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

Rahul R. Marathe and Sarah M. Ryan

Department of Industrial and Manufacturing Systems Engineering,
Iowa State University, Ames, IA 50011.

October 20, 2006

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. Shadow price values suggest that timing is driven mainly by the service level constraint while expansion size is determined by economics.

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

*corresponding author. E-mail: smryan@iastate.edu, Phone: (515) 294 4347
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 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.

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 occur because of the 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 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 investigates under what conditions it is optimal to allow shortages before initiating expansion.
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, 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.

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 with objective of cost and risk minimization. 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 Bender’s decomposition. Kurabuk and Yu (2003) considered a two stage model for a semiconductor manufacturer, where the strategic planning involves how much of which microelectronic technology to produce and also
decision about 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 semiconductor industry where the machine purchase, floor space and shell expansions were jointly optimized over a finite time horizon. They considered a 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 above papers in the following ways. The demand process we consider is the same as in Whitt (1981); however, that model did not include expansion lead times. 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. Chaouch and Buzacott (1994) considered starting the expansion either before or after the demand crosses the capacity position for a different demand model. 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. Using the results from options pricing theory, he showed that risk can be effectively accounted for in a wide range of operational planning models, particularly the linear capacity planning models. To our knowledge, this paper was the first to point 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 equation was found for the joint distribution of the Brownian motion and its maximum 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. Chuang (1996) also included some remarks about reducing the numerical computations by a clever change of variables. 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 to 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. This sets up an interesting optimization problem to jointly optimize both decision aspects, 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. To our knowledge, this is the only instance of quantifying undercapacity as an exotic option, and it yields economic insight. Finally, the optimization problem to minimize cost subject to the service level constraint is solved using a cutting plane algorithm for the timing and size policy parameters simultaneously. Numerical instances reveal how these two dimensions of the expansion policy interact, 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 will use consistent notation. Let $B(t)$ be a Brownian motion having drift $\mu$ and volatility $\sigma^2$ with $B(0) = 0$. The demand for the service is given by a GBM process $P(t) = P(0)e^{B(t)}$. 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 $\mu k$ and volatility $\sigma^2 k$. Hence, given $P(t)$, the logarithmic growth over a small period of time $\Delta t$ 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 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.
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. Note that for a positive drift ($\mu > 0$), with certainty, $T_n < \infty$ (Karlin and Taylor, 1975). For any realization $\omega$ of the Brownian motion $B(t)$ let $t_n = T_n(\omega)$. 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:

$$C_n(X) = kX^a, \quad (1)$$

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 proportion of its previous value. In this paper, we assume that each expansion begins when demand reaches some fixed proportion, $p$, of the capacity position, where $p$ may be less than, equal to, or greater than one. The sequence of capacity levels follows $K_n = vK_{n-1}, n \geq 1$, where $v \geq 1$. Ryan (2004) showed that for $p < 1$ with fixed lead times, the value of $p$ to guarantee a specified service
level can be found according to the Black-Scholes formula for pricing a European call option. Moreover, assuming this timing policy is followed, the expansion size policy minimizes the infinite horizon discounted cost under the expansion cost function in Equation (1). To allow use of the European call option value, the capacity shortage constraint was overly conservative in some cases. In this paper, we generalize the model and analysis to consider also the case \( p \geq 1 \), and we reformulate the constraint to be no more restrictive than necessary. This reformulation requires the use of more sophisticated mathematical analysis.

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](image)

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

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. \( \alpha, \beta, \gamma \), are used in different situations. The \( \alpha \) 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 \( \beta \) service level as the fraction of demand that is satisfied. Lastly, the \( \gamma \) service level is defined in terms of cumulative unsatisfied demand, which makes it applicable only in case of backorders. Because services generally cannot be backordered, our constraint is based on the \( \beta \) definition. In the capacity expansion problem, the proportion of demand that is satisfied during the \( n^{th} \) cycle is:

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

which is a random quantity. To be consistent with cost discounting, we discount future shortages to time \( t_n \).

