nabla\mu$ should be interpreted as a subgradient.

Figure 2 **The Optimal Ex-Post Activity Vector x for the Example Depends on the Capacity K and Demand D**



marginal operating-profit is the expected shadow vector $\boldsymbol{\Lambda}(\mathbf{K})$ of the linear program:

$$\boldsymbol{\Lambda}(\mathbf{K}) \overset{\text{def}}{=} \mathbb{E}\nabla_{\mathbf{K}}\pi(\mathbf{K}, \mathbf{D}) = \mathbb{E}\boldsymbol{\lambda}(\mathbf{K}, \mathbf{D}) = \sum_{\text{domains } i} \boldsymbol{\lambda}^i P(\Omega_i(\mathbf{K})), \tag{7}$$

where P is the demand forecast and $\boldsymbol{\lambda}^i$ is the constant $\boldsymbol{\lambda}(\mathbf{K}, \mathbf{D})$ for $\mathbf{D} \in \Omega_i(\mathbf{K})$. Thus, the newsvendor-network solution solves:

$$\nabla\mu(\mathbf{K}^*) = \mathbf{0}, \qquad \text{where } \nabla\mu(\mathbf{K}) = \boldsymbol{\Lambda}(\mathbf{K}) - \mathbf{c}_K. \tag{8}$$

Condition (8) for the optimal capacity portfolio is a generalization of the critical-fractile condition in single-dimensional newsvendor models. The marginal value, $\boldsymbol{\Lambda}(\mathbf{K})$, generalizes the expected "underage cost," or the expected opportunity cost of having insufficient capacity. Incorporating an additional demand-shortage penalty, $\mathbf{c}_P$, increases net value, $\mathbf{v}$, by $\mathbf{R}'_D\mathbf{c}_P$, thereby increasing the underage cost. In a multiresource, multiproduct setting, the underage cost depends not only on the resource, but also on the demand-vector scenario. Both resource sharing (of the test resource) and demand dependence (both drives are imperfect substitutes) introduce coupling between different resources' marginal value, an effect that is absent in the single-dimensional model. Underage cost is balanced with overage cost, as measured by the marginal cost of excess capacity.

The newsvendor solution $\mathbf{K}^*$ is now found easily by a marginal or steepest-ascent argument. Start from the capacity vector $\mathbf{K}^b = (300, 300, 600)$, which is the lowest-cost portfolio that enables meeting the most-likely demand $\mathbf{D}^2$, which equals $\mathbb{E}\mathbf{D}$. (In other words, $\mathbf{K}^b$ would be optimal if there were no uncertainty.) Now evaluate the marginal value of an increment $\nabla\mathbf{K} > 0$, using (8):

$$\nabla\mu(\mathbf{K}^b) = \begin{pmatrix} 0 \\ 300 \\ 0 \end{pmatrix} 0.25 + \begin{pmatrix} 400 \\ 0 \\ 0 \end{pmatrix} 0.25 - \begin{pmatrix} 30 \\ 20 \\ 80 \end{pmatrix} = \begin{pmatrix} 70 \\ 55 \\ -80 \end{pmatrix}.$$

Thus, increase K_1 as long as $\nabla_{K_1}\mu > 0$, or until $K_1 = 350$, beyond which P_4 (let P_i denote $P(\Omega_i(\mathbf{K}))$) becomes 0, and $P_3 = 1/4$, and $\nabla_{K_1}\mu = 100/4 - 30 = -5 < 0$. Second, increase K_2, as long as $\nabla_{K_2}\mu > 0$, or until $K_2 = 350$, beyond which point P_1 becomes 0, and $\nabla_{K_2}\mu = -20 < 0$. Third, increasing K_3 beyond 600 is suboptimal, as that would yield $P_2 = P_3 = 0$, and $\nabla_{K_3}\mu = -80 < 0$. Similarly, decreasing K_3 below 600 is suboptimal, as that would yield $P_2 = 1/2$, and $\nabla_{K_3}\mu = 400(1/2 + P_3) - 80 > 0$. Thus, we have arrived at the unique newsvendor-network solution: $\mathbf{K}^* = (350, 350, 600)$.

The corresponding contingent activity vectors are $\mathbf{x}(\mathbf{K}^*, \mathbf{D}^1) = (150, 350)$, $\mathbf{x}(\mathbf{K}^*, \mathbf{D}^2) = (300, 300)$, and $\mathbf{x}(\mathbf{K}^*, \mathbf{D}^3) = (350, 250)$. Associated state-dependent operating profits are $\pi(\mathbf{K}^*, \mathbf{D}^1) = \$165{,}000$, $\pi(\mathbf{K}^*, \mathbf{D}^2) = \$210{,}000$, and $\pi(\mathbf{K}^*, \mathbf{D}^3) = \$215{,}000$. Expected operating profit is \$200,000 while capacity investment costs are $C(\mathbf{K}^*) = \$40{,}000 + \$65{,}500 = \$105{,}500$, so that maximal expected value is $\mu(\mathbf{K}^*) = \$94{,}500$. It is easy to verify that $\pi(\mathbf{K}, \mathbf{S}, \mathbf{D})$ is submodular,[11] which gives insight directly into the sensitivity to demand forecast parameters: The maximal expected value increases in the mean demand vector, but decreases in any demand variance or covariance terms (including any pairwise demand correlation). Thus, if products are less substitutable, expected value will suffer.

[11] For example, to show that the (sub)gradient $\partial\pi/\partial D_1 = v'\partial x/\partial D_1$ is decreasing in D_2, one must consider three scenarios: (1) If D_1 is small, then as D_2 increases from 0, $v'\partial x/\partial D_1$ remains constant at v_1 throughout Ω_0 and Ω_1; (2) if D_1 is intermediate, then as D_2 increases from 0, $v'\partial x/\partial D_1$ remains constant at v_1 throughout Ω_0 and decreases to $v_1 - v_2$ in Ω_2; (3) finally, if D_1 is large, $v'\partial x/\partial D_1$ remains constant at 0 throughout Ω_4 and Ω_3.

Discussion of the Example (Risk-Neutral). This example highlights an important conceptual distinction between deterministic and stochastic capacity planning mentioned in the Introduction: The optimal investment strategy in a stochastic model typically involves some degree of capacity imbalance which can never be optimal in a deterministic model. Capacity balance means that it is possible to fully utilize all resources simultaneously, which means in a newsvendor network:

DEFINITION 1. A capacity portfolio $\mathbf{K}$ in a newsvendor network is *balanced* if there exists an activity vector $\mathbf{x} \geq \mathbf{0}$, such that $\mathbf{Ax} = \mathbf{K}$.

In the example, the condition for a balanced capacity portfolio simplifies to $K_1 + K_2 = K_3$. In the graphic representation of Figure 2, a balanced capacity portfolio yields a rectangular feasible region, while capacity imbalance yields a rectangle with a "cut-off corner." Clearly, the capacity vector $\mathbf{K}^b = (300, 300, 600)$ is balanced, while the newsvendor network solution $\mathbf{K}^*$ exhibits a relative "capacity imbalance" $\gamma = (K_1 + K_2 - K_3)/K_3 = 1/6 = 13\%$. Capacity balance is attractive from a cost perspective, as it allows for the possibility of simultaneous full utilization of all capacities, resulting in nice accounting efficiency metrics.

Another capacity plan that may show up in practice is a plan that minimizes lost sales. In some settings, marketing managers may state that "a customer lost once is lost forever," and advocate ample capacity to prevent that. We refer to such a plan that is attractive from a revenue perspective as a "total coverage" capacity plan $\mathbf{K}^c$. In the example, the "best" total coverage is $\mathbf{K}^c = (450, 350, 700)$. The newsvendor-network solution $\mathbf{K}^*$ provides the expected-profit-maximizing *compromise between these two conflicting incentives*.

This feature of compromising two conflicting incentives is just another interpretation of the well-known property of overage-underage balance in single-dimensional newsvendor models. The reason for unbalancing capacities, however, follows from extending this purely financial argument involving uncertainty to multiple dimensions and, thus, is a feature unique to capacity portfolios. In the example, the marginal value of increasing investment in testers beyond 600 does not outweigh its cost of \$80. Therefore, there is an optimal 25% probability of not being able to meet all demand, and the optimal aggregate service level is 75%. In a network, many capacity configurations can yield the same aggregate service level. *Safety capacity* is the excess over the capacity that would be optimal if there were no uncertainty. In the example, optimal safety capacity is $\mathbf{K}^* - \mathbf{K}^b = (50, 50, 0)$. Hence, another interpretation of the optimal capacity portfolio is that it specifies the optimal amounts and locations of safety capacity in the network, thereby also specifying optimal product service levels. Thus, optimally trading off the expected value of safety capacity at different resources with its cost, results in a 100% service level for Product 2, while the higher-margin Product 1 receives a service level of 75%. Risk-neutral financial optimization, thus, specifies optimal safety-capacity amounts and locations that typically result in an unbalanced capacity portfolio because it "hedges" optimally against uncertainty.

Ramifications to Practice (Risk-Neutral). While the newsvendor-network solution, by definition, yields the highest expected profit among all capacity portfolios, it has three properties that may impact the likelihood of its implementation. First, the manager recommending a newsvendor investment plan must explain to top management why they should authorize cash to be invested in a capacity portfolio that is known in advance to be *never* fully utilized. The reason is that counterbalancing capacities provides the best operational hedge: This intentionally unbalanced portfolio yields a network capacity configuration that maximizes expected value. Obviously, if the manager could observe the actual demand and all other uncertainty in advance, she could identify precise capacity needs and would invest in a balanced capacity portfolio. Under uncertainty, however, she should "hedge her bets" and invest in excess or "safety" capacity in some resources precisely because this yields better average performance. Thus, one key driver for capacity imbalance is uncertainty (as we shall see in §6.3, risk aversion is another).

Second, given that the newsvendor solution is typically not optimal for any ex-ante known demand, it cannot be the outcome of the "typical" sales-plan driven capacity requirement planning described

in the Introduction. To put this in perspective, the capacity literature has typically focused on single-resource investment for which a famous result was first shown in the seminal paper by Manne (1961) (which is reviewed in §5.4): Given stochastic demand forecasts, the optimal capacity can be found by an "equivalent deterministic problem" that considers the same capacity-investment problem but with a perhaps modified, but always deterministic, demand and with some parameters (typically the discount rate) modified to incorporate the effect of uncertainty. This result justifies the practice of sales-plan driven capacity planning or deterministic planning based on one scenario. With a single resource, it is clear that one can always find one modified demand for which deterministic planning would yield the optimal newsvendor solution. In contrast, such an equivalent deterministic problem—and, hence, related justification of practice—typically does *not* exist for a true multiresource capacity-portfolio problem (e.g., where some resources are shared among different products). Indeed, the unbalanced optimal capacity portfolio can only be obtained by carefully weighing upside versus downside for each capacity, which are expressed by intricate and coupled optimality conditions.

