The Impact of Customer Impatience on Production Control

Michael H. Veatch*

March 27, 2007

Abstract

Most analyses of make-to-stock production control assume that either all orders are eventually met (complete backordering) or that no customers are willing to wait (lost sales). We consider a more nuanced model of customer behavior, where the fraction of potential customers who place orders depends on the current backlog, and hence the lead time. A continuous one-part-type, single machine model with Markov modulated demand and deterministic production is considered. We show that the impact of customer impatience is captured by one quantity, the mean sojourn time in the backlog states. A simple procedure finds this quantity and the optimal policy, which has hedging point form. In applications, observing the durations of stockouts gives a practical method of incorporating the effect of customer impatience.

*Department of Mathematics, Gordon College, Wenham, MA 01984, (978) 867-4375, mike.veatch@gordon.edu

1 Introduction

Common formulations of the flow-rate control problem assume either that all customers are infinitely patient, but there is a cost associated with making them wait, or completely impatient and sales are lost if the order cannot be met from stock. For the single-part-type, single unreliable machine problem with constant demand, the complete backordering case is analyzed in Bielecki and Kumar (1988) and the lost sales case is studied in Hu (1995). Neither of these assumptions is completely satisfactory. In the complete backordering model, the penalty for making customers wait is a cost applied to the current number of parts backordered (the backlog). It is problematic to estimate this cost, even if it is linear in the backlog. There is usually no tangible cost associated with backlog (unless the contract includes a provision for a discount for delivery delay). The real costs of delivery delays are more likely to be loss of future sales, which are difficult to measure, and cancelled orders. The lost sales model requires only a unit profit parameter; however, in many contexts it is overly pessimistic to assume that customers have no waiting tolerance. One generalization of the lost sales model is to allow a limited backlog (Martinelli and Valigi 2004).

A more realistic model of customer behavior is considered in the companion paper Gershwin et al. (2004). If there is finished goods inventory on hand (surplus), customer orders are filled immediately. If there is a backlog, it is converted to a lead time which is quoted to the customer who either leaves immediately (balks) or decides to wait for their purchase. Customers have a distribution of lead time tolerances, so that the balking probability of a potential customer depends on the backlog when they arrive. In the context of flow-rate control, demand is continuous and this probability is interpreted as the fraction of potential demand that is not realized. Thus, customer behavior is modeled by a *defection function* giving the fraction of demand that "defects" as a function of the backlog. In some service systems, e.g., where people join a physical queue, arriving customers observe the backlog directly and make balking decisions based on it. However, the primary motivation for this model is manufacturing systems for which fairly accurate lead time quotes can be made.

Specifically, the model in Gershwin et al. (2004) assumes the machine is reliable and that a first-come-first-served queue discipline is used, so that leadtime is deterministic and proportional to backlog. Unlike Bielecki and Kumar, randomness in the model comes from demand, which is assumed to be Markov modulated with two levels. This demand model can be used as a rough approximation for Poisson demand or may be useful in modeling longer-term variations in purchase contracts or demand rates. See Fleming et al. (1987) and Perkins and Srikant (2001) for other models with Markov modulated demand.

The objective is to maximize long-run average profit, where revenue is diminished by customers who do not wait and the cost rate is linear in the surplus. The model allows an economic tradeoff to be made between using inventory to buffer against demand uncertainty and increasing lost sales for some of the more impatient customers. It is shown in Gershwin et al. (2004) that the optimal policy has the same hedging point form as the Bielecki and Kumar model. The maximum production rate is used until the surplus reaches a level called the hedging point. The production rate is then set to the demand rate and the surplus remains constant until demand changes to the higher level. When the defection function

is piece-wise constant, the stationary distribution of the system is found, allowing average profit to be expressed as a function of the hedging point and numerically optimized. General defection functions are approximated by a piece-wise constant function.

This paper provides a further analysis of the model in Gershwin et al. (2004). First, it is shown that customer behavior can understood by decomposing the system into the backlog dynamics, which depend on the defection function, and the surplus dynamics, which depend on the hedging point. The impact of customer behavior is captured by one quantity, the mean sojourn time in the backlog states, which depends on the demand model and production rate but not the hedging point. The profitability of defection functions can be completely ordered by this quantity. The numerical study in this paper uses a sigmoid for the defection function, as proposed in Tan and Gershwin (2004). However, because the defection function influences profit only through this mean sojourn time, the numerical insights apply to other defection functions. In fact, two simple models of customer behavior are constructed that are equivalent to a given defection function in the sense that, for a given demand model and production rate, they have the same mean sojourn time. One model has a constant defection function whenever there is a backlog, i.e., some fraction of customers have no patience. A second model assumes all customers wait until a backlog limit is reached, as in Martinelli and Valigi (2004).