To guarantee a service level of at least \( 1 - \delta \), where \( \delta \) is a small positive number, define the violation random
variable as:

\[ \xi_n \equiv \int_{t_n + L}^{T_n + L} e^{-r(t-t_n)} \max(P(t) - K_n, 0) dt - \delta \int_{t_n + L}^{T_n + L} e^{-r(t-t_n)} P(t) dt \]  

(3)

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

2.2 Dynamic programming formulation

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.

For a given capacity level of \( K \), let \( V_t(K) \) be the minimum expected cost, discounted to time \( t \), of expanding capacity over the infinite horizon while satisfying the service level constraint. Then for \( n \geq 1 \),

\[
V_{t_n}(K_{n-1}) = \min_{X_n, T_{n+1}} \left[ C_n(X_n) + E_{t_n}[e^{-r(T_{n+1}-t_n)}] V_{T_{n+1}}(K_n) \right]
\]

subject to \( E[\xi_n] \leq 0 \),

(4)

where \( K_n = K_{n-1} + X_n \).

Under the Whitt-Luss expansion policy, the problem is to find policy parameters that give the minimum infinite horizon expansion cost discounted to time 0, such that the service level in each expansion cycle is met:

\[ V_0(K_0) = \min \left[ E[e^{-rT_1}] V_{T_1}(K_0) \right] \]

(5)

In the next section, we express the service level in terms of the policy parameters using the concepts of the up-and-out barrier option and specify a closed-form nonlinear expression for the infinite horizon cost in the same terms.

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 equation—to evaluate the service level constraint (3) analytically. The infinite horizon expansion cost is obtained a closed form expression for the objective function in terms of the policy parameters.
3.1 Analysis of the service level constraint

Evaluation of the expected constraint violation $E[ξ_n]$ 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 $v$.

Birge (2000) first pointed out the correspondence between the capacity shortage and the call option. If all demand could be met, expected discounted revenue at time $t > t_n$ would be proportional to $e^{-r(t-t_n)} \int_{P(t)}^\infty P(t) dF(P(t))$, assuming a constant sale price. Given that sales may be limited by actual capacity because services can be neither stored nor backordered, the discounted expected shortage $e^{-r(t-t_n)} E[\max(P(t) - K(t), 0)]$ multiplied by the price represents the lost revenue. Competing service providers have no obligation to serve this excess demand but do have the option to do so in case $P(t) > K(t)$. Completing the analogy, the demand corresponds to the asset price, $K(t)$ is the strike price, and $t$ is the expiration date when the European option would be exercised.

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 undercapacity 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:

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

(6)

where the expectations are taken with respect to time $t_n$.

We now simplify each of the terms on right hand side of Equation (6) to obtain the constraint expression in terms of the decision variables, $p$ and $v$. The first of these terms is equal to:

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

(7)

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:

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

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

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

(8)

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 (8) is equivalent to:

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

(9)
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 (9) becomes:

$$I_1^n = \int_{t_n + L}^{\infty} e^{-ru} E \left[ \left( P(t_n)e^{W_n(t-t_n)} - K_n \right) 1 \left( W_n(t-t_n) \geq \ln \frac{K_n}{P(0)} - B(t_n) \right) ight] du \left( \max W_n(k) \leq \ln \frac{pK_n}{P(0)} - B(t) : 0 \leq k \leq t - L - t_n \right) dt,$$

because $P(t) = P(0)e^{B(t)} = P(t_n)e^{B(t-t_n)}$ 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 (9) becomes:

$$I_1^n = \int_{t_n + L}^{\infty} e^{-ru} E \left[ (Q_n(t) - K_n) 1(Q_n(t) \geq K_n) 1(\max Q_n(s) \leq pK_n : 0 \leq s \leq u - L) \right] du. \quad (10)$$

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 the knock-out barrier options, particularly the down-and-out call option. Appropriate changes could be made for the up-and-out 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
- $\mu$: Drift parameter
- $\sigma$: Volatility parameter
- $\gamma$: Growth rate, $\gamma = \mu + \frac{\sigma^2}{2}$