Deterministic planning, on the other hand, typically yields a balanced capacity portfolio. While there is little doubt that capacity-planning software capabilities will advance in the future to incorporate these intricate conditions, just like financial asset pricing software has, the fact that the conditions are coupled has broad implications for capacity-planning methods in practice. Good capacity-portfolio planning cannot be performed independently at separate locations with only a corporate sales plan as input, but must be coordinated throughout the organization.[12] Obviously, top management knows that traditional sales-plan driven capacity planning misses uncertainty and, therefore, heuristically incorporates uncertainty by perturbing its capacity proposals before implementing them. Adopting optimal stochastic capacity-portfolio planning, however, eliminates these heuristic perturbations and automates the incorporation of uncertainty.

It seems logical to conjecture that the financial and strategic value of adopting stochastic portfolio planning should be highest for firms in volatile industries that have some resource-sharing among different products.[13] (The third property that should be considered when adopting a newsvendor capacity solution is that it defines the maximal-risk capacity portfolio, which will be discussed in §6.3.)

4.3. Game-Theoretic Capacity Investment by Multiple Agents

In many settings, capacity-investment decisions are not made in a vacuum. At a minimum, those decisions interact with external customer demand and may depend on external supply markets. These customers may have access to other firms, and these suppliers may also supply those other firms. In short, a firm's capacity decisions typically depend on, or interact with, other economic agents' decisions. Thus, it seems natural for capacity investment models to incorporate the strategic behavior of self-interested agents. A classic textbook example is capacity preemption, where an incumbent overbuilds capacity to deter entry by signaling to potential competitors that it has a small marginal cost. Thus, information asymmetry enters the picture, as well as many other game-theoretic factors, when a processing network is partitioned such that each subnetwork is controlled by a different agent.

The capacity-portfolio solution methods discussed earlier now must be augmented with a fixed-point condition to solve for Nash equilibrium capacity-investment strategies. Complexity quickly mounts and great care must be exerted to specify a tractable multiperson model. Most game-theoretic capacity analysis has, therefore, been restricted to a stationary setting where a single capacity investment is optimal, emphasizing capacity type and size, rather than timing decisions.

[12] While the process could be decentralized using incentives such as transfer prices, determining the appropriate incentives seems to require a central planner to first solve the organization-wide capacity problem.

[13] For example, pharmaceutical Eli Lilly employs a full time staff of sophisticated stochastic capacity-portfolio planners that aid in the strategic planning of new facilities for new compounds (private conversations).

Single-resource, multiagent capacity models consider networks with essentially one productive resource capability controlled by one (set of) agent(s) who interact with another set of agents without processing capabilities. The "nonprocessing" agents typically have another economic asset that is of value to the producer. Examples of such assets include market information and access, as in the typical manufacturer-retailers relationship; design and marketing capabilities, as in the more recent setting where OEMs have outsourced production to contract manufacturers; or the buying decision, as in the direct relationship of manufacturer-customers. Cachon and Lariviere (1999) consider the manufacturer's capacity investment and allocation decisions to several downstream retailers that have private information. Armony and Plambeck (2002) consider the capacity investment of a manufacturer that sells through two distributors. When supply is scarce, customers may place duplicate orders, which leads the manufacturer to overestimate both the demand and cancellation rates. This typically leads the manufacturer to purchase too much capacity, but estimation errors may lead to underinvestment when the cost of capacity is high. Plambeck and Taylor (2001) investigate the impact of bargaining and industry structure on capacity investment and profitability when OEMs outsource manufacturing to contract manufacturers. Caldentey and Wein (2003) present contracts that are linear in backorder, inventory, and capacity levels to coordinate a manufacturer-retailer production-inventory system, including the capacity decision. While not a true multiagent capacity problem, Carr and Lovejoy (2000) analyze the manufacturer-customer relationship in a setting where demand management is relatively less costly than capacity adjustment, so that a capacitated firm will "choose a demand distribution" from a set of potential customer segments that is most profitable. Lovejoy and Li (2002) study how to best expand hospital operating room (OR) capacity, which can be had by building new ORs or extending the working hours in the current ORs, acknowledging the conflicting priorities of patients, surgeons and surgical staff, and hospital administrators.

Porteus and Whang (1991) consider a principal-agent formulation where capacity and demand are a function of the private effort of the manufacturing manager and marketing managers, respectively. They present an optimal incentive plan that is interpreted as requiring the firm's owner to make a futures market for capacity, paying the manufacturing manager the expected marginal value for each unit of capacity provided, receiving the realized marginal value from the marketing managers, and on average losing money in the process. That framework was extended by Kouvelis and Lariviere (2000), who allow prices to adapt as information evolves. In a newsvendor setting, managers in the first stage receive the expected shadow price, whereas later agents are charged the realized shadow price of the resources they utilize. This linear transfer-pricing system simplifies the nonlinear incentive scheme that coordinates the "global newsvendor" of Kouvelis and Gutierrez (1997), where two agents sell an identical product in two markets.

Multiresource, multiagent capacity models consider networks where several agents own processing capacity, which implies a capacity portfolio problem. The relationship can be "vertical," meaning that agents control different "stations" in a supply chain, such as in multiechelon systems, "horizontal," with parallel firms supplying a common market, or a mixture of both. Most articles consider horizontal competition, typically with univariate uncertainty. For example, Loch (1991) considers price and capacity decisions for a duopoly in a competitive queuing model, which is extended by Lederer and Li (1997) to perfect competition. Bashyam (1996) considers capacity expansion in a two-stage setting akin to a static newsvendor network but with private, Bayesian demand-information updating. Lippman and McCardle (1997) show how a newsvendor critical-fractile solution extends to a competitive setting with multiple agents supplying a single market with fixed-price and univariate demand uncertainty. That solution critically depends on the "splitting rules" that specify how initial demand is allocated among competing firms and how any excess demand is allocated among firms with remaining capacity (or inventory). Van Mieghem and Dada (1999) discuss how the relative timing of the three major operational

decisions—capacity, processing (inventory) quantity, and price—impact the sensitivity and profitability of those decisions under demand uncertainty for a monopoly, oligopoly, and perfect competition. The relative value of postponement seems to increase as the industry becomes more competitive. (Cf. Van Mieghem and Dada 1999, for a review of earlier capacity and pricing models.)

Subcontracting refers to the setting where both the contractor and supplier have productive capacity. With outsourcing, on the other hand, the contractor has no productive capacity and relies completely on the supplier(s). Subcontracting is an example of combined vertical-horizontal relationships in the supply chain if the subcontractor also has market access. Strategic capacity choice, by both subcontractor and manufacturer, is analyzed by Van Mieghem (1999) under multivariate demand uncertainty. The coordination potential of three contract types is investigated, including an incomplete bargaining contract. The impact of demand volatility and correlation on the option value is presented, as well as conditions for when outsourcing—the extreme solution point where the manufacturer invests in zero capacity—is optimal. Cachon and Harker (2002) consider a duopoly where firms face scale economies with univariate uncertainty. Firms compete with two instruments: explicit prices and delivered operational performance. The latter includes quality of service, which is directly driven by the capacity choice of the queue's processor. They also allow each firm to outsource its processing to a supplier and identify economies of scale as a strong motivator for outsourcing. Bernstein and DeCroix (2002) analyze a newsvendor-like capacity game in a single-product modular assembly system. The final assembler moves first by setting a price-only contract specifying the unit-price she will pay to subassemblers, who then set the unit-prices they will pay to their suppliers. In the second stage, all parties independently choose their capacity level. Finally, demand is observed and all parties produce the same number of units. Equilibrium capacities and prices are characterized and used to guide network design.

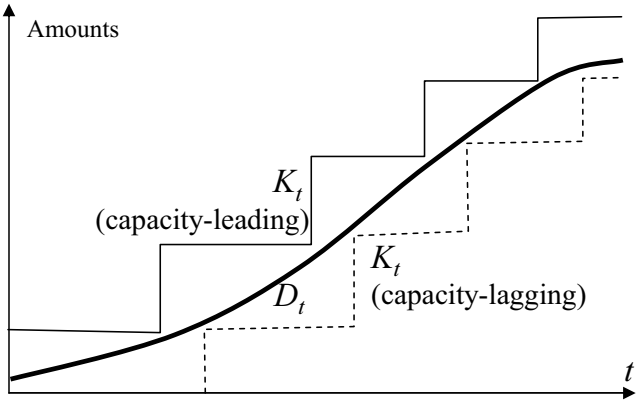
5. Optimal Capacity Adjustment over Time

This section reviews capacity investment problems that are dynamic, meaning they do not reduce to a single initial investment.

5.1. Generic Dynamic Capacity Strategies and Trade-offs

In addition to the portfolio configuration decisions on type and amounts reviewed above, dynamic capacity problems must also decide on the timing of capacity adjustments. Timing involves some general trade offs. For example, if one expects demand to increase, when should one increase capacity and by how much? In a single-product setting, the generic choice is a combination of the polar extremes of a *capacity-leading* strategy, where there are never demand shortages, and a *capacity-lagging* strategy, where there is never underutilized capacity, as shown Figure 3. Capacity lagging has the advantages of eliminating capacity-overage risk, being less dependent on accurate forecasting, and delaying capital expenditures. On the downside, however, it incurs lost sales and dissatisfied customers, which can invite entry by competitors; it has no ability to exploit the "upside" of a forecast; and it is very sensitive to any start-up problems with new capacity additions. Neither of the two polar extremes employ inventory. A hybrid strategy uses initial excess capacity to build inventory to sup-

Figure 3 Managing Capacity Over Time Involves Deciding the Capacity Adjustments' Timing and Magnitudes



ply later undercapacitated periods. While that is hard to do in a service setting (unless part of the service can be performed in advance of customer demand), it combines the advantages of high utilization and capturing all demand, at the cost (and risk) of holding inventory. Clearly, the trade-off is between the capacity investment costs and the inventory holding cost, and appropriate levels of both are driven by the trade-off between cost of underage versus overage (cf. the newsvendor-network discussion above). The appropriate strategy may also depend on where products are in their life cycle: Product-introduction requires capacity-leading, whereas maturity may move more towards inventory-smoothing or even lagging.

