1-2009

Capacity expansion under a service-level constraint for uncertain demand with lead times

Rahul R. Marathe
Indian Institute of Technology Madras

Sarah M. Ryan
Iowa State University, smryan@iastate.edu

Capacity expansion under a service level constraint for uncertain demand with lead times

Rahul R. Marathe [*] Sarah M. Ryan [†‡]

June 4, 2008

Abstract

For a service provider facing stochastic demand growth, expansion lead times and economies of scale complicate the expansion timing and sizing decisions. We formulate a model to minimize the infinite horizon expected discounted expansion cost under a service level constraint. The service level is defined as the proportion of demand over an expansion cycle that is satisfied by available capacity. For demand that follows a geometric Brownian motion process, we impose a stationary policy under which expansions are triggered by a fixed ratio of demand to the capacity position, i.e., the capacity that will be available when any current expansion project is completed, and each expansion increases capacity by the same proportion. The risk of capacity shortage during a cycle is estimated analytically using the value of an up-and-out partial barrier call option. A cutting plane procedure identifies the optimal values of the two expansion policy parameters simultaneously. Numerical instances illustrate that if demand grows slowly with low volatility and the expansion lead times are short, then it is optimal to delay the start of expansion beyond when demand exceeds the capacity position. Delays in initiating expansions are coupled with larger expansion sizes.

1 Introduction

We consider a service provider that owns capacity and wishes to meet a specified level of service over a long time horizon as demand increases with increasing uncertainty. Taking capacity simply as the ability to provide service, we treat it as a single resource residing at a single location. The capacity added does not deteriorate; that is, once the capacity is installed, we assume that it is available infinitely. Economies of scale motivate adding capacity in discrete chunks rather than continuously over time. There is an economic tradeoff between the scale economies and discounting of future expansion costs: discounting works to delay expansions, while the economies of scale favor one large present expansion over a series of

[*]Department of Management Studies, Indian Institute of Technology Madras, Chennai, India 600 036.
[†]Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA 50011.
[‡]corresponding author. E-mail: smryan@iastate.edu, Phone: (515) 294 4347.

smaller ones. Significant lead times, or delays, exist between the time the expansion decision is made and the time when the added capacity is actually available to satisfy the demand. Coupled with uncertainty in the demand growth, these create the risk of failing to provide adequate service during the lead times. We assume the lead times are fixed so that the sole randomness in the model comes from the demand process. The goal is to find the timing and sizes of future capacity expansions that minimize the infinite horizon expected discounted cost of expansion while maintaining the specified service level.

We model demand for the capacity as a geometric Brownian motion (GBM) process. Previously, we analyzed historical usage data for some services in the U.S. and found some that were consistent with this assumption. For example, the total airline passenger enplanements in the USA from 1985 to 2001 were statistically consistent with discrete observations of a GBM process (Marathe and Ryan, 2005). Pilots are a critical resource for satisfying this demand. There is evidence of pilot shortages nationwide (Barker, 2000; Donohue, 2000; Hopkins, 2001). An airline facing pilot shortages and committed to meeting a specified proportion of demand faces essentially the capacity expansion problem we study in this paper. Here, the human resource capacity can be considered non-deteriorating if expanding by one represents creating a permanent position and its cost covers the stream of individuals who will occupy it. The lead time would include recruiting and training the initial occupant. Usage of electric power was also found to be consistent with the GBM process, and substantial lead times exist for increasing either generating or transmission capacity. We acknowledge the improbability of demand following a GBM process with fixed parameters over a very long period of time in practice; hence, the main contributions of this paper concern theoretical understanding of the capacity expansion policy and its sensitivity to the problem parameters.

The majority of the models in the past have concentrated on the scenario when the capacity expansion project starts before (or immediately when) the demand for the resource reaches the capacity position. By capacity position, we mean the capacity that will be available after any current expansion project is completed. However, we envision cases where the service provider may want to delay the start of expansion until after the demand reaches the capacity position. This delay could be justified by specific parameter values (such as demand growth rate, volatility, expansion lead time length, etc.) observed in particular industries. Allowing this delay expands the choices the service provider can make according to the tradeoffs between expansion cost and shortage risk. Because the economic penalty for not meeting the demand is difficult to quantify, we develop a service level constraint in terms of the proportion of demand over each expansion cycle that is satisfied with the available capacity. The paper jointly optimizes the timing and sizes of expansions to answer the question: Is it optimal to allow shortages before initiating expansion? The answer is a conditional yes. If the service provider is operating in a low growth industry

(characterized by smaller demand drift and/or volatility values), or when the service provider does not have to meet very tight service levels; or when the expansion projects can be completed fairly fast (smaller expansion lead times), it is in fact optimal to accumulate some shortages before initiating the next expansion project. Larger expansion sizes compensate for the delays, so under these circumstances the overall strategy can be characterized as delayed and infrequent.

The GBM model has been used before to represent future demand in capacity studies. Whitt (1981) studied capacity utilization over time assuming demand followed a GBM. An indirect validation of the assumption was provided by Lieberman (1989), who showed in an empirical study of the chemical industry that actual capacity utilization matched the predictions from Whitt's proposed policy. This paper is motivated chiefly by Ryan (2004), in which the model was developed and analyzed under the assumption that no shortages were allowed to accumulate before starting any expansion. Numerical instances suggested that this assumption could preclude the optimal solution in some cases. Pak et al. (2004) studied the effects of technological change in the same environment.

The capacity expansion problem in general has been very well researched. As van Meighem (2003) points out, there are over 15,000 articles with 'capacity' in the title or keyword. This paper follows in the stream of the seminal paper by Manne (1961), who proposed a model to decide the expansion sizes in cases where the demand follows a linear deterministic or random-walk pattern. He also considered the effects of economies of scale and penalties for demand not being satisfied. Smith (1979) analyzed the addition of capacity from a finite set of available possible additions for a case of exponential demand. A turnpike theorem was developed which gives the structural characteristics of the optimal policy. Whitt (1981) studied the capacity utilization aspect of the problem. Using results for stochastic clearing processes he obtained the stationary distribution function for utilization under a particular expansion policy when demand follows a GBM process. The long-term expected utilization depends on both the size and timing parameters, as will be discussed further below. Bean and Smith (1985) analyzed the effect of the study horizon length on the solution to the deterministic capacity expansion problem. They developed an algorithm to determine the length of the horizon needed to identify an optimal first facility to install. Buzacott and Chaouch (1988) examined the effects of demand plateaus assuming the demand to follow an alternating renewal process. Bean et al. (1994) considered a generalization of Brownian motion demand, where demand was assumed to be either a nonlinear Brownian motion process or a non-Markovian birth and death process. Like Manne (1961), they showed that the problem can be transformed into an equivalent deterministic problem and that the effect of uncertainty in the demand is to reduce the interest rate. He and Pindyck (1992) modeled the value of the firm when the demand for capacity followed a GBM process. By considering the values

of capacity already in place and the firm's growth option (i.e. the potential for additional profits from adding capacity later), they found the optimal capacity investment. The authors also considered flexible capacity – capacity that could be used to satisfy more than one type of demand.

The risk induced by incorrect demand forecasting was considered by Marin and Salmeron (2001), where the electric power generation capacity expansion problem for stochastic demand was formulated to minimize cost and risk. The risk function defined was the penalty cost arising from changes to be made in the generation plan owing to changes in the demand. The resulting large problem was solved using Benders decomposition. Kurabuk and Yu (2003) considered a two stage model for a semiconductor manufacturer, where strategic planning involves how much of which microelectronic technology to produce and also the location of that production. At the tactical stage, corrections and reconfigurations were done based on more accurate demand and capacity information. This model considered capacity as an uncertain element because of variability in the manufacturing processes of high technology products.

All of the preceding analyses relied on the absence of lead times to rule out unplanned shortages. Other related work explicitly considers lead times. Davis et al. (1987) considered a capacity expansion problem to find the optimal timing and sizes of future expansion where the demand was a random point process (that is, the demand increased by discrete amounts at random times). They considered lead time to depend on the rate of investment and applied stochastic control theory to find the optimal expansion policy. Chaouch and Buzacott (1994) examined the same problem as Buzacott and Chaouch (1988) including lead times and also considered the two cases, where the capacity addition started before and after the demand reached the current capacity, respectively. Assuming proportional shortage costs, they set up a total cost minimization problem resulting from an infinite horizon dynamic programming formulation. Cakanyildirim et al. (2004) studied a capacity expansion problem for the semiconductor industry where the machine purchase, floor space and shell expansions were jointly optimized over a finite time horizon. They considered deterministic lead times for the machine purchases. A special polynomial-time algorithm was developed to find the optimal machine purchase times and the optimal times and sizes for both floor and shell expansions.