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

$$E \left[ (S_T - K) 1(S_T \geq K, \max S_t \leq H : 0 \leq t \leq t_1) \right] =$$

$$S_0 \Psi \left( d_1, -e_1, -\sqrt{\frac{t_1}{T}} \right) - \left( \frac{H}{S_0} \right)^{\frac{\sigma^2}{2} + 1} \Psi \left( f_1, -e_1', -\sqrt{\frac{t_1}{T}} \right)$$

$$- e^{-\gamma T} K \Psi \left( d_2, -e_2, -\sqrt{\frac{t_1}{T}} \right) + e^{-\gamma T} K \left( \frac{H}{S_0} \right)^{\frac{\sigma^2}{2} - 1} \Psi \left( f_2, -e_2', -\sqrt{\frac{t_1}{T}} \right),$$

(11)

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

$$d_1 = -\ln \left( \frac{K}{S_0} \right) + \left( \gamma + \frac{\sigma^2}{2} \right) T \frac{1}{\sigma \sqrt{T}} , \quad d_2 = d_1 - \sigma \sqrt{T}$$

$$e_1 = -\ln \left( \frac{H}{S_0} \right) + \left( \gamma + \frac{\sigma^2}{2} \right) \frac{t_1}{\sigma \sqrt{t_1}} , \quad e_2 = d_1 - \sigma \sqrt{t_1}$$

$$e_1' = e_1 + \frac{2 \ln \frac{H}{S_0}}{\sigma \sqrt{t_1}} , \quad e_2' = e_1' - \sigma \sqrt{t_1}$$

$$f_1 = -\ln \left( \frac{K}{S_0} \right) + 2 \ln \left( \frac{H}{S_0} \right) + \left( \gamma + \frac{\sigma^2}{2} \right) T \frac{1}{\sigma \sqrt{T}} , \quad f_2 = f_1 - \sigma \sqrt{T}$$

With respect to Equation (10), the terms of Heynen and Kat (1997) have following correspondence:

$$S_T \leftrightarrow Q_n(u), \quad K \leftrightarrow K_n, \quad t_1 \leftrightarrow u - L, \quad T \leftrightarrow u, \quad H \leftrightarrow pK_n, \quad S_0 \leftrightarrow P(t_n).$$
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 (10) and (11), we have:

$$I_n^1 = K_n \int_0^\infty e^{-ru} \left[ e^{\gamma u} \left( \frac{P}{v} \right) \Psi \left( -\ln \left( \frac{v}{P} \right) + \left( \gamma + \frac{\sigma^2}{2} \right) u \frac{\ln v + \left( \gamma + \frac{\sigma^2}{2} \right) (u - L)}{\sigma \sqrt{u}} , -\sqrt{\frac{u - L}{u}} \right) \right] du.$$  

$$- \frac{2}{v^2} e^{\gamma u} \left( \frac{P}{v} \right) \Psi \left( -\ln \left( \frac{v}{P} \right) + \left( \gamma - \frac{\sigma^2}{2} \right) u \frac{\ln v - \left( \gamma - \frac{\sigma^2}{2} \right) (u - L)}{\sigma \sqrt{u - L}} , -\sqrt{\frac{u - L}{u}} \right)$$  

$$- \frac{2}{v^2} \frac{e^{\gamma u}}{\sqrt{u}} \left( \frac{P}{v} \right) \Psi \left( -\ln \left( \frac{v}{P} \right) + \left( \gamma + \frac{\sigma^2}{2} \right) u \frac{\ln v + \left( \gamma + \frac{\sigma^2}{2} \right) (u - L)}{\sigma \sqrt{u}} , -\sqrt{\frac{u - L}{u}} \right)$$  

$$- \frac{2}{v^2} \frac{e^{\gamma u}}{\sqrt{u}} \left( \frac{P}{v} \right) \Psi \left( -\ln \left( \frac{v}{P} \right) + \left( \gamma - \frac{\sigma^2}{2} \right) u \frac{\ln v - \left( \gamma - \frac{\sigma^2}{2} \right) (u - L)}{\sigma \sqrt{u - L}} , -\sqrt{\frac{u - L}{u}} \right)$$