The magnitude of the capacity adjustment and its timing are also interdependent. A key question is whether one should have many small adjustments or few large adjustments, their polar extremes being continuous adjustment or a single, discrete adjustment. Typically, many small adjustments are the result of continuous improvement activities. Few, large adjustments, and, thus, discrete capacity changes, typically are caused by "frictions." As we shall review, the literature has identified several sources of friction: indivisibility (i.e., "lumpy capacity"), irreversibility (i.e., a kinked adjustment cost function as (3) with $\mathbf{c}_{K,t} > \mathbf{r}_{K,t}$) and nonconvexity (e.g., due to fixed cost and economies-of-scale in the adjustment cost function).

Finally, in a multiresource setting, the adjustment decisions of various resources are coupled and an important question is whether there exists a simple description of the coupled dynamics and, perhaps, a natural ordering of resources, such that capacity i is always adjusted before capacity j.

5.2. Structural Properties of Optimal Capacity Portfolio Dynamics

At each possible capacity adjustment time, the firm will base its investment decision on the information then available and on its assessment of the uncertain future. Mathematically, information availability and uncertainty, which are crucial to any investment strategy, are modeled by a standard probabilistic framework with a probability space $(\Omega, \mathcal{F}, P)$ and filtration $\mathbb{F} = \bigcup_t \mathcal{F}_t$ as primitives. The filtration $\mathbb{F}$ shows how information arrives and uncertainty is resolved as time passes, with $\mathcal{F}_t$ representing the information available at time t. Under the risk-neutral assumption, the firm's objective is to maximize its expected net present value, which is the discounted sum of operating profits $\pi_t(\mathbf{K}_t, \omega)$ minus adjustment costs $C_t(\mathbf{K}_t - \mathbf{K}_{t-1})$.

To solve the dynamic investment problem, the usual backward induction argument leads to Bellman optimality equations. Adopting a discrete-time formulation for simplicity, let $V_t(\mathbf{K}_{t-1}, \omega)$ denote the optimal value function when starting at the beginning of period t, with capacity vector $\mathbf{K}_{t-1}$ given state of the world ω. With concave operating-profit functions and convex adjustment costs, the optimal value functions $V_t(\cdot, \omega)$ inherit the concavity of the operating-profit functions π_t (Theorem 1 in Eberly and Van Mieghem 1997).

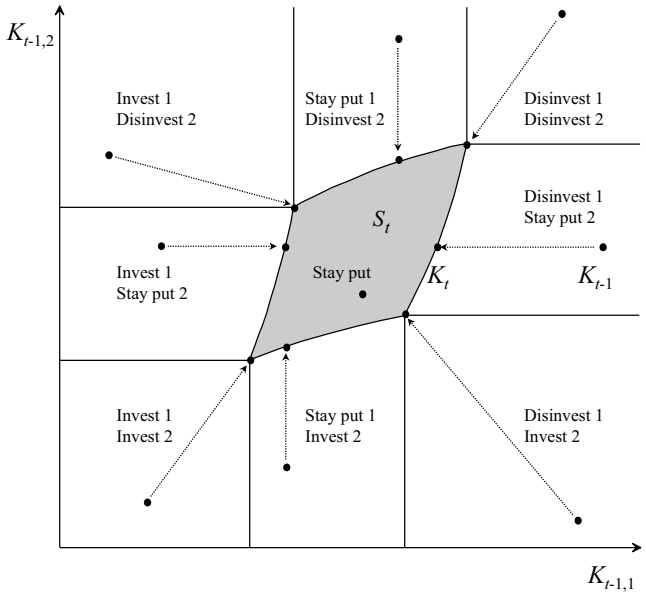
Concavity makes the control problem well behaved. Eberly and Van Mieghem (1997) show that concavity implies that it is optimal to invest according to a certain kind of control-limit policy that they call an ISD (Invest/Stay put/Disinvest) policy. Roughly speaking, an ISD policy adjusts each capacity i according to a control-limit policy defined by two critical numbers or "triggers," $K^L_{t,i} \le K^H_{t,i}$, which are functions of the critical numbers of other resources. These define three action zones: Capacity i is increased to $K^L_{t,i}$ if $K_{t-1,i} < K^L_{t,i}$ (invest), decreased to $K^H_{t,i}$ if $K_{t-1,i} > K^H_{t,i}$ (disinvest), and not adjusted otherwise (stay put). An ISD policy for the two-resource case ($n = 2$) has the structure shown in Figure 4.

How is such an optimal ISD strategy found? For ease of notation, let $g_t(\mathbf{K}_t, \omega)$ be the firm's expected net present value, evaluated at the beginning of period t and conditioned on the available information, given that capacities have been adjusted to $\mathbf{K}_t$ and an optimal (partial) investment strategy is implemented:

$$g_t(\mathbf{K}_t, \omega) = \pi_t(\mathbf{K}_t, \omega) + \delta \mathbb{E}[V_{t+1}(\mathbf{K}_t) \mid \mathcal{F}_t](\omega). \quad (9)$$

Eberly and Van Mieghem (1997) show that if the solution $\kappa_t(\mathbf{K}_{t-1}, \omega)$ to the concave optimization problem $\sup_{\mathbf{K}_t \in \mathbb{R}^n_+} \{g_t(\mathbf{K}_t, \omega) - C_t(\mathbf{K}_t - \mathbf{K}_{t-1})\}$ is unique, then κ_t is an ISD policy and is optimal for the general capacity investment problem for t, for every $t \in \{1, \ldots, T\}$

Figure 4 Structure of a Two-Dimensional ISD Policy Generated by a Supermodular Concave Function



and $\omega \in \Omega$. (Optimality extends to the infinite-horizon case under mild additional conditions.)

The concavity of g_t yields the following additional properties of an optimal ISD policy. The optimal ISD policy for period t is characterized by a connected set $S_t(\omega) \subset \mathbb{R}^n_+$ for each $\omega \in \Omega$, where

$$S_t(\omega) = \left\{\mathbf{K} \in \mathbb{R}^n_+ \colon \mathbf{r}_{K,t} \le \nabla g_t(\mathbf{K}, \omega) \le \mathbf{c}_{K,t}\right\}. \qquad (10)$$

If $\mathbf{K}_{t-1} \in S_t(\omega)$, no adjustments are made: $\mathbf{K}_t = \kappa_t(\mathbf{K}_{t-1}, \omega) = \mathbf{K}_{t-1}$; otherwise, $\mathbf{K}_{t-1}$ is adjusted to a point $\kappa_t(\mathbf{K}_{t-1}, \omega)$ on the boundary of $S_t(\omega)$. Thus, if the capacity vector $\mathbf{K}_{t-1}$ is within $S_t(\omega)$, it is optimal "to stay put" on all dimensions and to continue with the same capacity portfolio for the next period. Hence, the set S_t is also called the "region of inaction" or "continuation region." Its boundaries are increasing (decreasing) if all operating profit functions, $\pi_t(\cdot, \omega)$, and the salvage function, $f(\cdot, \omega)$, are supermodular for each ω. The capacities are then economic complements (substitutes) so that a higher optimal investment threshold in resource k justifies a higher (lower) optimal investment threshold in resource j, and vice versa.

5.3. Optimal Investment Dynamics and Sources of Friction

When the time period becomes arbitrarily small, one can instantaneously adjust capacity. Assuming sufficient regularity, a control problem in continuous time obtains where the central region S_t may move continuously over time. Only the initial capacity investment may represent an "impulse" control to the boundary of S_0 if the initial state $\mathbf{K}_0$ is outside S_0. All subsequent optimal capacity adjustments are "barrier" or "instantaneous" controls. No control is needed as long as $\mathbf{K}_t$ remains inside S_t and a period of inaction ensues. When $\mathbf{K}_t$ "hits" the boundary of S_t, the minimal amount of control is exercised to prevent $\mathbf{K}_t$ from "leaving" S_t. Thus, the optimal investment dynamics are similar to the dynamics of a point, $\mathbf{K}_t$, floating inside a (typically moving) domain, S_t, being reflected on its boundaries. This succession of periods of inaction followed by bursts of capacity adjustment coincides with empirically observed patterns.

The width of the continuation region along coordinate i is an increasing function of the amount of *irreversibility* $c_{K,t,i} - r_{K,t,i}$ of resource i investment. The continuation region shrinks to a single point, representing the optimal capacity vector at t, if investment is reversible (i.e., $\mathbf{r}_{K,t} = \mathbf{c}_{K,t}$). With such "frictionless" investment, capacity is almost always adjusted because the adjustment can be "undone" at no cost in the future. If investment is irreversible (i.e., $\mathbf{r}_{K,t} < \mathbf{c}_{K,t}$), S_t represents a hysteresis zone where optimal capacity at time t depends on previous capacity at time $t-1$, and capacity is adjusted only if period t's operating profit and information outlook are sufficiently different from the previous period. Thus, irreversibility introduces "friction" in the optimal capacity-investment policy which exhibits periods of action (invest or disinvest) interspersed with periods of inaction.

Indivisibility, such as discreteness in possible capacity adjustment sizes—"lumpy capacity"—are a second source of nonlinearities and friction that may increase the region of inaction and distort the structure of the optimal policy. Narongwanich et al. (2002) consider a single-product setting where new product generations are introduced stochastically over time,

and study whether one should invest in product-dedicated capacity, which is only used for one product life cycle, or reconfigurable capacity. They show that the ISD policy remains optimal if all resources have identical adjustment sizes. With different adjustment sizes, the optimal policy is ISD-like but with perturbations around its boundary, probably reflecting adjustments from integer restrictions.

Fixed costs in the capacity-adjustment cost function are a third source of friction and introduce additional nonlinearities that increase the region of inaction, as shown in Abel and Eberly (1998). Fixed capacity-adjustment costs have a similar effect as fixed ordering costs in inventory theory; indeed, ISD policies are related to a multidimensional generalization of the famous two-critical number (S, s) control policy. In general, nonconvex costs or nonconcave operating profits introduce frictions that will lead a firm's optimal policy to exhibit occasional large changes, or discontinuities, of the kind that arise with impulse controls. For example, Dixit (1995) considers a convex adjustment-cost function, but the production function is convex-concave, exhibiting initially increasing returns to scale, and then decreasing returns to scale. This produces similar effects to nonconvex capacity costs: The optimal capacity "jumps over" the increasing returns portion but is constrained by the eventual decreasing returns.