This paper differs from the previous literature in the following ways. The demand process we consider is the same as in Whitt (1981) and He and Pindyck (1992); however, those models did not include expansion lead times. We assume a variation on the stationary timing and size policy for capacity expansion proposed by Whitt (1981). While Ryan (2004) assumed that the next expansion starts before the current capacity position is reached, in this paper we also consider the possibility where the next expansion is started after demand exceeds the capacity position. We observe numerically some conditions under which the

policy of accumulating shortages before is optimal. Manne (1961) considered allowing initial shortages for linear demand growth with no expansion lead times. Chaouch and Buzacott (1994) considered starting the expansion either before or after the demand crosses the capacity position for linear demand demand punctuated by plateaus. Also, they employed a proportional penalty cost for not meeting the demand, whereas we define a service level constraint.

Similar to Ryan (2004), we use financial mathematics to estimate potential capacity shortages. Application of financial options theory to operational problems is a relatively new field of research. Birge (2000) applied the basic principles of risk-neutral valuation to general forms of constrained resource problems, such as capacity planning, and pointed out the correspondence between undercapacity and a call option. In our model, the potential for capacity shortages can be compared to the barrier options in finance – in particular, the up-and-out call option. As defined by Musiela and Rutkowski (1997), the generic term barrier options refers to the class of options whose payoff depends on whether or not the underlying prices hits a pre-specified barrier during the option's life. The idea of these options was discussed as early as the 1970's by Merton (1973) and Goldman et al. (1979), who analyzed "path dependent options." Rubinstein and Reiner (1991) and Rubinstein (1991) arrived at analytical formulas for various types of barrier options as a limiting case of a discrete time model. The price of the barrier option was found from the joint distribution of Brownian motion and its maximum in Chuang (1996). First, the joint distribution of the Brownian motion and its maximum was found for when the time intervals considered for the Brownian motion and its maximum are different. This result was then used to find the price of "partial barrier options" (Musiela and Rutkowski, 1997) – that is, barrier options in which the underlying price is monitored for barrier hits only during a prespecified portion of the option's lifetime. Similar results about the partial barrier option were obtained by Heynen and Kat (1997). They gave analytical expressions for all cases of barrier options viz. cash or nothing, asset or nothing, etc.

We solve the problem to optimize capacity expansion cost under the service level constraint by using a cutting plane method. These methods have been proposed for solving complex optimization problems (see Gomory (1963), Kelley (1960), Wolfe (1961), Zangwill (1969) for their development). More recently, Atlason et al. (2004) used a cutting plane algorithm to solve a call center staffing problem under a service level constraint. The service level expression in their model was evaluated by simulation of the call data whereas we evaluate the constraint analytically. Because the complexity of our analytical formula for constraint violation has complexity roughly similar to simulation, the cutting plane approach is well suited for both cases.

The first contribution of this paper is to generalize the model of Ryan (2004) such that the service provider could start each expansion either *before* or *after* the demand has crossed the capacity position, by balancing the total cost incurred against the service level achieved. We postulate that there could be situations where the service provider would prefer to accumulate certain shortages before starting the next capacity addition. We find the optimal level of either excess or shortage relative to capacity position that should trigger a new capacity addition, and also the amount of new capacity to be added. Secondly, the service level of Ryan (2004) defines capacity shortages per unit of capacity, whereas we define it over total demand during the expansion cycle, and we remove the previous conservative bias in its estimation. As a result, unlike in the previous model where the service level expression was dependent only on the timing of expansions, the new constraint expression involves both timing and size of expansion. Hence, the main analytical contribution of our work is the closed-form expression for the service level constraint such that problem can be solved optimally as a single non-linear program for either the excess or shortage cases. Both the timing and size decision aspects are optimized jointly rather than sequentially as in Ryan (2004). The reformulated service level is evaluated in terms of a partial up-and-out barrier call option rather than a simple European option. Analysis of the cost function shows that the two decision variables are complementary in most cases, and numerical results confirm that delays in initiating expansion are coupled with large expansion sizes. The optimization problem to minimize cost subject to the service level constraint is solved using a cutting plane algorithm. Numerical instances reveal how these two dimensions of the expansion policy interact and respond to parameter changes, and the optimal multiplier for constraint violation quantifies the economic impact of the service level constraint.

The paper is organized as follows: we discuss our capacity expansion model in Section 2. Here we define all the terms and conditions applicable to our model. We mathematically analyze the model in Section 3, where we formulate the service level constraint and the objective function in terms of two decision variables. In Section 4 we discuss the numerical method used to solve the optimization problem and also discuss the numerical results and the effects of various model parameters on the capacity expansion decision. Concluding remarks and future directions in Section 5 complete the paper.

2 Model

As our model is similar to Ryan (2004), we use consistent notation. Let $B(t)$ be a Brownian motion having drift μ and volatility σ^2 with $B(0) = 0$. The demand for the service is given by a GBM process $P(t) = P(0)e^{B(t)}$ with a cumulative distribution function $F(\cdot)$. This implies that, for any values of k and t, the ratio $\frac{P(t+k)}{P(t)}$ is a random quantity independent of all the values up to t; and its logarithm, $\ln \frac{P(t+k)}{P(t)}$

has a normal distribution with mean μk and volatility $\sigma^2 k$. Hence, given $P(t)$, the logarithmic growth over a small period of time Δt is given by $\ln \frac{P(t+\Delta t)}{P(t)} = \mu \Delta t + \sigma \sqrt{\Delta t} Z$, where Z is a standard normal random variable. We define $\gamma \equiv \mu + \frac{\sigma^2}{2}$.

The assumption of a GBM process for demand may be reasonable in cases where

- the demand growth during a period, as a percentage of total demand, has a lognormal distribution that is stationary over time, and
- these successive growth percentages are independent.

Where past measurements of demand are available, common statistical tests can be used to verify the former condition, and Ross (1999) outlined a simple procedure to test the latter. Marathe and Ryan (2005) found that historical usage of electric power and airline travel over decades met both conditions after seasonal effects were removed. On the other hand, although data availability limited the statistical tests that could be applied, the conditions were not met by time series that could serve as proxies for the demand for Internet and mobile telephone service due to their declining growth rates. While the expectation of a constant exponential rate of demand growth continuing indefinitely may not be realistic, this assumption allows the closed form evaluation of the service level constraint by analogy with a barrier option.

We assume that capacity additions occur at discrete time points and that a fixed lead time of L time units is required to install new capacity. The problem is to choose a sequence $\{(T_n, X_n), n \geq 1\}$ where the time at which n^{th} capacity expansion starts, T_n is a stopping time with respect to the Brownian motion $B(t)$ and X_n is the n^{th} increase in capacity. For any realization ω of the Brownian motion $B(t)$ let $t_n \equiv T_n(\omega)$. Let K_n be the installed capacity after n capacity additions are completed, where K_0 is the initial capacity. Then,

$$K_n = K_0 + \sum_{i=1}^{n} X_i.$$

The installed capacity at time t is given by:

$$K(t) = \begin{cases} K_0, & 0 \leq t < t_1 + L \\ K_n, & t_n + L \leq t < t_{n+1} + L \end{cases}$$

The capacity position at time t is given by:

$$\Pi(t) = \begin{cases} K_0, & 0 \leq t < t_1 \\ K_n, & t_n \leq t < t_{n+1} \end{cases}$$

We assume an economies of scale regime, under which the cost of installing capacity of size X is given by:

$$C(X) = kX^a, \tag{1}$$

where k is a constant and $a < 1$ is the economies of scale parameter. Costs are discounted continuously at rate $r > \gamma$.

We assume that the policy proposed by Whitt and Luss for the same demand function is modified to account for the lead times and its parameter is adjusted to allow planned shortages to occur. Whitt (1981) showed that, without lead times, their policy results in a stationary distribution for the capacity utilization and provided a simple formula for its expected value. In the Whitt-Luss policy, each new expansion occurs when demand reaches some fixed proportion (< 1) of current capacity, and after its instantaneous addition, the new capacity is a constant multiple of its previous value. Ryan (2004) showed that for $p < 1$ with lead times, the Black-Scholes option pricing formula could evaluate the expected shortages. Moreover, under this timing policy, the proportional increment policy (where the optimal expansion size is a fixed proportion of the capacity position) is optimal. Because the goal of our model is to extend the Ryan (2004) model to $p \geq 1$, we assume the same stationary timing and size policy. Although it has not been formally proven optimal, the policy is easily implementable and plausible. The expected amount of constraint violation in each expansion cycle is stationary and the infinite horizon expected discounted expansion cost reduces to a geometric series. Hence, in this paper, we assume that each expansion is initiated when demand reaches some fixed proportion, p, of the capacity position, where p may be less than, equal to, or greater than one. That is, for $n \geq 1$, $T_n = \min\{t \geq 0 : P(t) = pK_{n-1}\}$. Note that for a positive drift ($\mu > 0$), $T_n < \infty$ with probability one (Karlin and Taylor, 1975). The sequence of capacity levels follows $K_n = vK_{n-1}, n \geq 1$, where $v \geq 1$.

Figures 1 and 2 illustrate the policy and potential shortages seen at the realized time t_n, when demand first reaches the level pK_{n-1}, $p > 1$. The first decision variable, p, quantifies the level of shortage that triggers an expansion. The n^{th} capacity expansion has just started. Upon its completion, the total installed capacity will reach the level K_n after the lead time L. The second decision variable is the size of each expansion $v \equiv \frac{K_n}{K_{n-1}}$. The next expansion will start at the time when the demand $P(t)$ first reaches

the new position pK_n. This random variable T_{n+1} could be greater than $t_n + L$ as in Figure 1 or less than $t_n + L$, causing the successive lead times to overlap as in Figure 2.