We evaluate the second term on the right hand side of the Equation (6) in the same way as above. Let

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

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

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

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

$$I_n^2 = K_n \int_L^\infty e^{(\gamma-r)u} \left( \frac{P}{v} \right) \Phi \left( \frac{\ln v - \left( \gamma + \frac{\sigma^2}{2} \right) (u - L)}{\sigma \sqrt{u - L}} \right)$$  

$$- K_n \int_L^\infty e^{(\gamma-r)u} \frac{2}{v^2} \left( \frac{P}{v} \right) \Phi \left( -\ln v - \left( \gamma - \frac{\sigma^2}{2} \right) (u - L) \right)$$  

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

The final service level constraint in Equation (6) can now be expressed as: $E[\xi_n] = I_n^1 - \delta I_n^2 \leq 0$, which is true if and only if:

$$g(p, v) = \frac{I_n^1 - \delta I_n^2}{K_n} \leq 0$$

where the terms in the numerator are evaluated using Equations (12) and (13), respectively. By dividing the constraint violation during a cycle by the actual capacity, $g(p, v)$ is obtained in units of time.

15
After dividing out $K_n$ in Equation (14), the expected constraint violation does not depend on $n$. 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 optimization

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. Pak et al. (2004) showed that under the policy assumed here, the expected infinite horizon discounted cost is given by:

$$f(p, v) \equiv \frac{kK_0^a(v - 1)^a p^{-\lambda}}{1 - v^{\alpha - \lambda}},$$

(15)

where $K_0$ is the initial capacity and $\lambda = \sqrt{\frac{\mu^2}{\sigma^2} + \frac{2\sigma}{\alpha} - \frac{\mu}{\sigma^2}}$ (note that $\lambda > 1$ for $r > \gamma$).

This expression forms our objective function for the non-linear program in terms of the policy parameters, $p$ and $v$. The optimal values of the policy parameters must minimize this total cost of expansion and also must satisfy the constraint of maintaining the service level. Therefore, the optimization problem under the assumed capacity expansion policy is:

$$\min f(p, v)$$

subject to:

$$g(p, v) \leq 0$$

$$v \geq 1; p \geq 0.$$  

(16)

Here, the expression for the objective function is obtained from Equation (15) and the service level constraint expression is obtained from Equation (14). 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 size parameter $v$ to $v \geq 1$.

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 the Equation (15) 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 (16) mathematically states the capacity expansion problem. Its complexity is evident from the fact that it is a non-linear optimization problem with a rather 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 (16). As seen from Equation (14), 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 for our problem. The Lagrangean dual of the original optimization problem of Equation (16) 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).

4.1.1 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

$$a_0 + (a^i)^T [p^i, v^i]^T = -g(p^i, v^i) + b_j, \quad \forall i \in \{1, \cdots, k\}$$

$$a_0 + (a^j)^T [p^j, v^j]^T = -g(p^j, v^j) + b_j, \quad \forall i \in \{1, \cdots, k\}, \forall j \in \{1, \cdots, k\}, i \neq j$$

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:

$a_0 \in \mathbb{R}, i \in \{1, \cdots, k\}$: intercepts of the hyperplane;

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

$b_i^+, b_i^- \in \mathbb{R}, i \in \{1, \cdots, k\}$: 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.

4.1.2 Steps involved in cutting plane algorithm

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. Moreover, the Sub Problem constraints are linear with non-linear objective function.
Hence the total computation time to solve these Sub Problems is far less than the original problem. Finally,
Zangwill (1969) provides proof for finite convergence of the cutting plane algorithm, which means that the
optimal solution to the original problem in Equation (16) will eventually be found, provided that the problem
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:

\[
\begin{align*}
\max & \quad z \\
\text{subject to} & \quad z \leq f(p_j, v_j) + u g(p_j, v_j) \quad \text{for} \quad j = 1 \cdots k - 1 \\
& \quad u \geq 0.
\end{align*}
\]

Let \((z_k, u_k)\) be the optimal solution.