5.4. Dynamic Capacity Models: Univariate Uncertainty

Characterizing the continuation region, S_t, and its dynamics is a formidable problem. Only when uncertainty is generated by "nice and amenable" stochastic processes can it be analytically determined. While capacity-portfolio studies with multivariate uncertainty are typically restricted to an i.i.d. setting, dynamic studies are typically restricted to univariate uncertainty. In addition, most models assume a univariate stochastic process $X = \{X_t(\omega)\colon\ \omega \in \Omega,\ \ t \geq 0\}$ with independent increments. (While scenario-based or stochastic-programming models can also capture dynamics, here, attention goes to papers that analytically describe the investment dynamics.) Often, X is a Markov process, such as a Markov chain or Brownian motion, to maintain tractability.

Dynamic single-resource capacity models with stochastic demand seem to have started with the seminal work of Manne (1961), where demand follows a Brownian motion with positive drift and only capacity expansions are considered. Hence, a regenerative process obtains and it is optimal to always add capacity in the same increment ΔK, which is an increasing function of variance, whenever the demand backlog hits a given "trigger" level (the critical ISD number). The associated expansion times can be expressed in terms of the "hitting time" of the Brownian motion. Manne (1961) showed that the stochastic problem "does little—if anything—to complicate matters" compared to the deterministic problem with known demand. Specifically, he showed that the stochastic problem where capacity shortages are not allowed can be transformed into what has become known as an "equivalent deterministic problem" by replacing the discount rate by a decreasing function of the variance rate σ^2 of the demand process. Thus, the entire effect of uncertainty is captured by a single number: The modified or "equivalent" discount rate $r^* = (\sqrt{1+2r\sigma^2}-1)/\sigma^2$, where r is the riskless rate. (Such equivalent deterministic problem, however, does *not* exist for the multi-resource capacity-portfolio problem, as discussed in §4.2.) Equally important to practice, Manne (1961) showed that the cost function is relatively insensitive to the capacity increment, ΔK, so that errors in parameters are rather inconsequential (similar to the EOQ model).

Various extensions and modifications followed. For example, Giglio (1970) considered a different nonstationary, increasing demand structure of the form $D(t) = at + \epsilon_t$, where $\mathbb{E}\epsilon_t = 0$. Thus, demand mean and variance are independent of each other, and the variance of ϵ_t is either constant or linearly increasing in t. Freidenfelds (1981) considers a series of models where demand is Markovian and represented by birth-death processes. Bean et al. (1992) consider capacity expansion when demand is a semi-Markov process and no demand shortages are allowed. They define the equivalent deterministic problem and show that it exists whenever the demand is a transformed Brownian motion or a regenerative birth-and-death process. The equivalent deterministic problem, again, uses a lower

equivalent interest rate, and the equivalent deterministic demand can be found from a regression on observed demand. Davis et al. (1987) take a different and quite unique approach: They suppose that the decision maker exercises day-to-day control over a capacity installation project by controlling the rate of investment. The demand is a Poisson process and is met by consecutive construction of identical expansion projects of given size. Sophisticated stochastic control theory ultimately shows that the control is "bang-bang": Either carry out construction at maximal speed or do nothing and wait, similar to an ISD policy.

Angelus and Porteus (2002) consider expansion followed by contraction through a single product's life cycle. The driving stochastic process, X, represents demand, which is first stochastically increasing over a known interval of time, after which it is stochastically decreasing. They explicitly solve for the two critical numbers that define the continuation region, S_t, at each period. It is shown that it is optimal to change the service level (i.e., the probability of meeting demand) over time: Provide the lowest service level during the peak period and the highest during contraction periods. In addition, they provide initial insights into the much more complex setting where inventory is carried between periods.

Dynamic multiresource or capacity portfolio models are often continuous-time models and assume some special structure that simplifies the treatment of the source of (typically univariate) uncertainty, X. Eberly and Van Mieghem (1997) provide an example where X is univariate Markovian. In addition, the general operating-profit function is assumed stationary ($\pi_t = \pi$ and $C_t = C$), Markovian (in the sense that it only depends on the current capacity vector and current state of nature so that $\pi_t = \pi(\mathbf{K}_t, X_t)$), supermodular and linearly homogeneous in X and $\mathbf{K}$, and increasing in X. Homogeneity of degree 1 is powerful in reducing complexity because all dynamics can be described in a scaled coordinate system K_i/X_t or $k_i = \log(K_i/X_t)$. To appreciate the significance of this result, they show that the central domain is now fixed in $\mathbf{K}/X$-space: $S(X_t) = \{\mathbf{K} \in \mathbb{R}^n_+ : \mathbf{r} \le \nabla V(\mathbf{K}/X_t, 1) \le \mathbf{c}\}$. In k-space, the investment dynamics reduce to those of a "particle" moving along the line $\log(\mathbf{K}) - \log(X_t)$ in the interior of S and expanding or contracting capacities when hitting a boundary of S. The important result here is that, because S is fixed, in equilibrium only a few of the $2n$ faces of S are ever hit, and always in the same sequence. Therefore, the order and frequency of capacity adjustments—or their "flexibility" in economic terms—is endogenous in the capacity model: Some resources will be systematically adjusted more often than other resources and, thus, are systematically "more flexible" than other resources. That ordering follows the ordering of the width of S measured along the coordinate directions. As discussed above, the width of the region of inaction is increasing in the difference $\mathbf{c}_K - \mathbf{r}_K$. If the Markov process, X, is a univariate geometric Brownian motion and the operating-profit function, π, is derived from a constant returns-to-scale Cobb-Douglas production function and a constant elasticity demand function, the dynamics can be solved in closed form. It then is shown that the fractions $r_{K,i}/c_{K,i}$ endogenously determine the resource adjustment ordering or "flexibility."

The fixed flexibility ordering inspired the "bottleneck policies" of Çakanyildirim and Roundy (2002). Like Angelus and Porteus (2002), they consider a single product's life cycle, but assume a multiresource production process with lumpy capacity in a cost-minimization context. They show that the sequence in which resources are expanded in the first up-part of the cycle is the reverse of that in which they are contracted in the down part. Clearly, in a single-product setting with lumpy capacity, the resource adjustment sequence is easier to determine because only capacity adjustment of bottleneck resources can be optimal. However, the timing of those adjustments is nontrivial because it may be optimal to adjust several resources simultaneously. Çakanyildirim and Roundy (2002) provide an algorithm to calculate the adjustment times and, hence, determine the "clusters" of resources that are adjusted simultaneously. Hu and Roundy (2002) extend Çakanyildirim and Roundy (2002) and present an algorithm to calculate the adjustment times in a multiproduct, multiresource setting with stochastically increasing demand. In both papers, adjustment times are determined at

the beginning of the horizon (instead of dynamically over time). Narongwanich et al. (2002), reviewed above, analyze the choice between lumpy dedicated or reconfigurable capacity when new generations of a single product are introduced stochastically over time.

Analyses of specific instances of the general dynamic capacity-portfolio investment analysis with multivariate uncertainty, which is necessary to model correlations between product demands or input supplies, seem virtually nonexistent. A likely reason is that with multivariate uncertainty, the characterization of the continuation region in continuous time typically reduces to a partial-differential equation with nonstandard boundary conditions. Typically, analytic solutions are not available, so one must resort to numerical analysis. Discrete time does not seem much easier, except perhaps, if one considers only a few periods. Kouvelis and Milner (2002) analyze the two continuation regions in a two-period model where a single product is provided through a multiresource process comprised of either in-house production or outside sourcing. In addition to product demand, the outside supply capacity is also stochastic.

6. Risk Aversion and Hedging in Capacity Investment

This section reviews how risk aversion is incorporated in capacity problems and how financial and operational hedging can reduce the risk associated with capacity investments.

6.1. Moving Beyond Maximizing Expected Present Values

Capacity investment often involves substantial certain cash outlays in order to receive uncertain future rewards. It seems natural to consider the variability in payoffs in addition to the average payoff. From an economic theory perspective, however, it is not obvious that firms should care about risk. Indeed, this question amounts to asking: What is the objective of the firm? Without attempting to summarize the field of corporate finance, let us discuss some important issues. First, the objective depends on the ownership structure. For a privately owned company, the objective may be the maximization of the owners' expected utility. The appropriateness of that objective, however, depends on whether the owner is diversified or not. If a majority of an owner's assets are tied up in the firm, the capacity-investment decision may materially impact the owner's utility. If, on the other hand, the owner is well diversified,[14] the total variability of her asset portfolio will impact the owner's utility. In that case, the impact of the capacity-investment decision on expected utility will be driven by the covariance of the returns with those of the market. For a publicly owned company, the objective is typically stated as maximizing the market value of cash flows, where that value is priced in competitive markets. If the market is "perfect," the capital-asset pricing model dictates that the value is linear in the expected return if the firm's returns are uncorrelated with the market returns,[15] in which case, risk-hedging does not enhance value. If the market has imperfections, however, nonlinearities enter into the objective and firms do care about risk. Typical market imperfections arise from the cost of bankruptcy and related financial distress, cost of raising external capital, taxes, etc. In addition, agency concerns lead to risk aversion: To motivate risk-averse managers, the firm's owners may give them a stake in the firm. Finally, it is well documented that "executives (including those who are risk-seeking) make substantial effort to reduce or eliminate risk, usually by delaying decisions and by collecting more information," according to Pindyck and Rubinfeld (1989).

Given the significance of the cash outlays, the number of capacity-investment articles incorporating risk is surprisingly small. While Caldentey and Haugh (2003) present a more general model setting, the majority of the few available risk studies in operations management are done in an inventory context, as reviewed by Chen and Federgruen (2000), Ding and Kouvelis (2001), and Gaur and Seshadri (2002).

[14] Actually, diversification eliminates only "diversifiable risk," and nondiversifiable or "systematic" risk remains because the capacity return may depend on the overall economy.

[15] If the firm's returns are correlated with market returns, firm value can still be constructed to be linear in expected returns, if the expectation is taken using risk-adjusted state probabilities; i.e., under the equivalent Martingale measure (see later).

This is, however, an active research area and one can expect substantial future activity.

When firms care about risk, a fundamental problem is how to model risk behavior. According to Kreps (1990), the predominant models for choice under uncertainty are the von Neumann-Morgenstern preferences and expected utility theory, where probabilities are objective, and the Savage (or Anscombe-Aumann) model, where probabilities are subjective. Here, we adopt the von Neumann-Morgenstern paradigm that simplifies the choice over outcomes of *terminal wealth* to the maximization of the expected utility of those outcomes. Several new important considerations arise when adopting this paradigm:

First, endowment or wealth enters the model. While one maximizes the expected utility of terminal wealth, typically, the model primitive is initial endowment or wealth W (or a budget constraint). The investment options and returns link initial wealth to terminal wealth. In line with real options theory, it is important to incorporate all available investment options. At a minimum, the investor can choose to keep some of her money in a riskless asset with certain (risk-free) return r. For simplicity, consider the one-period investment problem,[16] in which case, terminal wealth equals $\pi(\mathbf{K}, \omega) + (1+r)(W - C(\mathbf{K}))$.