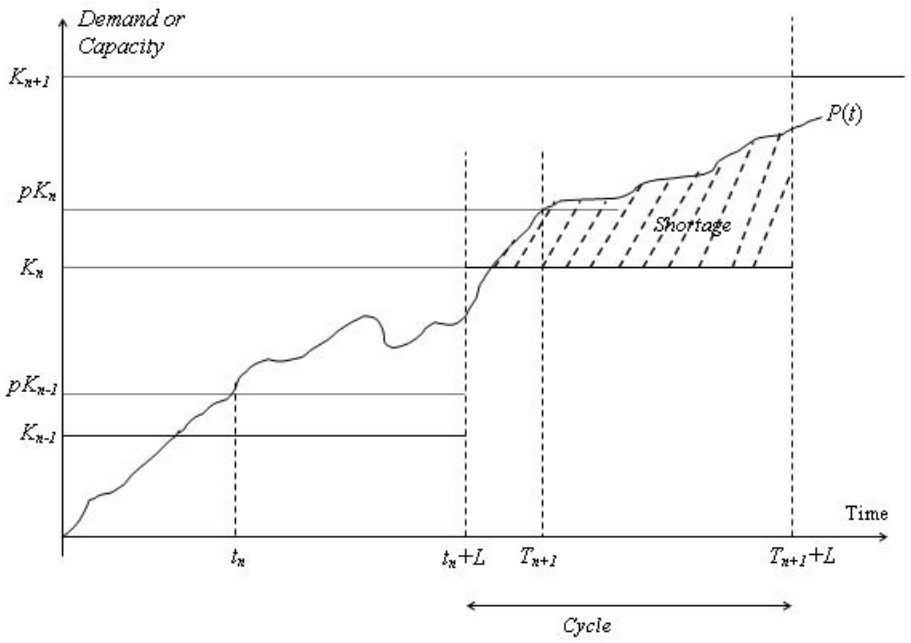

Figure 1: Capacity expansion policy and potential shortage when the next expansion starts after the end of the current expansion cycle

In the case illustrated in Figure 2 where the $(n+1)^{st}$ expansion is initiated before the n^{th} expansion project is finished, we wish to avoid double-counting the capacity shortages. We define a capacity cycle as the interval between the end of the current expansion project $t_n + L$ and the end of the next expansion project $T_{n+1} + L$, during which the actual capacity is K_n. Note that the cycle may be shorter than the lead time. Although not illustrated, the same two possibilities exist for the case when $p < 1$, i.e., the expansion is initiated *before* the demand reaches the current capacity position.

2.1 Formulation of the service level expression

For a generic capacity cycle, we formulate a service measure akin to the fill rate used in periodic (Sobel, 2004) and continuous review (Hadley and Whitin, 1963; Klemm, 1971) inventory models. At a generic expansion epoch t_n, the decision maker knows the demand, $P(t_n) = pK_{n-1}$ and estimates shortages during the interval $[t_n + L, T_{n+1} + L)$ with uncertain endpoint. In the inventory management literature, three different definitions of service levels, viz. α, β, γ, are used in different situations. The α measure, defined as probability of not being out of stock, does not reveal how much of the demand is actually satisfied. Schneider (1981) defines the β service level as the fraction of demand that is satisfied. Lastly, the γ service level is defined in terms of *cumulative* unsatisfied demand, which makes it applicable only

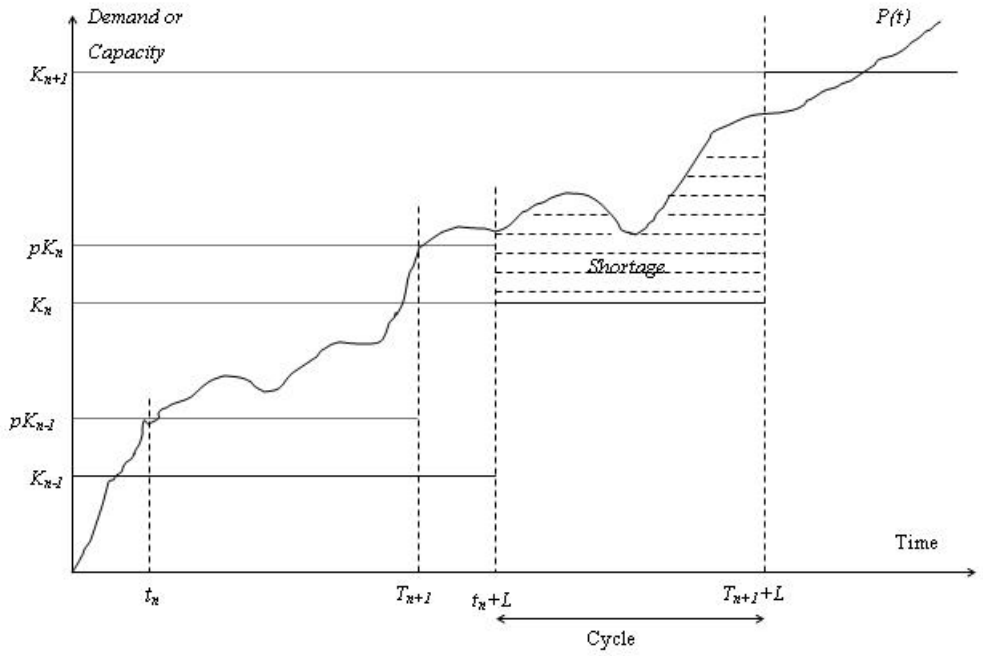

Figure 2: Capacity expansion policy and potential shortage when the next expansion starts before the end of the current expansion cycle

in case of backorders. Because services generally cannot be backordered, our constraint is based on the β definition. In the capacity expansion problem, the proportion of demand that is satisfied during the n^{th} cycle is:

$$\beta_n = \frac{\int_{t_n+L}^{T_{n+1}+L} \min[P(t), K_n]dt}{\int_{t_n+L}^{T_{n+1}+L} P(t)dt} = 1 - \frac{\int_{t_n+L}^{T_{n+1}+L} \max[P(t) - K_n, 0]dt}{\int_{t_n+L}^{T_{n+1}+L} P(t)dt}, \tag{2}$$

which is a random quantity. To guarantee a service level of at least $1 - \delta$, where δ is a small positive number, define the violation random variable as:

$$\xi_n \equiv \int_{t_n+L}^{T_{n+1}+L} e^{-r(t-t_n)} \max[P(t) - K_n, 0]dt - \delta \int_{t_n+L}^{T_{n+1}+L} e^{-r(t-t_n)} P(t)dt \tag{3}$$

Below, we show that the service level constraint, expressed as $E[\xi_n] \leq 0$, can be evaluated independently of n.

2.2 Optimization problem

We assume that the total cost of expansion is incurred at the beginning of the expansion project. Because the expansion lead time is fixed, there is no loss of generality.

Let $T(x)$ be the time when the demand first hits the level x. Under the assumed policy, the size of the n^{th} expansion is $X_n = K_n - K_{n-1}$ and it is initiated at the time $T(pK_{n-1})$, where $K_n = v^n K_0$. The

expected total cost, discounted to time 0, is given by

$$f(p,v) \equiv \sum_{n=1}^{\infty} E\left[e^{-rT(pv^{n-1}K_0)}\right] k\left(v^{n-1}(v-1)K_0\right)^a .$$

The problem is to minimize $f(p,v)$ subject to $E[\xi_n] \leq 0$ for all $n \geq 1$.

3 Mathematical Analysis

Having explained the model environment and discussed the policy parameters, we now analyze the mathematical model in detail. We use results from financial option pricing theory – particularly, the up-and-out barrier call option price formula – to evaluate the expected value of the service level violation (3) analytically. The infinite horizon expansion cost is obtained in closed form expression in terms of the policy parameters.

3.1 Analysis of the service level constraint

Evaluation of the expected constraint violation $E[\xi_n]$ is complicated by the nonlinear function of the stochastic process $P(t)$ in the first integral and the random upper limits for both integrals in Equation (3). Ryan (2004) estimated the future shortages, $\max[P(t) - K_n, 0], t > t_n$, by drawing an analogy with option pricing that allowed a straightforward application of the Black-Scholes valuation formula.

Under the assumption that $p < 1$, shortages can occur only during the lead time. This recognition allowed Ryan (2004) to integrate over an interval of fixed length L rather than the random length interval in Equation (3), at the expense of overestimating total shortages for cycles n such that $T_{n+1} < t_n + L$. In effect, the constraint as evaluated by this method had a conservative bias. Moreover, because the shortages were measured in relation to capacity rather than demand as in Equation (2), the resulting expression depended only on the timing parameter p and not on the size parameter v. In the following we employ methods of valuing exotic options to evaluate the expected shortage over an arbitrary cycle exactly in terms of p and v.