Now using the optimal value of the penalty variable \(u_k\), solve the Sub Problem:

\[
\theta_k = \min f(p, v) + u_k g(p, v)
\]

\[
\text{s.t} \quad p \geq 0, v \geq 1.
\]

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:

\[
z \leq f(p_k, v_k) + u g(p_k, v_k).
\]

Figure 4: Steps involved in cutting plane algorithm

is feasible.

4.2 Numerical Results

We start with the numerical analysis of the service level constraint formulated in Equation (14) then apply the cutting plane algorithm to the problem (16). We show how the cutting plane algorithm converges for one problem instance, and then explore the impact of changes in various model parameter.

4.2.1 Results regarding the service level

As the value of timing parameter \(p\) increases, in effect, we are delaying the start of expansion project further. Hence intuitively, the shortage violation function \(g(p, v)\) should become less and less negative and in fact become positive corresponding to constraint violation for large values of the timing parameter. Similarly, as the value of size parameter \(v\) increases, corresponding to larger expansions each time, we expect \(g(p, v)\) to become more and more negative (hence, more and more favorable). These behaviors are illustrated in Figures 5 and 6. For Figure 5, we fix the size factor \((v)\)=6; the values of the problem parameters were: drift \((\mu)\)=8%, volatility \((\sigma)\)=20%, lead time \((L)\)=2 years, interest rate \((r)\)=11%, whereas the numerical values for parameters in Figure 6 were the same except that instead of given value of the size parameter, we set the value of timing parameter \((p)\) to 1.001.

As seen from these two figures, numerically the constraint expression behaves the way it is expected, i.e., it becomes more negative with decreasing values of timing parameter and/or with increasing values of size.
parameter. However, Equation (15), shows that the cost decreases with higher values of timing parameter. Also, the cost increases with the size parameter (with exception for very small $v$). The pull in opposite direction for the values of timing and size parameter is what sets up an interesting optimization problem.

The cutting plane algorithm used here approximates the dual of the original capacity expansion problem (of Equation (16)). Hence while solving the Master Problem of the cutting plane algorithm, 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 then be considered as the penalty (in terms of dollars per unit time) of not meeting the demand at the specified service level.

### 4.2.2 Optimization results

We applied the dual cutting plane algorithm described in Section 4.1 to Equation (16). While solving the sub problems of Figure 4, we added a dummy constraint of $p \leq 2$ to reduce the number of iterations
required for convergence. As seen from Table 1 below, the first iteration of algorithm minimizes the total cost (Equation (15)) subject to the constraints on the decision variables \((0 \leq p \leq 2, v \geq 1)\). The parameter values used for this hypothetical instance of the problem were: drift \((\mu) = 2\%\), volatility \((\sigma) = 20\%\), lead time \((L) = 2\) years, interest rate \((r) = 13\%\) and economies of scale parameter \((a) = 0.99\). The service level was assumed to be 95%, meaning that the shortages were limited to 5% of the total demand during the expansion cycle \((\delta = 0.05)\). The initial feasible point was \((p_0, v_0) = (1.4, 2)\). The successive iterations and the convergence of the cutting plane algorithm are summarized in Table 1. The level of accuracy used for all the numerical studies was up to 3 decimal places. See Appendix A for the details of computational method used. Since the Master Problem is a simple linear program, its optimal solution to that is obtained nearly instantaneously. However, the Sub Problem at each iteration involves the expression \(g(p, v)\), which includes integration of bivariate normal distribution functions. The Sub Problem at each iteration was solved by applying ‘NMinimum’ function of Mathematica 5.1 (Wolfram, 2004). The average computational time required to solve each Sub Problem is approximately 2 hours on a Intel© Pentium IV personal computer with Windows XP operating system and 1 GB of memory.