Second, can one invest beyond the initial endowment? Theoretically, this possibility is automatically incorporated *if* one assumes a perfect capital market in which one can borrow without limitations at the risk-free interest rate. While we shall assume so for simplicity, in practice this does not hold, and the debate then moves to how one finances the investment. Choosing the "right" mix between new equity or debt defines the capital structure of the firm, which is beyond the scope of this paper. Stochastic capacity models assume (often implicitly) either a perfect capital market or investment costs that are relatively small compared to the value of the firm, such that capital structure is unaffected. (Refer to Stenbacka and Tombak 2002, for an overview of financing complications.) Then, a risk-sensitive investor will choose a capacity vector that maximizes the expected utility $U(\mathbf{K})$ of terminal wealth, where

$$U(\mathbf{K}) = \mathbb{E}u(\pi(\mathbf{K}, \omega) + (1+r)(W - C(\mathbf{K}))), \qquad (11)$$

and the utility function $u(\cdot)$ is strictly increasing (such that the investor prefers "more over less") and concave (such that the investor is "risk averse," meaning that the expected value is preferred over the risky outcome).

Third, what is the form of the utility function u? Among the myriad forms, $u(x) = -e^{-z_o x}$ $(z_o > 0)$ is theoretically and mathematically appealing. It models a decision maker with constant coefficient of absolute risk aversion $z(x) = -u''(x)/u'(x) = z_o$. One of its appealing features is that the capacity decision becomes independent of initial wealth (if one can borrow without limitations). Indeed, the expected utility in (11) simplifies to

$$U(\mathbf{K}; z_0) = -e^{-z_o(1+r)W} \cdot \mathbb{E}\exp\big(-z_o(\pi(\mathbf{K}, \omega) - (1+r)C(\mathbf{K}))\big).$$

A second appealing simplification happens if the operating profits $\pi(\mathbf{K}, \omega)$ are normally distributed. Denote its mean and variance by $\Pi(\mathbf{K})$ and $\sigma^2(\mathbf{K})$. Invoking the characteristic function of the normal distribution $\phi_\pi(t) = \mathbb{E}\exp(i\pi) = \exp(it\Pi - (1/2)\sigma^2 t^2)$ then yields:

$$\begin{aligned} &U(\mathbf{K}; z_0) \\ &\quad = -e^{-z_o(1+r)(W - C(\mathbf{K}))}\phi_\pi(iz_o) \\ &\quad = -\exp\big(-z_o(1+r)(W - C(\mathbf{K})) - z_o\Pi(\mathbf{K}) + \tfrac{1}{2}\sigma^2(\mathbf{K})z_o^2\big). \end{aligned}$$

In summary, under the assumptions of perfect capital markets, constant absolute risk aversion, and normally distributed operating profits, the von Neumann-Morgenstern framework of maximizing $U(\mathbf{K})$ is equivalent to maximizing

$$U_{MV}(\mathbf{K}; z) = \mu(\mathbf{K}) - z\sigma^2(\mathbf{K}), \qquad (12)$$

where $z = z_o/2$ is called the "risk parameter," and μ and σ^2 are the mean and variance of firm value expressed in end-of-period monetary units:

$$\mu(\mathbf{K}) = \mathbb{E}\pi(\mathbf{K}, \omega) - (1+r)C(\mathbf{K}) = \Pi(\mathbf{K}) - (1+r)C(\mathbf{K}).$$
$$\sigma^2(\mathbf{K}) = \mathbb{E}\pi^2(\mathbf{K}, \omega) - (\mathbb{E}\pi(\mathbf{K}, \omega))^2.$$

[16] Incorporating risk considerations over sample time paths is vastly more complicated and beyond the scope of this paper. Classic references include Merton (1971) and Kreps and Porteus (1978). Schroder and Skiadas (1999) present state-of-the-art formulations and results.

(Recall that $\delta = (1+r)^{-1}$, so that these definitions are equivalent to (6), which was expressed in present-value monetary units.) The objective (12) mirrors the celebrated mean-variance formulation of financial-portfolio theory developed and reviewed by Markowitz (1991).

Mean-variance formulations have two significant benefits: They are *implementable* (i.e., only two moments are required, which can be estimated) and *useful* in the sense that they provide "good recommendations," even when the decision maker does not know her utility function. The benefit of usefulness relates to the important concept of the Pareto-optimal, or efficient, frontier, which can be defined as follows. For generality's sake, let $\mathcal{R}(\mathbf{K})$ denote a measure of risk of the capacity investment $\mathbf{K}$ and assume the family of "quasi-utility" functions

$$U_{\mathcal{R}}(\mathbf{K}; z) = \mu(\mathbf{K}) - z\mathcal{R}(\mathbf{K}), \tag{13}$$

with risk parameter $z \geq 0$. Review the traditional definitions:

Definition 2. Capacity portfolio $\mathbf{K}$ is $\mathcal{R}$-efficient if and only if there does not exist another portfolio $\mathbf{K}'$ such that either $\mu(\mathbf{K}') > \mu(\mathbf{K})$ while $\mathcal{R}(\mathbf{K}') = \mathcal{R}(\mathbf{K})$, or $\mathcal{R}(\mathbf{K}') < \mathcal{R}(\mathbf{K})$ while $\mu(\mathbf{K}') = \mu(\mathbf{K})$.

Definition 3. The efficient frontier $\mathcal{F}_{\mathcal{R}}$ is the set of risk-return pairs of $\mathcal{R}$-efficient portfolios:

$$\mathcal{F}_{\mathcal{R}} = \big\{(\mathcal{R}(\mathbf{K}), \mu(\mathbf{K})) \colon \mathbf{K} \text{ is } \mathcal{R}\text{-efficient}\big\}.$$

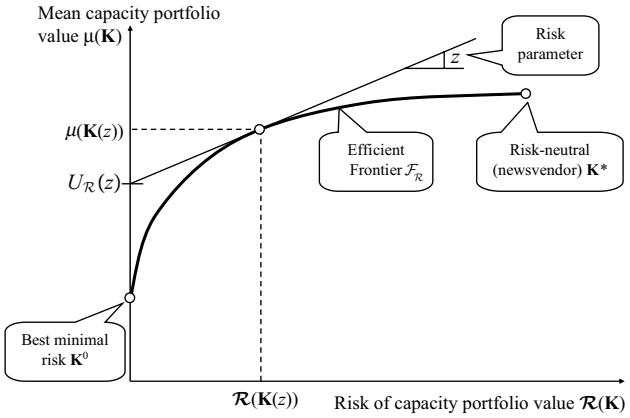
Loosely speaking, under mild regularity conditions, the frontier is the northwest boundary of the set of all risk-return pairs $\{(\mathcal{R}(\mathbf{K}), \mu(\mathbf{K})) \colon \mathbf{K} \in \mathcal{K}\}$. Depending on the risk measure, process structure, and uncertainty, the frontier may be a concave function. Then there is a one-to-one correspondence to efficient investments and investments that maximize the (quasi-)utility $U_{\mathcal{R}}(\mathbf{K}; z)$, as illustrated in Figure 5. In that case, we can define $\mathbf{K}(z)$ as the efficient capacity portfolios that maximize $U_{\mathcal{R}}(\mathbf{K}; z)$:

$$\mathbf{K}(z) = \arg\max_{\mathbf{K} \in \mathcal{R}^R} U_{\mathcal{R}}(\mathbf{K}; z) \quad \text{and}$$
$$U_{\mathcal{R}}(z) = U_{\mathcal{R}}(\mathbf{K}(z); z). \tag{14}$$

If the quasi-utility $U_{\mathcal{R}}(\mathbf{K}; z)$ is strictly concave in $\mathbf{K}$, the maximizer, $\mathbf{K}(z)$, is unique. Then, $\mathbf{K}(\cdot)$ is a path or curve in $\mathbf{K}$-space that summarizes everything one needs to know to implement the optimal investment recommendation. Hence, we shall refer to $\mathbf{K}(z)$ as the *efficient capacity path* or the relevant set of "good recommendations" referred to above, which are those that are on the efficient frontier. To conclude the risk-averse capacity portfolio problem, all that is needed is a reasonable estimate of the decision maker's risk attitude z. Thus, the efficient path and frontier embody the trade-off between expected return and risk. In practice, the decision maker can be presented with (a subset of) the efficient investment path, starting from the risk-neutral solution $\mathbf{K}^*$ and moving to the lowest risk efficient plan by tightening the risk parameter z. Moving from von-Neumann-Morgenstern preferences to mean-variance preferences, however, cannot be done, in general. Indeed, the expectation of $u(\pi)$, typically, cannot be expressed in terms of the first and second moment only, except under the special assumptions above or when the utility function u is concave quadratic. A debate has been held on the appropriateness of mean-variance formulations. Markowitz (1991) states:

> So, equipped with database, computer algorithms and methods of estimation, the modern portfolio theorist is able to trace the mean-variance frontiers for large universes of securities. But is this the right thing to do for the investor? In particular, are mean and variance proper and sufficient criteria for portfolio choice?...

Figure 5 **The Risk-Return Frontier $\mathcal{F}_{\mathcal{R}}$ and Its Correspondence to the Efficient Capacity Portfolio $\mathbf{K}(z)$ and Maximal Value $U_{\mathcal{R}}(z)$ of the Quasi-Utility Function $U_{\mathcal{R}}(\cdot; z)$ for a Fixed Risk Parameter z**



> We seek a set of rules which investors can follow in fact—at least investors with sufficient computational resources. Thus we prefer an approximate method which is computationally feasible to a precise one which cannot be computed. I believe that this is the point at which Kenneth Arrow's work on the economics of uncertainty diverges from mine. He sought a precise and general solution. I sought as good an approximation as could be implemented. I believe both lines of inquiry are valuable.

Thus, while mean-variance preferences are consistent with utility theory only under very limiting assumptions, the crucial question is: How much utility do we "give up" when maximizing the mean-variance quasi-utility? Levy and Markowitz (1979), and Kroll et al. (1984) studied this question in the following sense: If you know the expected value and variance of a probability distribution of return on a portfolio, can you guess fairly closely its expected utility? They found that the correlation between the predicted expected utilities and the actual expected utilities was extremely high, usually exceeding 0.99. (This was the case for a variety of utility functions and nonnormal probabilities.) In other words, there is evidence that mean-variance formulations suggest "good recommendations" in the sense that they yield close-to-maximal utility.