One implication of this decomposition is that the optimal hedging point can be computed more efficiently. Instead of solving the full model repeatedly while searching for the optimal hedging point, the mean sojourn time is computed once. Then an equivalent model is used to compute average profit that replaces the backlog dynamics with a single state, simplifying the optimization.

More importantly, the decomposition provides a simple way of estimating the model parameters. Instead of estimating the defection function, one can estimate the mean sojourn time in the backlog states. In applications, the duration of stockouts is usually readily available. Simply computing their mean gives an estimate of the mean sojourn time, which fully captures the impact of the defection function on average profit. Estimating the mean duration of stockouts is vastly easier than estimating the defection function. Customers that balk because of the lead time may not be recorded. Even if they are recorded, estimating a function requires much more data than estimating a single quantity. The usual approach is to use some parameterized class of functions. More research would be needed to justify using a particular class of function.

Empirical studies of customer responses to lead times, such as Anderson et al. (2003) which studies a mail-order catalogue, have observed that customers are more likely to cancel their orders when the anticipated delay is longer but have not identified specific forms of defection functions. Customer impatience in telephone call centers, where the queue is invisible to the customer, has been studied extensively, e.g., Bolotin (1994) and Brown et al. (2005). It is observed in Whitt (2005) that call center performance is sensitive to the distribution of the customer abandonment time, which is not nearly exponential. When call centers announce the anticipated delay to customers, they make a balking decision as in our model, but may also abandon later; Whitt (1999) gives a queueing analysis.

The layout of the paper is as follows. Section 2 presents the flow control model with defection. The decomposition into backlog and surplus behavior is shown in Section 3 and used to construct simpler models in Section 4 and to find the optimal policy in Section 5. Numerical examples are presented in Section 6.

2 The defection model

We consider the one-part-type, one-machine make-to-stock production control problem analyzed in Gershwin et al. (2004). The demand rate at time t is denoted $d(t)$. It is governed by an exogenous Poisson process $D(t)$ on the states L (low) and H (high). When $D(t) = L$, $d(t) = \mu_L$ and transitions to H occur with rate λ_{LH}; when $D(t) = H$, $d(t) = \mu_H$ and transitions to L occur with rate λ_{HL}. Given a demand rate, demand is continuous and deterministic. We assume that the production capacity $\underline{u}$ is sufficient to meet the demand when it is low but insufficient when it is high, i.e., $\mu_L < \underline{u} < \mu_H$. The average demand rate is

$$E(d) = \frac{\lambda_{LH}}{\lambda_{HL} + \lambda_{LH}}\mu_H + \frac{\lambda_{HL}}{\lambda_{HL} + \lambda_{LH}}\mu_L. \tag{1}$$

The continuous inventory level at time t is $x(t)$, with $x < 0$ representing backlog. Assume that customers are served first-in-first-out. Arrivals at time t observe $x(t)$ and can infer their (deterministic) waiting time to be $x^-(t)/\underline{u}$, where $x^- = \max\{0, -x\}$. Customers have various waiting tolerances, represented by a *defection function*: $B(x)$ is the fraction of potential customers who choose not to order when the backlog is x. This function is assumed to be non-increasing and satisfies

$$\begin{aligned} 0 \le B(x) \le 1 \\ B(x) = 0 \quad \text{for} \quad x \ge 0 \\ \lim_{x \to -\infty} B(x) > \frac{E(d) - \underline{u}}{E(d)}. \end{aligned} \tag{2}$$

The second condition says that no customers defect when there is zero waiting time. The third, which can be rewritten $E(d)(1 - \lim_{x\to-\infty} B(x)) < \underline{u}$, says that for sufficiently large backlogs, the average order rate is less than capacity. This condition is required for stability. Note that $B(x) = 1$ for all $x < 0$ corresponds to a lost sales model (no customers wait) and $B(x) = 0$ corresponds to complete backordering (all customers wait).