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Constraint Added ((z \leq f(p_k, v_k) + u g(p_k, v_k)))</th>
<th>Master Problem solution ((z_k, u_k))</th>
<th>Sub Problem ((p_k, v_k), \theta_k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(z \leq 0.921 - 0.02u)</td>
<td>(0.921, 0)</td>
<td>(2, 1.009), 0.223</td>
</tr>
<tr>
<td>2</td>
<td>(z \leq 0.223 + 2.027u)</td>
<td>(0.914, 0.341)</td>
<td>(1.954, 2.178), 0.687</td>
</tr>
<tr>
<td>3</td>
<td>(z \leq 0.5 + 0.553u)</td>
<td>(0.906, 0.734)</td>
<td>(1.990, 2.944), 0.821</td>
</tr>
<tr>
<td>4</td>
<td>(z \leq 0.646 + 0.238u)</td>
<td>(0.899, 1.060)</td>
<td>(1.151, 1.038), 0.853</td>
</tr>
<tr>
<td>5</td>
<td>(z \leq 0.7218 + 0.124u)</td>
<td>(0.893, 1.380)</td>
<td>(1.280, 1.530), 0.873</td>
</tr>
<tr>
<td>6</td>
<td>(z \leq 0.845 + 0.021u)</td>
<td>(0.883, 1.867)</td>
<td>(1.260, 1.630), 0.869</td>
</tr>
<tr>
<td>7</td>
<td>(z \leq 0.934 - 0.034u)</td>
<td>(0.878, 1.636)</td>
<td>(1.441, 2.059), 0.878</td>
</tr>
</tbody>
</table>

At the end of 7th iteration, the objective function value of the Master Problem is approximately equal to the optimal objective function value to the Sub Problem; hence, we stop and say that the cutting plane algorithm has converged. The optimal solution in this instance is the solution to the Sub Problem in the last iteration, that is, \(p^* = 1.441, v^* = 2.059\). This means that the service provider should start the new capacity expansion project when the demand hits 144.1% of the current capacity position (thereby allowing initial shortage of 44.1% over the current capacity position) and the expansion should be such that after the project is completed, the capacity available is 205.9% of current capacity position. The optimal value of the objective function \(f(p^*, v^*)\) is found to be 0.892 and the value of shortage violation expression \(g(p^*, v^*)\) is -0.002. The scaled penalty cost of not satisfying the service level \((u)\) is 1.636/year.
In a similar fashion the cutting plane algorithm was implemented for different values of the problem parameters and the results of these tests are summarized below. The default values of the problem parameters used were the hypothetical values above. Here, we discuss the effects of problem parameters on the optimal timing variable \( p \) and not on the optimal size variable \( v \) because there was no clear trend for the optimal size parameter as we varied other problem parameters one at a time.

**Effect of the drift parameter**

To test how the decision regarding the timing of the new expansions is affected due to change in the drift parameter of the demand process, the drift parameter values were varied keeping the other parameters values as default. Then values of drift parameter were tested in the cutting plane algorithm and optimal values of timing and size parameters were obtained.

Table 2: Effect of demand drift on optimal decision variables

<table>
<thead>
<tr>
<th>( \mu )</th>
<th>Optimal ( p^* )</th>
<th>Optimal ( v^* )</th>
<th>( f(p^<em>, v^</em>) )</th>
<th>( g(p^<em>, v^</em>) )</th>
<th>( u^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>1.441</td>
<td>2.058</td>
<td>0.892</td>
<td>-0.002</td>
<td>1.636</td>
</tr>
<tr>
<td>0.05</td>
<td>1.083</td>
<td>1.363</td>
<td>1.905</td>
<td>-0.037</td>
<td>3.774</td>
</tr>
<tr>
<td>0.08</td>
<td>0.989</td>
<td>1.011</td>
<td>4.218</td>
<td>-0.0004</td>
<td>10.000</td>
</tr>
<tr>
<td>0.1</td>
<td>0.956</td>
<td>1.632</td>
<td>17.232</td>
<td>-0.087</td>
<td>43.078</td>
</tr>
</tbody>
</table>