An up-and-out barrier option expires when the asset price first reaches a specified level from below; it is termed a partial barrier option if expiration results only from crossing the barrier before a time limit earlier than the expiration date. The seller of such an option is exposed to less risk (compared to an ordinary option) from a rapidly increasing asset value. Similarly, our decision maker at time t_n is liable for shortages within the time interval $[t_n + L, T_{n+1} + L)$ only. An increase in demand to the level pK_n causes the shortage option in the current cycle to expire by triggering the start of the next expansion

cycle. Applying the barrier option value to the capacity shortage provides a mechanism to handle the random upper limit for the integrals in Equation (3), but also represents the limited liability assigned to the decision at time t_n in view of later expansions.

The expected amount of constraint violation in Equation (3) is:

$$E[\xi_n] = E\left[\int_{t_n+L}^{T_{n+1}+L} e^{-r(t-t_n)} \max[P(t) - K_n, 0]dt\right] - \delta E\left[\int_{t_n+L}^{T_{n+1}+L} e^{-r(t-t_n)} P(t)dt\right], \tag{4}$$

where the expectations are taken at time t_n.

We now simplify each of the terms on right hand side of Equation (4) to obtain the constraint expression in terms of the decision variables, p and v. The first of these terms is equal to:

$$I_n^1 = E_{t_n}\left[\int_{t_n+L}^{T_{n+1}+L} e^{-r(t-t_n)}[P(t) - K_n]1(P(t) \geq K_n)dt\right] \tag{5}$$

where $1(x)$ is an indicator function such that $1(x) = 1$ if x is true and 0 otherwise. The upper limit $T_{n+1} + L$ is a random term because T_{n+1} is the time (unknown at time t_n) at which the demand will hit the value of pK_n for the first time.

To obtain deterministic integration limits in I_n^1, we introduce another indicator function $1(t \leq T_{n+1}+L)$ and remove the upper limit of integration. This step is justified because:

$$t \leq T_{n+1} + L \Leftrightarrow t - L \leq \min[t \geq 0 : P(t) = pK_n] \Leftrightarrow \max P(s) \leq pK_n, \forall s \leq t - L.$$

Therefore, $1(t \leq T_{n+1} + L) = 1(\max P(s) \leq pK_n, \forall s \leq t - L)$, and

$$I_n^1 = \int_{t_n+L}^{\infty} e^{-r(t-t_n)} E_{t_n}\left[(P(t) - K_n)1(P(t) \geq K_n) \times 1(\max P(s) \leq pK_n : 0 \leq s \leq t - L)\right] dt. \tag{6}$$

Next, given knowledge of events up to time t_n, using the Markov property we can shift the origin to time t_n and find the expected value in terms of a translated Brownian motion. In terms of the underlying standard Brownian motion, Equation (6) is equivalent to:

$$I_n^1 = \int_{t_n+L}^{\infty} e^{-r(t-t_n)} \quad E_{t_n}\left[[P(0)e^{B(t)} - K_n]1\left(B(t - t_n + t_n) \geq \ln \tfrac{K_n}{P(0)}\right)\right.$$
$$\left. 1\left(\max B(s) \leq \ln \tfrac{pK_n}{P(0)} : 0 \leq s \leq t - L\right)\right] dt. \tag{7}$$

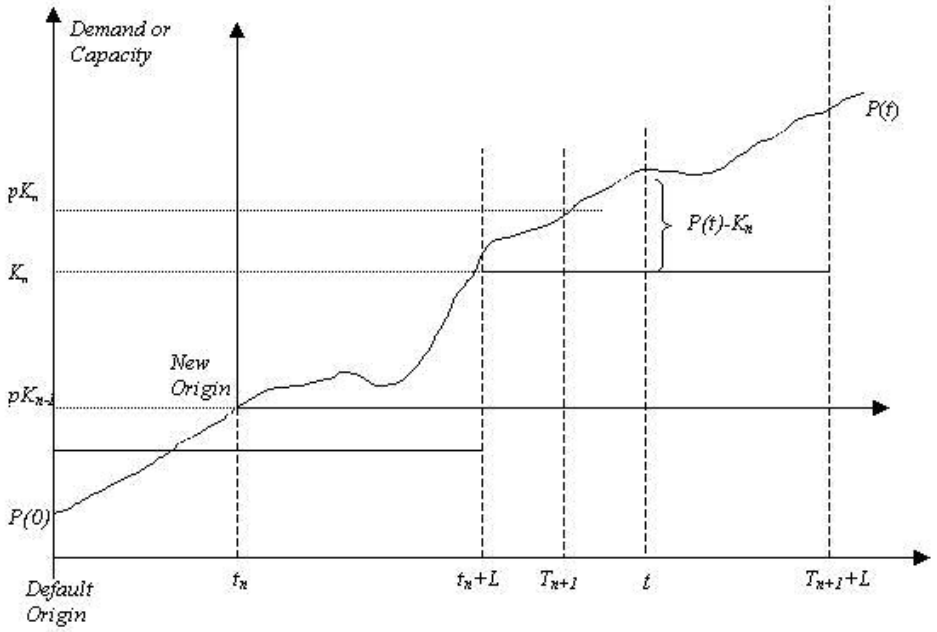

Figure 3: Important time instances of the $P(t)$ process

Define a new Brownian motion $W_n(t) \equiv B(t+t_n) - B(t_n)$, which has the same drift and volatility as the $B(t)$ process (Karlin and Taylor, 1975). In terms of this new process, Equation (7) becomes:

$$I_n^1 = \int_{t_n+L}^{\infty} e^{-r(t-t_n)} E_{t_n} \left[[P(t_n)e^{W_n(t-t_n)} - K_n] 1\left(W_n(t-t_n) \geq \ln \frac{K_n}{P(0)} - B(t_n) \right) \right.$$
$$\left. 1\left(\max W_n(k) \leq \ln \frac{pK_n}{P(0)} - B(t_n) : 0 \leq k \leq t-L-t_n \right) \right] dt,$$

because $P(t) = P(0)e^{B(t)} = P(t_n)e^{B(t)-B(t_n)}$ has the same distribution as $P(t_n)e^{W_n(t-t_n)}$ given $P(t_n)$.

Define $Q_n(t) \equiv P(t_n)e^{W_n(t)}$ as a GBM with respect to the Brownian motion $W_n(t)$. Also, we define a new variable, $u \equiv t - t_n$, and finally Equation (7) becomes:

$$I_n^1 = \int_L^{\infty} e^{-ru} E\left[(Q_n(u) - K_n) 1(Q_n(u) \geq K_n) 1(\max Q_n(s) \leq pK_n : 0 \leq s \leq u-L) \right] du. \tag{8}$$

The integral in this equation can be evaluated by simplifying the joint probability of the Brownian motion and its maximum over different time periods. Chuang (1996) first presented this joint probability distribution, and then used it to value knock-out barrier options, particularly the down-and-out call option. Appropriate changes could be made for the up-and-out call option. The expression for partial up-and-out barrier call option is discussed in Appendix A.

The expansion policy specifies $v = \frac{K_n}{K_{n-1}}$ as the ratio of successive capacity levels. We also know that $P(t_n) = pK_{n-1}$, because t_n is the expansion epoch determined by demand reaching the level of pK_{n-1}. Now exploiting the correspondence between Equations (8) and (17), we have:

$$\begin{aligned}
I_n^1 &= K_n \int_L^\infty e^{-ru} \left[e^{\gamma u} \left(\frac{p}{v}\right) \Psi \left(\frac{-\ln\left(\frac{v}{p}\right) + \left(\gamma + \frac{\sigma^2}{2}\right) u}{\sigma\sqrt{u}}, \frac{\ln v + \left(\gamma + \frac{\sigma^2}{2}\right)(u-L)}{\sigma\sqrt{(u-L)}}, -\sqrt{\frac{u-L}{u}} \right) \right. \\
&- v^{\frac{2\gamma}{\sigma^2}+1} e^{\gamma u} \left(\frac{p}{v}\right) \Psi \left(\frac{-\ln\left(\frac{v}{p}\right) + 2\ln v + \left(\gamma + \frac{\sigma^2}{2}\right) u}{\sigma\sqrt{u}}, \frac{-\ln v + \left(\gamma + \frac{\sigma^2}{2}\right)(u-L)}{\sigma\sqrt{(u-L)}}, -\sqrt{\frac{u-L}{u}} \right) \\
&- \Psi \left(\frac{-\ln\left(\frac{v}{p}\right) + \left(\gamma - \frac{\sigma^2}{2}\right) u}{\sigma\sqrt{u}}, \frac{\ln v - \left(\gamma - \frac{\sigma^2}{2}\right)(u-L)}{\sigma\sqrt{(u-L)}}, -\sqrt{\frac{u-L}{u}} \right) \\
&- \left. v^{\frac{2\gamma}{\sigma^2}-1} \Psi \left(\frac{-\ln\left(\frac{v}{p}\right) + 2\ln v + \left(\gamma - \frac{\sigma^2}{2}\right) u}{\sigma\sqrt{u}}, \frac{-\ln v - \left(\gamma - \frac{\sigma^2}{2}\right)(u-L)}{\sigma\sqrt{(u-L)}}, -\sqrt{\frac{u-L}{u}} \right) \right] du. \qquad (9)
\end{aligned}$$