The decision variable is the production rate at time t, $u(t)$. The dynamics of $x(t)$ are given by

$$\frac{dx(t)}{dt} = u(t) - d(t)(1 - B(x(t))) \tag{3}$$

at points where the derivative exists. The quantity $d(t)(1 - B(x(t)))$ is the rate at which sales take place, i.e., the demand rate minus the defection rate. There is a reward A per unit sold and an inventory holding cost rate g^+ per unit per time. Defection is the only penalty for backorders; there is no backlog cost rate. Let u denote the control policy $\{u(t)\}$. The

expected profit of policy u in the interval $[0, T]$ is

$$J^u(x, D; T) = E_{x,D} \int_0^T [Ad(t)(1 - B(x(t))) - g^+ x^+(t)]dt,$$

where $E_{x,D}$ denotes expectation with respect to the initial state $x(0) = x$, $D(0) = D$ and $x^+ = \max\{0, x\}$. Similarly, the long-run average profit of policy u is defined as

$$J^u = \liminf_{T\to\infty} \frac{1}{T} J^u(x, D; T).$$

The production control problem is

$$J^* = \max_u J^u \tag{4}$$

subject to (3), $u(t)$ non-anticipating with

$$0 \le u(t) \le \underline{u} \tag{5}$$

and $d(t)$ a Poisson process as defined above.

It is shown in Gershwin et al. (2004) that the optimal policy for (3) - (5) has hedging point form: For some $z \ge 0$,

$$u^*(x, D) = \begin{cases} \mu_L & \text{if } x = z \text{ and } D = L \\ \underline{u} & \text{otherwise.} \end{cases} \tag{6}$$

The policy on the transient states $x > z$ does not affect the average profit. Finding the stationary distribution of (x, D) under a hedging point policy for general $B(x)$ appears to require numerical methods. Gershwin et al. (2004) finds this distribution when $B(x)$ is piece-wise constant and the recurrent states are bounded below. For a lower bound to exist, we must strengthen (2) to

$$B(x) > \frac{\mu_H - \underline{u}}{\mu_H} \quad \text{for some } x. \tag{7}$$

We can interpret (7) as saying that even in the high demand state enough of the customers are impatient that the system is stable. The minimum inventory level that will be reached is $x^* = \max\{x : \underline{u} - \mu_H(1 - B(x)) \ge 0\}$ in the following sense. From (3), $dx(t)/dt \ge 0$ for $x(t) = x^*$ (since $\mu_H > \mu_L$). Hence, if $x(0) \ge x^*$ then $x(t) \ge x^*$ for all t. Note that $x^* \le 0$. Numerical tests suggest that J is convex in z (as is known for the complete backordering model) and the optimal z is readily found by searching. However, the stationary distribution must be computed for each z. We introduce a decomposition so that this distribution only need be computed once. It also provides a complete ordering of defection functions.

3 Decomposition of backlog and surplus behavior

Consider a policy (6) with hedging point z. Since $x^* = 0$ is just a lost sales model, we assume in this section that $x^* < 0$. Define the mean sojourn times in backlog and nonbacklog states, respectively, as

$$\begin{aligned} \tau_- &= E_{0,H} \min\{t > 0 : x(t) \geq 0\} \\ \tau_+ &= E_{0,L} \min\{t > 0 : x(t) < 0\}. \end{aligned}$$

From now on we consider only the stationary distribution of $x(t)$ and $D(t)$ under the given policy and let X, D, and the demand rate d be random variables with this stationary distribution. Let

$$\begin{aligned} p_z &= P(X = z, D = L) \\ p^* &= P(X = x^*, D = H) \end{aligned}$$

and $f_D(x)$ be the marginal density in demand state D. Note that the system can enter the backlog states only from $(0, H)$ and the non-backlog states only at $(0, L)$. Thus the process $Y(t) = 1_{\{x(t)<0\}}$, which keeps track of whether or not there is a backlog, is a semi-Markov process; its mean times between transitions are τ_- and τ_+. We can define a renewal process (on Y or the original system) with a renewal occuring upon hitting state $(0, L)$ at the end of each backlog period. The mean renewal time is the sum of the two sojourn times. The quantity τ_- is important because average profit depends on the defection function only through τ_-.