As the drift parameter increases, the optimal value of the timing variable \( p \) decreases. Recall that the decision variable \( p \) represents the initiation of the expansion project in the sense that we initiate the \((n+1)^{st}\) expansion project when the demand process hits the value \( pK_n \), where \( K_n \) was the capacity position after \( n \) expansions. This means that the increase in drift is prompting earlier initiation of the new expansion project. From Table 2, we also observe that the value of the objective function, which is the total cost of expansion, also increases as the drift parameter increases. This occurs because the infinite horizon cost of expansion increases with decreasing timing variable \( p \). Higher drift values for the demand process implies a high growth industry. Hence, we can see that for high growth industries the optimal timing parameter value is small and the cost of meeting the service level constraint is high.

**Effect of the volatility parameter**

Because demand fluctuations are common, it is critical to study the effects of this demand volatility on the decision variables.

From Table 3, we can see that as the volatility of the demand process increases, it forces the service provider to initiate expansion earlier and also the optimal size of the future expansions gets smaller. Hence for an industry where the demand experienced is highly fluctuating, it is optimal to start the capacity expansions
earlier and not wait for initial shortages to accumulate. The optimal expansion cost grows along with the demand volatility. A service provider in a more stable industry, however, has the luxury of delaying the expansion and can in fact tolerate initial shortages before the start of the expansion project. This trend of the optimal solution, though, is the opposite of what is observed in the financial options theory. There, the volatility of the stock price is considered favorable and can be exploited by the writer of the option because the price of the call increases with increase in volatility.

**Effect of the lead time length**

We also studied the effects of the length of the expansion lead time on the optimal starting time of the expansion project. From Table 4, longer time intervals to complete expansion projects require expansion with smaller initial shortages, in order to maintain the given service level. Note that the optimal size parameter does not increase monotonically with \( L \).

<table>
<thead>
<tr>
<th>( L )</th>
<th>Optimal ( p^* )</th>
<th>Optimal ( v^* )</th>
<th>( f(p^<em>, v^</em>) )</th>
<th>( g(p^<em>, v^</em>) )</th>
<th>( u^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.477</td>
<td>1.819</td>
<td>0.747</td>
<td>-0.02</td>
<td>0.993</td>
</tr>
<tr>
<td>2</td>
<td>1.441</td>
<td>2.058</td>
<td>0.892</td>
<td>-0.002</td>
<td>1.636</td>
</tr>
<tr>
<td>2.5</td>
<td>1.130</td>
<td>1.411</td>
<td>1.006</td>
<td>-0.025</td>
<td>2.524</td>
</tr>
<tr>
<td>4</td>
<td>0.995</td>
<td>1.006</td>
<td>0.963</td>
<td>-0.0013</td>
<td>3.471</td>
</tr>
<tr>
<td>5</td>
<td>0.997</td>
<td>1.004</td>
<td>0.962</td>
<td>-4.5e-5</td>
<td>3.498</td>
</tr>
</tbody>
</table>

**Effect of the allowed shortage**

Lastly, we observe the effect of allowed shortage value (\( \delta \)) on the optimal values of decision variables, more particularly, the optimal value of the timing variable. The numerical values for the other problem parameters used to study this case were: drift (\( \mu \))=5%, volatility (\( \sigma \))=20%, interest rate (\( r \))=10%, lead time (\( L \))=0.5%, and economies of scale parameter (\( a \))=0.9. These numerical values were selected from the Ryan (2004) numerical analysis. Table 5 shows that as we allow more shortages during the expansion cycle, the optimal timing variable value increases. A similar trend was observed in Ryan (2004) though we note that the service level equation in Ryan (2004) is different than ours (See Appendix B for details). On the other hand, when the service provider can allow larger total shortages during the expansion cycle, then the new expansions could be started later. Not surprisingly, relaxing the service level goal, reduces the optimal expansion cost.
Table 5: Effect of allowable shortage on optimal decision variables