We evaluate the second term on the right hand side of Equation (4) in the same way. Let

$$I_n^2 = E\left[\int_{t_n+L}^{T_{n+1}+L} e^{-r(t-t_n)} P(t) dt \right].$$

After steps similar to those for I_n^1 we have,

$$I_n^2 = \int_L^\infty e^{-ru} E\left[Q_n(u) 1(\max Q_n(s) \le pK_n : 0 \le s \le u - L) \right] du.$$

Once again, using Heynen and Kat (1997), this expression is equal to:

$$\begin{aligned}
I_n^2 &= K_n \int_L^\infty e^{(\gamma - r)u} \left(\frac{p}{v}\right) \Phi \left(\frac{\ln(v) - \left(\gamma + \frac{\sigma^2}{2}\right)(u-L)}{\sigma\sqrt{u-L}} \right) du \\
&- K_n \int_L^\infty e^{(\gamma - r)u} v^{\frac{2\gamma}{\sigma^2}+1} \left(\frac{p}{v}\right) \Phi \left(\frac{-\ln(v) - \left(\gamma + \frac{\sigma^2}{2}\right)(u-L)}{\sigma\sqrt{u-L}} \right) du \qquad (10)
\end{aligned}$$

where $\Phi(\cdot)$ is the standard normal distribution function.

The final service level constraint in Equation (4) can now be expressed as $E[\xi_n] = I_n^1 - \delta I_n^2 \le 0$, which is true if and only if

$$g(p, v) \equiv \frac{I_n^1 - \delta I_n^2}{K_n} \le 0, \qquad (11)$$

where the terms in the numerator are evaluated using Equations (9) and (10), respectively. By dividing the constraint violation during a cycle by the actual capacity, $g(p, v)$ is obtained in units of time.

After dividing out K_n in Equation (11), the expected constraint violation does not depend on n, because I_n^1 and I_n^2 depend on the cycle only through K_n. Hence, the expression for the service level constraint is the same for all the expansion cycles. The expected shortage in our model depends on both the decision variables. In fact, this is consistent with the expression for the capacity utilization without lead times Whitt (1981), which involves both the timing and the size parameter of the expansion policy.

3.2 Infinite horizon expansion cost

Within the feasible region, the optimal values of the policy parameters p and v are those that minimize the capacity expansion cost. The objective function represents the infinite horizon discounted cost of expansion for installing these capacity units at future time instances. From Karlin and Taylor (1975),

$$E\left[e^{-rT(pv^{n-1}K_0)}\right] = \left(\frac{P(0)}{pv^{n-1}K_0}\right)^{\lambda}, \quad \text{where } \lambda = \sqrt{\frac{\mu^2}{\sigma^4} + \frac{2r}{\sigma^2}} - \frac{\mu}{\sigma^2}, \tag{12}$$

and $\lambda > 1$ because $r > \gamma$. Therefore, as in Pak et al. (2004),

$$f(p,v) = k\left((v-1)K_0\right)^a \left(\frac{P(0)}{pK_0}\right)^{\lambda} \sum_{n=1}^{\infty} \left(v^{a-\lambda}\right)^{n-1} = \frac{k'(v-1)^a p^{-\lambda}}{1-v^{a-\lambda}}, \tag{13}$$

where $k' \equiv kK_0^{a-\lambda}(P(0))^{\lambda}$, because $a < 1 < \lambda$.

This expression forms our objective function for the non-linear program in terms of the policy parameters, p and v. Therefore, the optimization problem under the assumed capacity expansion policy is:

$$\begin{aligned} &\min f(p,v) \\ \text{subject to:} \\ &g(p,v) \leq 0 \\ &v \geq 1; p \geq 0. \end{aligned} \tag{14}$$

Here, the expression for the objective function is obtained from Equation (13) and the service level constraint expression is obtained from Equation (11). The constraint $p \geq 0$ includes both $p > 1$, in which case the service provider has to start the next expansion project with some initial shortages; and $p \leq 1$, meaning the next expansion project is started no later than when the demand hits the current capacity position as in Ryan (2004). Also, since we are considering only expansions, we restrict values of the size parameter to $v \geq 1$.

3.3 Qualitative analysis

If the objective function of a minimization problem is submodular then its decision variables are complementary (Amir, 2003). The function $f(p, v)$ is submodular if $\frac{\partial^2 f(p,v)}{\partial p \partial v} \leq 0$. From Equation (13), we have:

$$\frac{\partial^2 f(p,v)}{\partial p \partial v} = k'^{-1} \frac{\partial^2 f}{\partial p \partial v} = -\lambda p^{-\lambda-1} \left[\frac{a(v-1)^{a-1}(1-v^{a-\lambda}) - (\lambda - a)(v-1)^a v^{a-\lambda-1}}{(1-v^{a-\lambda})^2}, \right] \tag{15}$$

which is nonnegative if

$$\lambda(v-1) + a \leq a v^{\lambda - a + 1}. \tag{16}$$

The inequality holds for v sufficiently large. For reasonable values of the problem parameters, v slightly larger than one is sufficient. Therefore, we expect p^* and v^* to move in the same direction as the problem parameters vary. A firm can adopt either a strategy of delaying expansion until significant shortage occurs and compensating with large expansions, or one of small expansions, each of which is initiated when little or no shortage relative to capacity position has occurred. We term the former strategy "delayed and infrequent" and the latter strategy "proactive and frequent." This complementarity is consistent with results of the model in Manne (1961), where demand followed a linear Brownian motion, but not with those of Chaouch and Buzacott (1994) with demand plateaus. Attempts at analytically establishing monotone comparative statics, which depend on the sign of $\frac{\partial^2 f(p,v;s)}{\partial u \partial s}$, where $u = p$ or v and s is a parameter such as μ or σ, were unsuccessful. Hence, we draw further managerial insights from numerical analysis.

The next section describes a numerical method to solve this problem along with its results. Prior to computation, we can predict the qualitative behavior of the objective and constraint functions with respect to p and v. In the n^{th} cycle, for a fixed value of v, a small value of p reduces the time when demand is likely to cross the pK_n barrier that prompts an expansion. This limits the shortage risk but also increases the discounted cost by shortening the time interval to the next expansion. It is clear from Equation (13) that increasing p lowers the cost. For p fixed, a higher value of v reduces the shortage risk by starting each cycle with more excess capacity. It also lengthens the cycle, delaying the next expansion at the expense of a large present expansion cost. The optimal v for fixed p achieves a tradeoff between cost discounting and economies of scale.

4 Solution Methodology and Numerical Results

Equation (14) mathematically states the capacity expansion problem. Its complexity stems from the non-linear objective function and difficult constraint expression. In this section, we discuss the solution methodology used to find optimal values of the decision variables. Also, we numerically solve this

optimization problem under various instances of the problem parameters and discuss the results.

4.1 Optimization technique: Cutting plane method

We use the well-known cutting plane algorithm to solve the optimization problem in Equation (14). As seen from Equation (11), the constraint equation involves integrals of bivariate normal distribution functions, and hence, finding the partial derivatives of the constraint equation is difficult. Because the gradient of the constraint equation cannot be found readily, the usual gradient-based optimization methods cannot be used. The Lagrangean dual of the original optimization problem of Equation (14) is impractical because of the complexity of the constraint equation. Hence, to approximate the Lagrangean dual problem, we use the cutting plane algorithm (which Zangwill (1969) calls the 'dual cutting plane algorithm'), which bypasses finding feasible directions at each step of the problem (Bazaraa et al., 1993). A numerical method of checking for convexity is used to verify the conditions for the algorithm's convergence. The details of that numerical method can be found in Appendix B.

Following Bazaraa et al. (1993), the steps of the dual cutting plane algorithm as it applies to our problem are shown in Figure 4. The Master Problem is a linear program, the solution for which gives an upper bound for the solution to the Sub Problems. Because the Sub Problem constraints are linear with non-linear objective function, they are solved much faster than the original problem. Zangwill (1969) provided a proof of the finite convergence of the cutting plane algorithm, which means that the optimal solution to the original problem in Equation (14) will eventually be found, provided that the problem is feasible.

The cutting plane algorithm used here approximates the dual of the original capacity expansion problem. Therefore, upon solving the Master Problem, the value of the decision variable u we obtain is the Lagrangean multiplier of the shortage violation constraint. This optimal value of the variable can be considered as the penalty (in terms of dollars per unit time) of not meeting the demand at the specified service level.