Nevertheless, while widely used and appealing to practice, mean-variance preferences have serious limitations. Most importantly, they treat positive deviations from the mean ("upside") symmetrically with negative deviations ("downside"). Indeed, variances give equal weight to deviations above or below the average. This implies that dominating strategies may not lie on the frontier: There may be investment plans that yield returns higher, or at least as high, in each state of nature as an investment plan on the frontier. That has led to the suggestion of other approaches to incorporate risk, including asymmetric preferences such as reviewed in Nawrocki (1999):

$$\mathscr{R}_1(\mathbf{K}) = \mathbb{E}((\mu(\mathbf{K}) - \pi(\mathbf{K}))^+)^2$$
"below-mean semivariance,"

$$\mathscr{R}_2(\mathbf{K}) = \mathbb{E}((t - \pi(\mathbf{K}))^+)^2$$
"below-target t semivariance,"

$$\mathscr{R}_3(\mathbf{K}) = \mathbb{E}(t - \pi(\mathbf{K}))^+$$
"expected below-target t risk."

Eppen et al.'s (1989) argument for their use of expected downside loss, $\mathscr{R}_3$, is based on the fact that the histogram of profits with actual data show that the distribution of profits was not nearly normal or even symmetric and that $\mathscr{R}_3$ only adds a simple linear constraint (while variance makes the problem nonlinear and is computationally less tractable). Another related approach, called *robust planning* (cf. Paraskevopoulos et al. 1991, Laguna 1998, Malcolm and Zenios 1994) is to have the risk measure, say $\mathscr{R}_4$, denote an increasing function in the "uncertainty-sensitivity of the return $\pi(\mathbf{K}, \omega)$." For example, with multivariate uncertainty (e.g., ω represents a "noise" vector), one suggestion for $\mathscr{R}_4(\mathbf{K})$ is $J_\omega' \Sigma J_\omega$, where $J_\omega = \nabla_\omega \pi(\mathbf{K}, \omega)|_{\omega=0}$ and Σ is the covariance matrix of ω. A closer view reveals that robust planning with this sensitivity term is intimately connected to mean-variance analysis. Its aim was to provide a practical approach for handling noisy data and uncertainty and it originated from a stochastic control, as opposed to utility maximization, perspective. The ideas of robust planning have recently been adopted in economics by Hansen and Sargent (2001) and others, as summarized in the module on "robustness to uncertainty" (May 2001 issue of *American Economic Review*), which also gives arguments against this more heuristic approach to uncertainty and ambiguity.

Instead of direct maximization of expected utility or other more heuristic risk objectives, risk sensitivity can also be incorporated through an option valuation framework in a financial market equilibrium model if an arbitrage-free market for trading and short-selling capacity assets exists. This approach is discussed next.

6.2. Mitigating Capacity-Investment Risk with Financial Hedging

Risk-averse decision makers may be interested in mitigating risk in the capacity-investment decision. Mitigating risk, or hedging, involves taking counterbalancing actions so that, loosely speaking, the future value varies less over the possible states of nature. If these counterbalancing actions involve trading financial instruments, including short-selling, futures, options, and other financial derivatives, we call this *financial hedging*. If, on the other hand, no

financial instruments are involved in the counterbalancing actions, we speak of *operational hedging*, which is the focus of the next section.

The relationship between investment in financial instruments and capacity-investment is immediate by recognizing that the effect of capacity is identical to selling a call option to the firm's demand above its capacity. Birge (2000) shows how this analogy is used to value capacity in the presence of demand uncertainty, and how it can be integrated in a linear capacity-investment model. Financial hedging yields an elegant approach to price present values using risk-neutral discounting and to incorporate risk without having to resort to utility functions. The basic idea is to construct a "perfect hedge," which is a portfolio that provides a constant future value in any state of nature and, therefore, can be priced using risk-free discounting.[17]

The classic example of a perfect hedge is to buy a risky asset (e.g., α shares in a company's stock at share price S_0) today, and to sell a ticket that entitles its bearer to buy one share of stock at a terminal date T, if she wishes, for a specified "strike" price s. (Thus, the ticket is a European call option with future value $X_c = (S_T - s)^+$, where S_T is the stock price at time T.) For simplicity, consider a two-date economy with dates indexed $t = 0, T$ and with uncertainty at time T modeled by two states of nature. Call the two states "up" and "down," so that S_T is either S_0u or S_0d, where $u > d$ are the up and down percentages. The only interesting case is to specify a strike price s such that $S_0d \le s \le S_0u$. The future value of the portfolio then is either $\alpha S_0u - (S_0u - s)$ in the up state when we pay $S_0u - s$ to the owner of the call, or αS_0d in the down state when the call is worthless. Clearly, both payoffs are equal if α equals $\alpha^* = (S_0u - s)/(S_0u - S_0d)$, which yields a perfect hedge. With the perfect hedge, the portfolio value is riskless and, in an arbitrage-free market, its expected value must increase at the same rate as a risk-free asset with return r. Therefore, the present value of the portfolio is found using risk-free discounting to be $\delta\alpha^*S_0d$, where, as before, $\delta = (1+r)^{-1}$ is the period's risk-neutral discount factor. In market equilibrium, there is no "free lunch," so that the present value of the portfolio must equal $\alpha^*S_0 - p_c$, the present value of α^* shares and a sold call. Thus, this arbitrage condition prices the call at $p_c = \delta\alpha^*S_0(1 + r - d)$.

The interesting facts here are that: (1) This price is independent of the particular risk-attitude of the buyer of the call, and (2) the perfect hedging share α^* is independent of uncertainty. Pursuing the latter fact, one can choose any probability measure over the future states without affecting the option price. A most useful choice is to adopt the famous *equivalent Martingale measure*[18] of Harrison and Kreps (1979), whose expectation operator $\mathbb{E}^*$ is such that the price p_X of any claim with future value X is $p_X = \delta\mathbb{E}^*(X)$. In other words, the equivalent Martingale measure separates risk from time valuation by adjusting state probabilities such that its expectation of future value is risk adjusted while discounting can be risk neutral. In our example, the equivalent Martingale measure simplifies to the "risk-neutral state probabilities" p^* and $1 - p^*$ for the up and down state, respectively. One easily verifies that there exist a unique p^*, such that $p_c = \delta\mathbb{E}^*(X_c) = \delta p^*(Su - s)$ and $S_0 = \delta\mathbb{E}^*(S_T) = \delta(p^*S_0u + (1-p^*)S_0d)$. The results from this basic idea extend to continuous time and continuous state-space, yielding the celebrated Black-Scholes formula.

To summarize, option pricing of financial hedging provides a powerful tool to value risky assets via risk-neutral discounting using the equivalent Martingale measure. It circumvents the need to specify utility functions or risk-adjusted discount rates. Such powerful results come at a cost, however. In our setting, it requires the existence of a market that: (1) trades options on the firm's and all its competitors' capacities, (2) is perfect (it allows continuous-time trading without transaction fees and without restrictions on short selling), (3) is arbitrage-free, and (4) is complete (there exists a self-financing trading strategy

[17] The theoretical justifications come from the fundamental theorems of asset pricing stating that (1) such perfect hedge is possible if the market is "complete," and (2) its price is unique if the market is "arbitrage-free."

[18] The fundamental theorems of asset pricing can be restated in terms of the equivalent Martingale measure (EMM) as (1) the market is arbitrage-free iff there exists an EMM, and (2) in an arbitrage-free market the EMM is unique iff the market is complete.

that replicates the firm's cash flows so that a perfect hedge can be constructed). (While the option pricing framework is no longer consistent with utility maximization when some of the conditions are relaxed, it still provides practical use. For example, even imperfect hedges are valuable and, thus, so is hedging using financial instruments that are correlated with the firm's profits, as shown by Gaur and Seshadri 2002.) With the advent of the Internet, such capacity options markets may become reality, but probably only for certain specific industries. Financial hedging requires writing an unambiguous contract that specifies capacity usages in a form that is divisible, tradeable, and enforceable. While possible in some single-commodity settings, such contract specification is not obvious in an idiosyncratic multiproduct, multiresource setting where actual capacity depends not only on investment levels, but also on practically any aspect of operating that process.

6.3. Mitigating Capacity-Investment Risk with Operational Hedging

Besides financial hedging, mitigating risk can also be accomplished by operational hedging, which involves counterbalancing actions that do not require trading financial instruments. Operational hedging, sometimes called *natural hedging*, includes various types of processing flexibility such as dual-sourcing, component commonality, having the option to run overtime, dynamic substitution, routing, transshipping, or shifting processing among different types of capital, locations, or subcontractors, holding safety stocks, having warranty guarantees, etc. An interesting characteristic of operational hedging is that it may allow one to actually exploit uncertainty. For example, while a firm may hedge currency risk with forward contracts, a firm that has production locations in two countries may be able to ex-post shift production to the preferential country, as analyzed by Huchzermeier and Cohen (1996). Similar benefits accrue to the "global newsvendor" of Kouvelis and Gutierrez (1997) with one production location but an ex-post transshipment option between countries. Clearly, such operational hedging may involve additional costs. For example, multilocation processing incurs a loss of scale, requires procurement from a wider supply base, slows down the learning curve process, and may produce less-consistent quality.

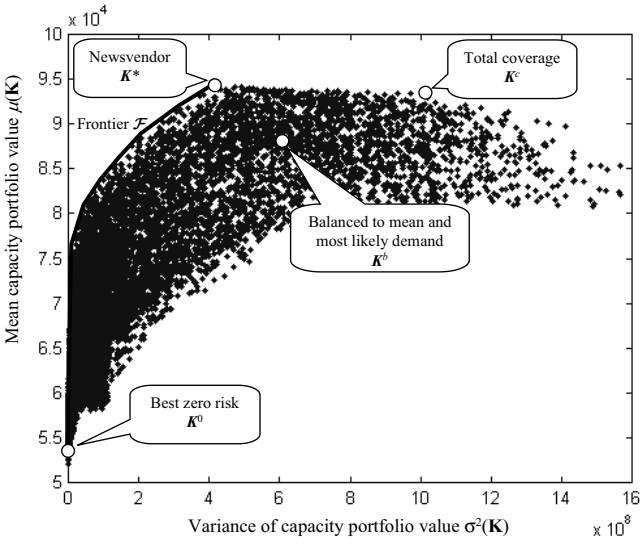
Obviously, a firm could simultaneously use both financial and operational hedging. Allayannis et al. (2001) empirically support the hypothesis that simply having multicountry production provides additional value to financial hedging. (Their econometric model does not incorporate any ex-post operational decisions.) Ding and Kouvelis (2001) explore the use of both hedging instruments in a univariate, single product, two-stage model. A domestic manufacturer must ex-ante invest in domestic capacity to produce goods that will be sold in a foreign market. In addition to a financial currency forward contract, the firm can ex-post decide its processing level. Chod et al. (2003) study the interdependence of operational hedging, using postponement of capacity and pricing, and financial hedging, using a financial derivative contract whose payoff is imperfectly correlated with the operating profit. While intuition may suggest that operational and financial hedging may be substitutes, they show that they actually can be complements.