<table>
<thead>
<tr>
<th>δ</th>
<th>Optimal $p^*$</th>
<th>Optimal $v^*$</th>
<th>$f(p^<em>, v^</em>)$</th>
<th>$g(p^<em>, v^</em>)$</th>
<th>$u^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>1.360</td>
<td>1.770</td>
<td>2.520</td>
<td>-0.023</td>
<td>2.157</td>
</tr>
<tr>
<td>0.02</td>
<td>1.396</td>
<td>1.796</td>
<td>2.517</td>
<td>-0.0635</td>
<td>1.762</td>
</tr>
<tr>
<td>0.03</td>
<td>1.447</td>
<td>1.856</td>
<td>2.381</td>
<td>-0.0327</td>
<td>1.554</td>
</tr>
<tr>
<td>0.04</td>
<td>1.483</td>
<td>1.8486</td>
<td>2.30</td>
<td>-0.0318</td>
<td>1.338</td>
</tr>
</tbody>
</table>

5 Conclusions and future work

If the demand drift and/or volatility is sufficiently low and/or if the the expansion lead times are short, the service provider can afford to wait longer before initiating the capacity expansion. This result is similar to the trend observed in Ryan (2004). However, while modifying the service level definition itself, our model extends the capacity expansion model to find the conditions under which the service provider could wait until after some initial shortages have been accumulated. Higher values for the shadow price of the service level constraint when the optimal timing parameter is smaller suggest that timing is driven more by the constraint than by cost minimization. The optimal size of expansion does not necessarily vary monotonically with changing model parameter. Based on these observations, we conjecture that timing is driven mainly by the service level constraint while expansion size is determined by economic considerations.

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. Secondly, in the current model, only capacity expansions were considered. There are practical examples 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 Computational method details

The software package Mathematica 5.1 (Wolfram, 2004) was used to obtain numerical results for the model. Evaluating Equation (14) involves integrating the bivariate normal distribution functions over infinite
regions. The method prescribed by Rose and Smith (1996) was used to code the service level expression. The authors set up a function $MVN[x, \mu, \text{var}]$ to calculate the $n$-dimensional multivariate normal distribution for the vector $x = [x_1, x_2, \cdots, x_n]$ defined everywhere in the $n$-dimensional real space, the mean vector $\mu = [\mu_1, \cdots, \mu_n]$ and symmetric positive-definite variance-covariance matrix $\text{var}$.

$$MVN[x, \mu, \text{var}] = (2\pi)^{-n/2} \sqrt{\det(\text{var}^{-1})} e^{\frac{1}{2}(x-\mu)^T \text{var} (x-\mu)}.$$ 

where $T$ represents transpose of the vector.

Then for a standard normal bivariate normal density function ($x = [x_1, x_2], n = 2$), with zero mean vector, variance elements unity, and correlation coefficient $\eta$, the $MVN$ function returns:

$$MVN[x, \mu, \text{var}] = \frac{e^{(x_1^2 - 2\eta x_1 x_2 + x_2^2)/(2\eta^2 - 1)} \sqrt{\frac{1}{1-\eta^2}}}{2\pi}.$$ 

This density function can then be integrated to find the probability distribution function.

**B Service level definition used in Ryan (2004)**

In Ryan (2004), the service level was defined as the total unsatisfied demand per unit of capacity over an expansion cycle. With notation similar to our model, for $t_n + L \leq t \leq T_{n+1} + L$, the shortage at time $t$ as a proportion of installed capacity was defined as:

$$S^n(t) = \max(P(t) - K_n, 0) / K_n.$$ 

And from this definition, the service level constraint was formulated as:

$$E_{t_n} \left[ \int_{\max(t_n+L,T_{n+1})}^{T_{n+1}+L} S^n(u)du \right] \leq \delta.$$ 

Note that because of this definition of the service level, the timing policy could be separated from the increment policy. That is, the timing parameter $p$ could be identified optimally, independent of the size parameter $v$, by applying the price of a European call option.

**References**