4.2 Numerical results

We applied the dual cutting plane algorithm to problem (14). The (hypothetical) default parameter values used were: drift (μ) = 2%, volatility (σ) = 20%, lead time (L) = 2 years, interest rate (r) = 13% and economies of scale parameter (a) = 0.99. The required service level was assumed to be 95%, meaning that the shortages were limited to $\delta = 5\%$ of the total demand during the expansion cycle. When solving the Sub Problems of Figure 4, we added a dummy constraint of $p \leq 2$ to reduce the number of iterations

Initialization step: Select an initial feasible point (p_0, v_0).
For each iteration k, solve the Master Problem for z and u, which is given as:

$$\begin{aligned} & \max \quad z \\ \text{subject to} \quad & z \leq f(p_j, v_j) + ug(p_j, v_j) \quad \text{for} \quad j = 0 \cdots k-1 \\ & u \geq 0. \end{aligned}$$

Let (z_k, u_k) be the optimal solution.
Now using the optimal value of the penalty variable u_k, solve the Sub Problem:

$$\begin{aligned} \theta_k = \min f(p, v) + u_k g(p, v) \\ \text{s.t} \quad p \geq 0, v \geq 1. \end{aligned}$$

Let (p_k, v_k) be the optimal solution to the Sub Problem.
If $z_k = \theta_k$, stop. Otherwise continue with the Master Problem with added constraint: $z \leq f(p_k, v_k) + ug(p_k, v_k)$.

Figure 4: Steps involved in cutting plane algorithm

required for convergence. In the first iteration of the Master Problem, the algorithm minimizes the total cost in Equation (13) subject to the constraints $0 \leq p \leq 2, v \geq 1$. Because the Master Problem is a simple linear program, its optimal solution is obtained nearly instantaneously. However, the Sub Problem at each iteration requires multiple evaluations of the expression $g(p, v)$, which involves integration of bivariate normal distribution functions. These Sub Problems were solved by applying 'NMinimum' function of Mathematica 5.1 (Wolfram, 2004).

To uncover the effects of the different problem parameters on the optimal solution, we conducted sensitivity studies in which a single parameter at a time was varied and the rest held constant at the default values.

4.2.1 Effect of the allowed shortage

Figure 5 displays several isocost curves along with boundaries of the feasible region for two values of δ. The isocost curves illustrate how cost can be reduced by increasing p, i.e., by delaying each expansion, and decreasing v so that expansions are smaller and more frequent. The constraint curves indicate that a given service level can be achieved exactly with either a combination of high values of p and v or a combination where both values are low. As with cost minimization, the two decision variables appear complementary with respect to satisfying the service level constraint. For an allowable shortage of 4%, the cost curve of $f(p, v) = 0.92$ is nearly tangential to the feasible region boundary at $p = 1.174$ and $v = 1.400$. If the service level constraint is relaxed to allow shortage of 6%, the optimal values of both

decision variables increase to $p = 1.355$ and $v = 1.693$, at a lower cost $f = 0.83$.

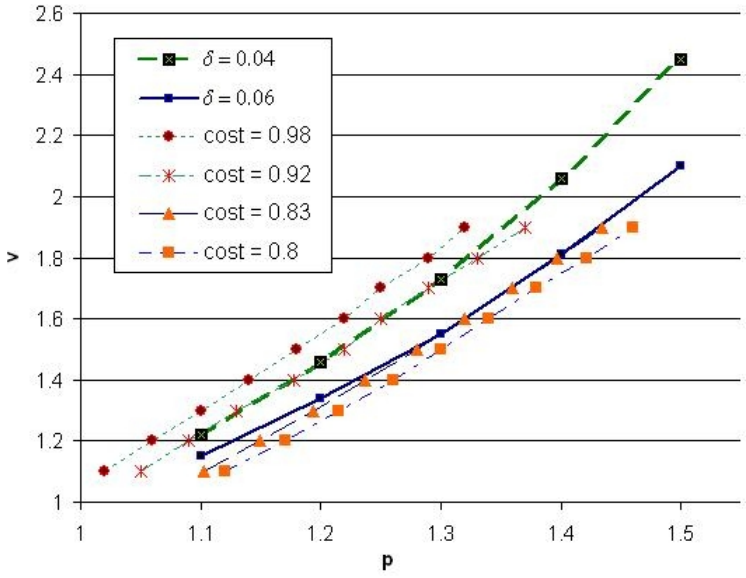

Figure 5: Effect of allowed shortage on optimal decision variable values and cost.

Table 1 summarizes these and other results and shows that relaxing the service level constraint moves the optimal strategy in the direction of delayed and infrequent expansions. It is not surprising that a larger value for the allowed shortage, δ, reduces both the optimal cost and the constraint violation penalty u^*, nor that it increases the optimal value of p. The results confirm the simultaneous increase in both expansion delay and expansion size, which is also observed in response to changes in other problem parameters below.

Table 1: Effect of allowed shortage on optimal decision variables

δ	Optimal p^*	Optimal v^*	f(p^*, v^*)	g(p^*, v^*)	u^*
0.03	1.093	1.372	1.054	-0.031	3.286
0.04	1.174	1.400	0.926	-0.002	1.945
0.05	1.270	1.560	0.877	-0.0007	1.569
0.06	1.355	1.693	0.832	-0.000	1.325

4.2.2 Effect of the lead time length

Like the allowed shortage, the length of the lead time affects the feasible region but not the cost function. Its effect on the the optimal expansion policy is shown in Table 2. Longer time intervals to complete expansion projects require expansion to begin with smaller initial shortages to maintain the given service level. However, the optimal size parameter actually decreases with L. This result differs from that in Ryan (2004) where, due to the sequential decision making process, an increase in the lead time resulted

only in more proactive expansions but had no effect on optimal expansion sizes.

Table 2: Effect of lead time on optimal decision variables

L	Optimal p^*	Optimal v^*	f(p^*, v^*)	g(p^*, v^*)	u^*
2	1.270	1.560	0.877	-0.0007	1.569
3	1.004	1.014	0.925	-0.0003	2.996
4	0.996	1.006	0.964	-0.001	3.398
5	0.995	1.004	0.967	-0.001	3.598

4.2.3 Effect of economies of scale

Economies of scale play no role in feasibility, but influence the impact of expansion size on the cost. The expansion cost was modeled by a power law that considers the benefits of adding capacity in larger increments, thereby taking advantage of economies of scale. We expect that as we decrease the parameter $a < 1$, the optimal size of expansion increases. Table 3 shows that the optimal value of p also increases, so that the large and infrequent expansions are also delayed. Because the timing and size decisions were taken separately in the Ryan (2004) analysis, there the increase in economies of scale effect did not change the optimal value of the timing variable.

Table 3: Effect of economies of scale on optimal decision variables

a	Optimal p^*	Optimal v^*	f(p^*, v^*)	g(p^*, v^*)	u^*
0.99	1.270	1.560	0.877	-0.0007	1.569
0.9	1.457	2.171	0.864	-0.015	1.676
0.8	1.530	2.670	0.855	-0.045	2.243
0.75	1.533	2.840	0.850	-0.006	2.523

4.2.4 Effect of the drift parameter

The service firm's management may be able to change capacity expansion parameters such as a and L by research and development, negotiations with suppliers, or adjustments in internal operations. Demand parameters are often more difficult to control or forecast, and they influence both the objective function and the service level constraint. The drift parameter of the demand process corresponds to the average growth rate. Table 4 shows how the optimal timing and size parameters change in response to an increase in the drift. We see that increased expected growth results in smaller optimal values of both decision variables. That is, the firm's response to a higher average rate of growth should be to make proactive and small (i.e., frequent) expansions. The optimal cost and shortage penalties both increase with expected growth in demand. In additional numerical test not displayed, the same trend in the optimal decision variables with respect to the drift parameter was observed for smaller lead times and a variety of values of the economy of scale parameter.

Table 4: Effect of demand drift on optimal decision variables

μ	Optimal p^*	Optimal v^*	f(p^*, v^*)	g(p^*, v^*)	u^*
0.02	1.270	1.560	0.877	-0.0007	1.569
0.05	1.092	1.443	1.968	-0.053	4.023
0.08	0.994	1.011	4.235	-0.001	9.604
0.1	0.987	1.009	13.127	-0.004	29.418

4.2.5 Effect of the volatility parameter

Because demand fluctuations are common, it is critical to study the effects of demand volatility on the decision variables. This parameter also can be interpreted as uncertainty in the demand forecast, which increases the farther forecasts are extended. Table 5 shows that demand volatility prompts the service provider to initiate expansions early relative to the capacity position and repeat them frequently. That is, fluctuations in demand have a similar effect on the expansion policy and its cost as does high positive drift. A service provider facing more predictable demand, however, has the luxury of delaying the expansion and even can tolerate shortages before the start of the expansion project. Delays in expansion (increased p) are coupled with larger increments (increased v). The increased risk due to volatility increases the optimal expected cost.

Table 5: Effect of demand volatility on optimal decision variables

σ	Optimal p^*	Optimal v^*	f(p^*, v^*)	g(p^*, v^*)	u^*
0.20	1.270	1.560	0.877	-0.0007	1.569
0.25	1.131	1.332	1.393	-0.008	2.919
0.3	1.005	1.030	2.095	-0.001	5.132
0.35	1.000	1.021	3.374	-0.002	7.397
0.4	0.997	1.018	6.345	-0.008	12.545

However, unlike with the drift parameter, the change in volatility affects the trend in the optimal decision variables differently, depending on the lead time length. At lower values of the lead time, increase in demand volatility causes delayed and bigger expansions. On the other hand, with long lead times, an increase in demand volatility triggers earlier and smaller sized expansions. The results are shown in Table 6.