Sometimes, however, complementing operational hedging with financial hedging may not be possible. For example, the planning horizon for a production facility may exceed 10 years. While operational hedging can be used, it is unlikely that financial hedging over that time-horizon is available. Financial hedging of capacity is also problematic if there does not exists a capacity futures market with the characteristics described above.

In a capacity portfolio setting, capacity imbalance is a natural form of operational hedging. As discussed in the Introduction and in §4.2, such imbalance is worthwhile even in a risk-neutral setting. The impact of operational hedging through capacity imbalance in a risk-averse, mean-variance setting is analyzed in Van Mieghem (2003b) and some new questions are addressed. For example, of a practical concern, one wants to know the direction along which one should adjust the risk-neutral capacity vector, so as to trace the frontier and achieve a risk-optimal operational capacity hedge. Related, from a design and performance analysis perspective, one wants to know how effective operational hedging is for a given network, i.e., how much return must be given up for a decrease

in risk? An answer to these two questions is formulated mathematically and interpreted in Van Mieghem (2003b). That paper also gives the first evidence that optimal relative capacity imbalance increases in risk aversion, but less significantly so when correlation is high. Moreover, it shows that increasing risk aversion sometimes leads to an increase in some optimal capacity levels, which is not optimal in single-resource models. Finally, its analysis under risk aversion highlights the third concern, in addition to the two discussed above in §4.2, that should be considered when adopting a risk-neutral newsvendor capacity solution: The newsvendor solution $\mathbf{K}^*$ defines the maximal-risk extreme point of the efficient frontier of risk-return capacity configurations. Consider all possible capacity plans—that is any nonnegative capacity three-vector $\mathbf{K} \in \mathbb{R}^3_+$ in the example of the introduction—and plot its associated expected profit and variance. The result for the example is shown in Figure 6 using the capacity plans defined above in §4.2. By definition, the risk-neutral optimizing newsvendor-network capacity $\mathbf{K}^*$ has highest expected profit and, thus, is the right-end-point of the frontier. Hence, a risk-averse decision maker may find it attractive to deviate from the newsvendor solution to trade-off some risk for return.

Figure 6 Risk-Return Scatter Plot and Efficient Frontier for the Capacity Portfolio Investment Problem of the Example



Note. Each point corresponds to a specific capacity investment vector $\mathbf{K}$.

As shown, giving up a small percentage in returns can cut risk by an order of magnitude.

7. Concluding Remarks

What inferences can be made for future stochastic capacity-portfolio investment? From an application and organizational perspective, the requirements of stochastic capacity portfolio optimization on methods and systems for capacity planning are substantial. It requires that forecasting expands its point forecast with either supplemental scenarios, or with a covariance matrix. The coupled nature of the optimality equations suggest a coordinated, firm-level approach. It would be interesting to see whether decomposition approaches are feasible that would allow some decentralized decision making at the plant level.

Related, research is needed to model the fact that demand is partly endogenous and partly exogenous, as discussed above. Such models could assess the "efficiency loss" that results from adopting sales-plan driven capacity planning. The problem here would be to determine the "best" deterministic sales plan assuming an optimal mechanism design for the sales force and manufacturing incentives. An agency[19] problem is embedded: Sales and marketing may agree on a plan, but it typically does not reflect an unbiased assessment of what will be sold. Kouvelis and Lariviere (2000) may provide a starting point for such research. In addition, sales faces a very nonsymmetric penalty function with harsher punishment for falling short of the sales plan. Consequently, manufacturing does need to hedge its capacity portfolio somehow to respond to deviations from the sales plan. In short, it appears that the need for operational hedging by purposely unbalancing the capacity portfolio, as discussed in this paper, may be robust to this more realistic model formulation. Future research is needed to confirm this conjecture.

From a theoretical perspective, capacity portfolio research is still in its infancy and many models remain un- or under-explored, especially queuing formulations. Game-theoretic and risk-averse capacity portfolio analysis seem to rapidly gain in popularity. Much

[19] The unbalanced optimal newsvendor capacity portfolio seems to imply that a first-best deterministic sales plan does not exist.

remains to be done and many obvious questions are still unanswered. For example, it is unclear what intuition we can build to answer the obvious questions: Where should one put safety capacity in the processing network and how does that answer change when risk aversion increases? An obvious initial step is to consider these questions in the most simple capacity-portfolio problems. Eventually, analysis will develop intuition.

Fortunately and unfortunately, capacity-portfolio models rapidly become complex. Complexity is unfortunate because it often makes superior analytical solutions elusive. Thus, simulation-based optimization becomes the natural second-best option and is expected to increase in popularity. At the same time, complexity is fortunate as study is worthwhile with a potential impact on practice. Compared to the impact of financial portfolio analysis, even a fraction would be substantial.

Acknowledgments

The author is grateful to the former Editor-in-Chief Leroy B. Schwarz for encouraging him to start, improve, and finish (!) this review paper. The author also thanks the senior reviewers, Evan Porteus and Gerard Cachon, for many excellent suggestions. This paper also benefited from feedback by Philipp Afeche, Sunil Chopra, Jim Dai, Yi Ding, Avinash Dixit, Janice Eberly, Stephen Graves, Michael Harrison, Martin Lariviere, Armony Mor, Roger Myerson, Erica Plambeck, Suresh Sethi, and Costis Skiadas.

References

Abel, A. B., J. C. Eberly. 1998. The mix and scale of factors with irreversibility and fixed costs of investment. McCallum, Plosser, eds. *Carnegie-Rochester Conference Series on Public Policy*, Vol. 48. Elsevier Science, 101–135.

Allayannis, G., J. Ihrig, J. P. Weston. 2001. Exchange-rate hedging: Financial vs. operational strategies. *Amer. Econom. Rev.* **91**(2) 391–395.

Angelus, A., E. L. Porteus. 2002. Simultaneous capacity and production management of short-life-cycle, produce-to-stock goods under stochastic demand. *Management Sci.* **48**(3) 399–413.

Armony, M., E. L. Plambeck. 2002. The impact of duplicate orders on demand estimation and capacity investment. Technical report, Research Paper 1750, Graduate School of Business, Stanford University.

Arrow, K. J. 1968. Optimal capital policy with irreversible investment. J. N. Wolfe, ed. *Value, Capital and Growth. Papers in Honour of Sir John Hicks.* Edinburgh University Press, Edinburgh, 1–19.

Atamtürk, A., D. S. Hochbaum. 2001. Capacity acquisition, subcontracting, and lot sizing. *Management Sci.* **47**(8) 1081–1100.

Bashyam, T. C. A. 1996. Competitive capacity expansion under demand uncertainty. *Eur. J. Oper. Res.* **95** 89–114.

Bassok, Y., R. Anupindi, R. Akella. 1999. Single-period multiproduct inventory models with substitution. *Oper. Res.* **47**(2) 632–642.

Bean, J. C., J. L. Higle, R. L. Smith. 1992. Capacity expansion under stochastic demands. *Oper. Res.* **40**(Suppl. no. 2) S210–S216.

Bernstein, F., G. A. DeCroix. 2002. Decentralized pricing and capacity decisions in a multitier system with modular assembly. Technical report, Fuqua, Duke University, NC.

Birge, J. R. 2000. Option methods for incorporating risk into linear capacity planning models. *Manufacturing Service Oper. Management* **2**(1) 19–31.

Bish, E., Q. Wang. 2002. Optimal investment strategies for flexible resources: Considering pricing and correlated demands. Technical report, Virginia Polytechnic Institute and State University, VA.

Boyaci, T., S. Ray. 2003. Product differentiation and capacity cost interaction in time and price sensitive markets. *Manufacturing Service Oper. Management* **5**(1) 18–36.

Bradley, J. R., B. C. Arntzen. 1999. The simultaneous planning of production, capacity and inventory in seasonal demand environments. *Oper. Res.* **47**(6) 795–806.

——, P. W. Glynn. 2002. Managing capacity and inventory jointly in manufacturing systems. *Management Sci.* **48**(2) 273–288.

Burnetas, A., S. Gilbert. 2001. Future capacity procurements under unknown demand and increasing cost. *Management Sci.* **47**(7) 979–992.

Cachon, G. P., P. T. Harker. 2002. Competition and outsourcing with scale economies. *Management Sci.* **48**(10) 1314–1333.

——, M. A. Lariviere. 1999. Capacity choice and allocation: Strategic behavior and supply chain contracting. *Management Sci.* **45**(8) 1091–1108.

Çakanyildirim, M., R. O. Roundy. 2002. Optimal capacity expansion and contraction under demand uncertainty. Technical report, University of Texas at Dallas, Dallas, TX.

Caldentey, R., M. Haugh. 2003. Optimal control and hedging of operations in the presence of financial markets. Working paper, Stern School of Business, New York University, New York.

——, L. M. Wein. 2003. Analysis of a decentralized production-inventory system. *Manufacturing Service Oper. Management* **5**(1) 1–17.

Carr, S., W. Lovejoy. 2000. The inverse newsvendor problem: Choosing an optimal demand portfolio for capacitated resources. *Management Sci.* **46**(7) 912–927.

Chand, S., V. N. Hsu, S. Sethi. 2002. Forecast, solution, and rolling horizons in Harvard operations management problems: A classified bibliography. *Manufacturing Service Oper. Management* **4**(1) 25–43.

Chen, F., A. Federgruen. 2000. Mean-variance analysis of basic inventory models. Technical report, Graduate School of Business, Columbia University, New York.

Chod, J., N. Rudi, J. A. Van Mieghem. 2003. Financial hedging of stochastic capacity investment: Complementarity with operational flexibility. Working paper, Northwestern University, Evanston, IL.

Dai, J. G., J. H. Vande Vate. 2000. The stability of two-station multitype fluid networks. *Oper. Res.* **48**(5) 721–744.