Table 6: Effect of demand volatility at various lead time values

	$L = 0.5$			$L = 1$			$L = 1.5$			$L = 2$		
σ	p^*	v^*	$f(p^*, v^*)$	p^*	v^*	$f(p^*, v^*)$	p^*	v^*	$f(p^*, v^*)$	p^*	v^*	$f(p^*, v^*)$
0.2	1.586	1.877	0.664	1.489	1.811	0.731	1.385	1.718	0.806	1.270	1.560	0.876
0.3	1.612	1.916	1.615	1.433	1.768	1.808	1.237	1.509	1.998	1.005	1.030	2.095

Table 7 confirms this combined effect for the default parameter values ($\mu = 0.05, \sigma = 0.2, r = 0.1, a = 0.9, \delta = 0.001$) from Ryan (2004). For large lead times, an increase in demand volatility requires the more proactive and frequent strategy. With smaller lead times, an increase in demand volatility causes delayed and infrequent expansions. For moderate lead times, the effect of demand volatility is similar to that seen in the Ryan (2004) results – that is, to have more proactive but less frequent (larger) expansions. The effect on p^* of increasing σ when L is small possibly can be explained as the ability to "wait and see" about volatile demand when small lead times permit a fast reaction. The larger p^* is coupled with larger v^* as in most of the other results. Also, if $L = 0$, one can choose $p = 1$ with no risk of shortage, and previous work such as in Bean et al. (1994) has shown that the impact of greater uncertainty is to increase the optimal expansion increment.

Table 7: Effect of demand volatility at various lead time values for parameter values from Ryan (2004)

	$L = 0.5$			$L = 1$			$L = 1.5$			$L = 2$		
σ	p^*	v^*	$f(p^*, v^*)$	p^*	v^*	$f(p^*, v^*)$	p^*	v^*	$f(p^*, v^*)$	p^*	v^*	$f(p^*, v^*)$
0.2	1.340	1.760	2.562	1.243	1.708	2.784	1.148	1.620	3.012	1.037	1.437	3.262
0.283	1.460	2.009	5.833	1.298	1.897	6.457	1.142	1.680	7.054	0.998	1.302	7.596

As a final comparison with the model of Ryan (2004), Table 8 quantifies the cost savings our model achieves for a sampling of parameter values. The timing and size parameters found by Ryan's sequential procedure are denoted p' and v', respectively. The substantial cost reductions achieved by the solution (p^*, v^*) to problem (14) can be attributed in varying proportions to satisfying the shortage constraint more nearly as an equality, optimizing simultaneously rather than sequentially, and allowing $p > 1$.

Table 8: Comparison with results from Ryan (2004).

Parameters					Ryan (2004) results				Our results			
μ	σ	L	δ	a	p'	v'	$f(p', v')$	$g(p', v')$	p^*	v^*	$f(p^*, v^*)$	$g(p^*, v^*)$
0.02	0.2	2	0.05	0.99	0.823	1.009	1.437	-0.049	1.270	1.560	0.876	-0.000
0.05	0.2	2	0.05	0.99	0.790	1.012	2.560	-0.052	1.092	1.443	1.968	-0.053
0.02	0.4	2	0.05	0.99	0.570	1.017	12.140	-0.064	0.997	1.018	6.345	-0.008
0.02	0.2	0.5	0.05	0.99	1.000	1.009	0.955	-0.007	1.586	1.877	0.664	-0.003
0.02	0.2	4	0.05	0.99	0.619	1.009	2.608	-0.047	0.996	1.006	0.964	-0.001
0.02	0.2	2	0.06	0.99	0.843	1.009	1.368	-0.049	1.355	1.693	0.832	-0.000
0.02	0.2	2	0.05	0.75	0.823	1.275	2.047	-0.106	1.533	2.840	0.850	-0.006

5 Conclusions and Future Work

The formulation of the service level constraint in this paper allows for expansion policies that either anticipate demand reaching the capacity position or react to demand having exceeded it. Its evaluation

by using barrier option pricing tools is exact, and therefore the numerical results in this paper supersede those in Ryan (2004), where timing and size decisions were made sequentially and evaluation of the service level constraint could err on the side of caution. We found analytically and numerically that the optimal expansion parameters are complementary. The delayed and infrequent expansion strategy that corresponds to large values of both parameters is optimal when greater shortages are permissible, lead times are short, economies of scale are significant, average demand growth is small, and/or demand volatility is low. The opposite strategy, of small and frequent expansions that are initiated proactively, is optimal when the problem parameters reflect a more stringent service level, smaller economies of scale, and greater risk of shortage from the combination of long lead times and faster or more volatile demand growth. Some unexpected behaviors, particularly in the combined effects of volatility and lead time, could be explored further.

Using this model, the service provider can optimize the parameters of the expansion policy according to numerical values of the model parameters observed in the industry in which the service provider operates. Relaxing the assumptions of the model suggests new directions in which this base model can be extended. One of the most important assumptions made was that the demand process follows the GBM process. Although this may be true for some industries, some bumpy demand processes may be more closely represented by a probability distribution that incorporates sudden changes in demand values, for example, a GBM process with jumps. Also, the assumption that the demand follows a GBM process over an infinite horizon could be modified; and the problem could be solved over a finite time horizon. Market saturation could be modeled by a process with nonhomogeneous drift. The present capacity expansion problem assumes non-deteriorating capacity. Geometric deterioration in the capacity *position* can be included by adding the deterioration rate to the growth rate as in Whitt (1981); however, allowing deterioration only in installed capacity complicates the analysis significantly. In the current model, only capacity *expansions* were considered. There are practical examples, especially where demand does not always grow, where reducing the capacity over a period of time may be profitable and hence, considering capacity *contraction* might prove worthwhile. Lastly, a deterministic fixed lead time was considered for expansion. A probability distribution could be considered for the lead time to make it more realistic and the effect of stochastic lead time on the capacity expansion problem could be analyzed.

Appendix

A Partial up-and-out barrier call option

A barrier option is a path dependent option where the payoff depends not only on the final price of the underlying asset but also on whether or not the underlying asset has reached some other "barrier" price during the life of the option (Rubinstein and Reiner, 1991). Heynen and Kat (1997) give an explicit analytical equation for up-and-out call option value. Notations used by them are specified here. It can be shown that the results obtained by using Chuang (1996) are exactly the same as in Heynen and Kat (1997). For the up-and-out option, define:

S_0: Initial price of the stock,

t_1: Arbitrary time before the expiration when the monitoring ends

T: Expiration time

K: Strike price

H: Barrier price

μ: Drift parameter

σ: Volatility parameter

γ: Growth rate, $\gamma = \mu + \frac{\sigma^2}{2}$

Then assuming the stock price follows a GBM process with drift μ and volatility σ^2, the price of the up-and-out call option is given by:

$$\begin{aligned} e^{-\gamma T}\left(E\left[(S_T-K)1(S_T \geq K, \max S_t \leq H : 0 \leq t \leq t_1)\right]\right) = \\ S_0\Psi\left(d_1, -e_1, -\sqrt{\tfrac{t_1}{T}}\right) - \left(\tfrac{H}{S_0}\right)^{\frac{2\gamma}{\sigma^2}+1}\Psi\left(f_1, -e_1', -\sqrt{\tfrac{t_1}{T}}\right) \\ -e^{-\gamma T}K\Psi\left(d_2, -e_2, -\sqrt{\tfrac{t_1}{T}}\right) + e^{-\gamma T}K\left(\tfrac{H}{S_0}\right)^{\frac{2\gamma}{\sigma^2}-1}\Psi\left(f_2, -e_2', -\sqrt{\tfrac{t_1}{T}}\right), \end{aligned} \tag{17}$$

where $\Psi(x, y, \rho)$ is the bivariate standard normal distribution function with correlation coefficient ρ and,

$$\begin{aligned} d_1 &= \frac{-\ln\left(\frac{K}{S_0}\right) + \left(\gamma + \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}} \quad , \qquad d_2 = d_1 - \sigma\sqrt{T} \\ e_1 &= \frac{-\ln\left(\frac{H}{S_0}\right) + \left(\gamma + \frac{\sigma^2}{2}\right)t_1}{\sigma\sqrt{t_1}} \quad , \qquad e_2 = d_1 - \sigma\sqrt{t_1} \\ e_1' &= e_1 + \frac{2\ln\frac{H}{S_0}}{\sigma\sqrt{t_1}} \quad , \qquad e_2' = e_1' - \sigma\sqrt{t_1} \\ f_1 &= \frac{-\ln\left(\frac{K}{S_0}\right) + 2\ln\left(\frac{H}{S_0}\right) + \left(\gamma + \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}} \quad , \qquad f_2 = f_1 - \sigma\sqrt{T} \end{aligned}$$

With respect to Equation (8), the terms of Heynen and Kat (1997) have following correspondence:

$$S_T \leftrightarrow Q_n(u), \quad K \leftrightarrow K_n, \quad t_1 \leftrightarrow u - L, \quad T \leftrightarrow u, \quad H \leftrightarrow pK_n, \quad S_0 \leftrightarrow P(t_n).$$

B Convexity

A necessary condition for the convergence of the cutting plane algorithm in a minimization problem is the convexity of the objective function and the constraint expression. Toward that, we were only able to find numerical evidence of pseudo-convexity of the objective function. Also, owing to the complexity of the constraint equation, analytical proof of convexity is difficult. Therefore, we use a technique that provides us with some evidence of convexity of the objective function as well as the constraint equation.

Atlason et al. (2004) discussed a numerical method for checking whether a function is concave. Via Theorem 9 of their work, they proposed solving a relatively simple linear program (LP) to check for concavity of any function. This method can be used to check convexity of a function by just a change of sign. The LP proposed by Atlason et al. (2004) changes given function values so that a supporting hyperplane for the convex hull of the points can be fitted through each sampled point. The objective of this LP is to minimize the change in the function values that needs to be made to accomplish this goal. The LP to test the convexity of the service level constraint expression of our problem is formulated as:

$$\min \sum_{i=1}^{k} |b_i|$$

subject to

$$a_{0_i} + \left(a^i\right) \begin{bmatrix} p^i & v^i \end{bmatrix}^T = -g(p^i, v^i) + b_j, \quad \forall i \in \{1, \cdots, k\}$$

$$a_{0_i} + \left(a^i\right) \begin{bmatrix} p^j & v^j \end{bmatrix}^T = -g(p^j, v^j) + b_j, \quad \forall i \in \{1, \cdots, k\}, \quad \forall j \in \{1, \cdots, k\}, i \neq j$$

Here, k is the number of points sampled. To linearize the objective function, the standard trick of writing $b_i = b_i^+ - b_i^-$ can be adopted. Then $|b_i| = b_i^+ + b_i^-$, where b_i^+ and b_i^- are nonnegative. The decision variables are:

$a_{0_i} \in \mathcal{R}, i \in \{1, \cdots, k\}$: intercepts of the hyperplane;

$a^i \in \mathcal{R}^2, i \in \{1, \cdots, k\}$: slopes of the hyperplane and

$b_i^+, b_i^- \in \mathcal{R}, i \in \{1, \cdots, k\}$: change in the function values.

Atlason et al. (2004) also proved that when the optimal objective value of the LP is 0, there exists a concave function that has the same value as the function in question at all the points sampled. We applied this method to our objective function and constraint equation with a change of sign. Because the

solution of that linear program had zero objective value, there exists a convex function that has the same value as the function in question at all the sampled points. At every instance, the constraint function and the objective function for our problem passed this test of convexity.

Acknowledgments

We are grateful to the anonymous referees for valuable suggestions and to Ananda Weerasinghe for several helpful conversations.

References

R. Amir. Supermodularity and complementarity in economics: an elementary survey. *Southern Economic Journal*, 7(3):636–660, 2003.

J. Atlason, M.A. Epelman, and S.G. Henderson. Call center staffing with simulation and cutting plane algorithm. *Annals of Operations Research*, 127:333–358, 2004.

L. Barker. Pilot shortages: How to reduce their impact on rural and smaller market. Committee on Commerce, Science and Transportation, Subcommittee on Aviation and Aeronautics, 2000. Retrieved in June 2005 from URL: http://commerce.senate.gov/hearings/0725bar.pdf.

M.S. Bazaraa, H.D. Sherali, and C.M. Shetty. *Nonlinear programming: Theory and algorithms*. John Wiley and Sons, New York, 1993.

J.C. Bean and R.L. Smith. Optimal capacity expansion over infinite horizon. *Management Science*, 31 (12):1523–1532, 1985.

J.C. Bean, J.L. Higle, and R.L. Smith. Capacity expansion under stochastic demand. *Operations Research*, 40(2):S210–S216, 1994.

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

J.A. Buzacott and A.B. Chaouch. Capacity expansion with interrupted demand growth. *European Journal of Operations Research*, 34:19–26, 1988.

M. Cakanyildirim, R.O. Roundy, and S.C. Wood. Optimal machine capacity expansions with nested limitations under stochastic demand. *Naval Research Logistics*, 51:217–241, 2004.

A.B. Chaouch and J.A. Buzacott. The effect of lead time on plant timing and size. *Production and Operations Management*, 3(1):38–54, 1994.

C.S. Chuang. Joint distribution of brownian motion and its maximum, with a generalization to correlated bm and application to barrier options. *Statistics and Probability Letters*, 28:81–90, 1996.

M.H.A. Davis, M.A.H. Dempster, S.P. Sethi, and D. Vernes. Optimal capacity expansion under uncertainty. *Advances in Applied Probability*, 19:156–176, 1987.

G.L. Donohue. Testimony before the house of representatives. Committee on Science, Subcommittee on Space and Aeronautics, 2000. Retrieved in June 2005 from URL: http://house.gov/science/donohue.html.

B.M. Goldman, H.B. Sosin, and M.A. Gatto. Path dependent options: Buy at low; sell at high. *The Journal of Finance*, XXXIV(5):1111–1127, 1979.

R.E. Gomory. An algorithm for integer solutions to linear programs. In R.L. Graves and P. Wolfe, editors, *Recent advances in mathematical programming*. McGraw–Hill Book Company, Inc., New York, 1963. Originally appeared in 1959 as a Princeton-IBM Mathematical Research Project technical report.

G. Hadley and T.M. Whitin. *Analysis of inventory systems*. Prentice-Hall Inc, Englewood Cliffs, New Jersey, 1963.

H. He and R.S. Pindyck. Investment in flexible production capacity. *Journal of Economic Dynamics and Control*, 16:575–599, 1992.

R.C. Heynen and H.M. Kat. Chapter six: Barrier options. In L. Clewlow and C. Strikland, editors, *Exotic Options: The State of the Art*. International Thompson Business Press, London and Boston, 1997.

G.E. Hopkins. A short history of pilot shortages. *Airline Pilot*, February 2001.

S. Karlin and T.M. Taylor. *A first course in stochastic processes*. Academic Press, New York, 1975.

J.E. Kelley. The cutting plane method for solving convex programs. *Journal of Society of Industrial Applications of Mathematics*, 8(4):704–712, 1960.

H. Klemm. On the operating characteristic 'service level'. In *Inventory Control and Water Storage*, pages 169–178. North-Holland Publishing Company, Amsterdam, 1971.

S. Kurabuk and S.D. Yu. Coordinating strategic capacity planning in semiconductor industry. *Operations Research*, 51(6):839–849, 2003.

M.B. Lieberman. Capacity utilization: Theoretical models and empirical tests. *European Journal of Operations Research*, 40:155–168, 1989.

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

R.R. Marathe and S.M. Ryan. On the validity of geometric brownian motion assumption. *The Engineering Economist*, 50(2):159–192, 2005.

A. Marin and J. Salmeron. A risk function for the stochastic modeling of electric capacity expansion. *Naval Research Logistics*, 48:662–683, 2001.

R. Merton. Theory of rational option pricing. *Bell Journal of Economics and Management Science*, 4(1): 141–183, 1973.

M. Musiela and M. Rutkowski. *Martingale methods in financial modeling*. Springer Verlag, Berlin, 1997.

D. Pak, N. Pornsalnuwat, and S.M. Ryan. The effect of technological improvement on capacity expansion for uncertain exponential demand with lead times. *The Engineering Economist*, 49:95–118, 2004.

S.M. Ross. *An Introduction to Mathematical Finance*. Cambridge University Press, Cambridge, UK, New York, 1999.

M. Rubinstein. Exotic options. Unpublished manuscript, University of California at Berkeley, 1991.

M. Rubinstein and E. Reiner. Breaking down the barriers. *Risk*, 4:28–35, 1991.

S.M. Ryan. Capacity expansion for random exponential demand growth with lead time. *Management Science*, 50(6):740–748, 2004.

H. Schneider. Effect of service-levels on order-points or order-levels of inventory models. *International Journal of Production Research*, 19(6):615–631, 1981.

R.L. Smith. Turnpike results for single location capacity management. *Management Science*, 25(5): 474–484, 1979.

M.J. Sobel. Fill rates of single-stage and multistage supply system. *Manufacturing and Service Operations Management*, 6(1):149–160, 2004.

J.A. van Meighem. Capacity management, investment and hedging: review and recent developments. *Manufacturing and Service Operations Management*, 5(4):269–302, 2003.

W. Whitt. The stationary distribution of a stochastic clearing process. *Operations Research*, 29(2): 294–308, 1981.

D.E. Wolfe. Accelerating the cutting plane method for nonlinear programming. *Journal of Society of Industrial Applications of Mathematics*, 9(3):481–488, 1961.

Wolfram. Mathematica 5.1, 2004. Software package, URL: www.wolfram.com.

W.I. Zangwill. *Nonlinear programming: A unified approach*. Prentice Hall, Englewood Cliffs, New Jersey, 1969.