Davis, M. H. A., M. A. H. Dempster, S. P. Sethi, D. Vermes. 1987. Optimal capacity expansion under uncertainty. *Adv. Appl. Probab.* **19** 156–176.

Ding, Q., P. Kouvelis. 2001. On the interaction of production and financial hedging decisions in global markets. Technical report, Washington University in St. Louis.

Dixit, A. K. 1995. Irreversible investment and scale economies. *J. Econom. Dynamics Control* **19** 327–350.

——, R. S. Pindyck. 1994. *Investment Under Uncertainty*. Princeton University Press, Princeton, NJ.

Eberly, J. C., J. A. Van Mieghem. 1997. Multi-factor dynamic investment under uncertainty. *J. Econom. Theory* **75**(2) 345–387.

Eppen, G. D., R. K. Martin, L. Schrage. 1989. A scenario approach to capacity planning. *Oper. Res.* **37**(4) 517–527.

Erlenkotter, D., S. Sethi, N. Okada. 1989. Planning for surprise: Water resources development under demand and supply uncertainty. I. The general model. *Management Sci.* **35**(2) 149–163.

Fine, C. H., R. M. Freund. 1990. Optimal investment in product-flexible manufacturing capacity. *Management Sci.* **36**(4) 449–466.

Freidenfelds, J. 1981. *Capacity Expansion: Analysis of Simple Models with Applications*. North-Holland, New York.

Gaur, V., S. Seshadri. 2002. Hedging inventory risk through market instruments. Technical report, New York University, New York.

Giglio, R. J. 1970. Stochastic capacity models. *Management Sci.* **17**(3) 174–184.

Graves, S. C. 2002. Manufacturing planning and control. P. Pardalos, M. Resende, eds. *Handbook of Applied Optimization*. Oxford University Press, NY, 728–746.

——, S. P. Willems. 2000. Optimizing strategic safety stock placement in supply chains. *Manufacturing Service Oper. Management* **2**(1) 68–83.

Hansen, L. P., T. J. Sargent. 2001. Robust control and model uncertainty. *Amer. Econom. Rev.* **91**(2) 60–65.

Harrison, J. M., D. M. Kreps. 1979. Martingales and arbitrage in multiperiod securities markets. *J. Econom. Theory* **20** 381–408.

——, J. A. Van Mieghem. 1999. Multi-resource investment strategies: Operational hedging under demand uncertainty. *Eur. J. Oper. Res.* **113**(1) 17–29.

He, H., R. S. Pindyck. 1992. Investments in flexible production capacity. *J. Econom. Dynamics Control* **16** 575–599.

Hiller, R. S., J. J. Shapiro. 1986. Optimal capacity expansion planning when there are learning effects. *Management Sci.* **32**(9) 1153–1163.

Hopp, W. J., M. L. Spearman. 1996. *Factory Physics*. Richard D. Irwin, Boston, MA.

Hu, W. T., R. O. Roundy. 2002. A continuous-time strategy capacity planning model. Technical report, ORIE, Cornell University, New York.

Hu, X., I. Duenyas, R. Kapuscinski. 2002. Advance demand information and safety capacity as a hedge against demand and capacity uncertainty. Technical report, University of Michigan, MI.

Huchzermeier, A., M. A. Cohen. 1996. Valuing operational flexibility under exchange rate risk. *Oper. Res.* **44**(1) 100–113.

Jordan, W. C., S. C. Graves. 1995. Principles on the benefits of manufacturing process flexibility. *Management Sci.* **41**(4) 577–594.

Kapuscinski, R., S. Tayur. 1998. Optimal policies and simulation-based optimization for multistage capacitated production inventory systems. M. Magazine, S. Tayur, R. Ganeshan, eds. *Quantitative Methods for Supply Chain Management*. Kluwer, Boston, MA.

Kouvelis, P., G. Gutierrez. 1997. The newsvendor problem in a global market: Optimal centralized and decentralized control policies for a two-market stochastic inventory system. *Management Sci.* **43**(5) 571–585.

——, M. Lariviere. 2000. Decentralizing cross-functional decisions: Coordination through internal markets. *Management Sci.* **46**(8) 1049–1058.

——, J. Milner. 2002. Supply chain capacity and outsourcing decisions: The dynamic interplay of demand and supply uncertainty. *IIE Trans.* **34**(8) 717–728.

Kreps, D. M. 1990. *A Course in Microeconomic Theory*. Princeton University Press, Princeton, NJ.

——, E. L. Porteus. 1978. Temporal resolution of uncertainty and dynamic choice theory. *Econometrica* **46** 185–200.

Kroll, Y., H. Levy, H. M. Markowitz. 1984. Mean-variance versus direct utility maximization. *J. Finance* **39**(1) 47–61.

Kulkarni, S. S., M. J. Magazine, A. S. Raturi. 2002. How does risk-pooling impact manufacturing network configuration? Working paper, University of Cincinnati, Cincinnati, OH.

Laguna, M. 1998. Applying robust optimization to capacity expansion of one location intelecommunications with demand uncertainty. *Management Sci.* **44**(11S) 101–110.

Lederer, P. J., L. Li. 1997. Pricing, production, scheduling and delivery-time competition. *Oper. Res.* **45**(3) 407–420.

Levy, H., H. M. Markowitz. 1979. Approximating expected utility by a function of mean and variance. *Amer. Econom. Rev.* **69** 308–317.

Li, S., D. Tirupati. 1994. Dynamic capacity expansion problem with multiple products: Technology selection and timing of capacity additions. *Oper. Res.* **42**(5) 958–976.

Lippman, S. A., K. F. McCardle. 1997. The competitive newsboy. *Oper. Res.* **45** 54–65.

Loch, C. H. 1991. Pricing in markets sensitive to delay. Ph.D. thesis, Stanford University, Stanford, CA.

Lovejoy, W. S., Y. Li. 2002. Hospital operating room capacity expansion. *Management Sci.* **48**(11) 1369–1387.

Luss, H. 1982. Operations research and capacity expansion problems: A survey. *Oper. Res.* **30** 907–947.

Malcolm, S. A., S. A. Zenios. 1994. Robust optimization for power systems capacity expansion under uncertainty. *J. Oper. Res. Soc.* **45**(9) 1040–1049.

Manne, A. S. 1961. Capacity expansion and probabilistic growth. *Econometrica* **29**(4) 632–649.

Markowitz, H. M. 1991. Foundations of portfolio theory. *J. Finance* **46**(2) 469–477.

Mendelson, H. 1985. Pricing computer services: Queueing effects. *Comm. ACM* **28**(3) 312–321.

Merriam-Webster's Collegiate Dictionary. 1998. 10th ed. Merriam-Webster, Springfield, MA.

Merton, R. 1971. Optimum consumption and portfolio rules in a continuous-time model. *J. Econom. Theory* **3** 373–413.

Nahmias, S. 1993. *Production and Operations Analysis*, 2nd ed. Richard D. Irwin, Boston, MA.

Narongwanich, W., I. Duenyas, J. R. Birge. 2002. Optimal portfolio of reconfigurable and dedicated capacity under uncertainty. Technical report, University of Michigan, MI.

Nawrocki, D. 1999. A brief history of downside risk measures. *J. Investing* **8**(3) 9–26.

Netessine, S., G. Dobson, R. A. Shumsky. 2002. Flexible service capacity: Optimal investment and the impact of demand correlation. *Oper. Res.* **50**(2) 375–388.

——, N. Rudi. 2003. Centralized and competitive inventory models with demand substitution. *Oper. Res.* **51**(2) 329–335.

Paraskevopoulos, D., E. Karakitsos, B. Rustem. 1991. Robust capacity planning under uncertainty. *Management Sci.* **37**(7) 787–800.

Pindyck, R. S., D. L. Rubinfeld. 1989. *Microeconomics*. Macmillan, New York.

Plambeck, E. L., T. Taylor. 2001. Sell the plant? The impact of contract manufacturing on innovation, capacity, and profitability. Technical report, Stanford Graduate School of Business, Stanford, CA.

Porteus, E. L., S. Whang. 1991. On manufacturing/marketing incentives. *Management Sci.* **37**(9) 1166–1181.

Rajagopalan, S. 1998. Capacity expansion and equipment replacement: A unified approach. *Oper. Res.* **46**(6) 846–857.

——, M. R. Singh, T. E. Morton. 1998. Capacity expansion and replacement in growing markets with uncertain technological breakthroughs. *Management Sci.* **44**(1) 12–30.

——, J. M. Swaminathan. 2001. A coordinated production planning model with capacity expansion and inventory management. *Management Sci.* **47**(11) 1562–1580.

Ryan, S. M. 2002. Capacity expansion for random exponential demand growth with lead times. Iowa State University. Working paper, 1–23, available at *www.public.iastate.edu/smryan/*.

——. 2003. Capacity expansion with lead times and correlated random demand. *Naval Res. Logist.* **50**(2) 167–183.

Schroder, M., C. Skiadas. 1999. Optimal consumption and portfolio selection with stochastic differential utility. *J. Econom. Theory* **89** 68–126.

Sethi, S. P., H. Yan, H. Zhang, Q. Zhang. 2002. Optimal and hierarchical controls in dynamic stochastic manufacturing systems: A survey. *Manufacturing Service Oper. Management* **4**(2) 133–170.

Stenbacka, R., M. Tombak. 2002. Investment, capital structure, and complementarities between debt and new equity. *Management Sci.* **48**(2) 257–272.

Van Mieghem, J. A. 1998a. Investment strategies for flexible resources. *Management Sci.* **44**(8) 1071–1078.

——. 1998b. Seagate technologies: Operational hedging. Kellogg School of Management Case Study. Northwestern University, Evanston, IL.

——. 1999. Coordinating investment, production and subcontracting. *Management Sci.* **45**(7) 954–971.

——. 2003a. Component commonality strategies: Value drivers and equivalence with flexible capacity. Technical report, Northwestern University, Evanston, IL. (Available at *www.kellogg.northwestern.edu/faculty/VanMieghem.*)

——. 2003b. Risk-averse newsvendor networks: Mean-variance analysis of operational hedging with capacity or inventory. Working paper, Northwestern University.

——, M. Dada. 1999. Price versus production postponement: Capacity and competition. *Management Sci.* **45**(12) 1631–1649.

——, N. Rudi. 2002. Newsvendor networks: Inventory management and capacity investment with discretionary activities. *Manufacturing Service Oper. Management* **4**(4) 313–335.

Received: July 11, 2002; days with author: 205; revisions: 2; average review cycle time: 30 days; Senior Editors: Evan Porteus and Gérard Cachon.