Lemma 1 *Consider two systems using the same hedging point policy but with different defection functions, with $x^* < 0$ for both. If τ_- is the same for both systems, then they have the same stationary distribution on $X \in [0, \infty)$.*

Proof. Let $f_D^+(\cdot)$ and p_z^+ be the density and mass conditioned on the event $\{X \geq 0\}$. Note that neither these nor τ_+ depend on $B(\cdot)$ because there is a renewal upon entry into $[0, \infty)$. The stationary distribution of $Y(t)$ is

$$P(Y = 0) = P(X \geq 0) = \frac{\tau_+}{\tau_- + \tau_+}. \tag{8}$$

Since

$$\begin{aligned} f_D(x) &= P(X \geq 0) f_D^+(x) \\ p_z &= P(X \geq 0) p_z^+ \end{aligned} \tag{9}$$

the lemma follows. ■

The lemma implies that two systems with different $B(\cdot)$ but the same τ_- have the same average holding cost, since $E(X^+) = P(X \geq 0)E(X^+|X \geq 0)$. The average profit is

$$\begin{aligned} J &= A[E(d) - E(dB(X))] - g^+E(X^+) \\ &= A[\underline{u}(1-p_z) + \mu_L p_z] - g^+E(X^+) \\ &= A[\underline{u} - (\underline{u} - \mu_L)p_z] - g^+E(X^+). \end{aligned} \tag{10}$$

The second equality holds because, for stable systems, we can account for revenue in terms of service instead of arrivals. By (8) and (9), p_z and revenue depend on $B(\cdot)$ only through τ_-. We have established the following theorem.

Theorem 1 *Consider two systems using the same hedging point policy but with different defection functions, with $x^* < 0$ for both. If τ_- is the same for both systems, then they have the same average profit J.*

The quantity τ_- can be used to rank the set of all defection functions. A customer population with smaller τ_- is less patient and will have less profit (for a given demand model, production rate, and hedging point). Also, applying the decomposition argument again, τ_- does not depend on the hedging point, making it easier to estimate in practice. In Section 4 we show how to construct simple defection functions that have the same τ_-. In Section 5 we use τ_- to simplify the computation of optimal policies.

4 Equivalent balking and limited backlog models

Two simple defection functions are the *balking* model

$$B(x) = \begin{cases} b, & x < 0 \\ 0, & x \geq 0, \end{cases} \tag{11}$$

i.e., a fraction b of customers have no patience, and the *limited backlog* model

$$B(x) = \begin{cases} 1, & x \leq x^* \\ 0, & x > x^*, \end{cases} \tag{12}$$

i.e., all customers place orders subject to the backlog being limited to x^*. Note that this definition of x^* is a special case of the lower bound defined in Section 2. As b decreases from 1, τ_- for the balking model increases continuously from 0. Hence, for a general defection function, there is a balking model with the same τ_-. The same is true of the limited backlog model as x^* is varied. These models are equivalent to general defection models with the same τ_- in the sense that, given a hedging point, they have the same average profit.

To find the equivalent balking rate b for (11), first compute $p_0 = P(x = 0) = P(x \geq 0)$ under the zero hedging point policy as in Gershwin et al. (2004). Under this policy, $\tau_+ =$

$1/\lambda_{LH}$ and from (8)

$$\tau_- = \frac{1-p_0}{p_0\lambda_{LH}}. \tag{13}$$

Since (13) also holds for the balking model, requiring the two models to have the same τ_- implies that they also have the same p_0. Equating the average service rate and average demand rate for the balking model,

$$\underline{u}(1-p_0) + \mu_L p_0 = \mu_L p_0 + [\mu_H P(D=H) + \mu_L(P(D=L) - p_0)](1-b).$$

Solving for b and using (1),

$$b = \frac{E(d) - \underline{u} + (\underline{u} - \mu_L)p_0}{E(d) - \mu_L p_0}. \tag{14}$$

Now consider the limited backlog model (12). To find the equivalent backlog limit x^*, equate p_0 for these two models. For simplicity, we use the same notation as before, e.g., p_0, to refer to the limited backlog model. Following Gershwin et al. (2004), the rate of change of $x(t)$ when $x^* < x(t) < 0$ in the high (H) and low (L) demand states, respectively, is

$$\Delta^H = \underline{u} - \mu_H \tag{15}$$

$$\Delta^L = \underline{u} - \mu_L \tag{16}$$

The stationary distribution satisfies the differential equations

$$\Delta^H \frac{df_H(x)}{dx} = -\lambda_{HL} f_H(x) + \lambda_{LH} f_L(x), \quad x^* < x < 0$$

and a similar equation for f_L. Flow balance across the point x requires

$$f_L(x) = -\frac{\Delta^H}{\Delta^L} f_H(x). \tag{17}$$

Combining these gives a single differential equation $df_H(x)/dx = \alpha f_H(x)$ with solution

$$f_H(x) = ce^{\alpha x}, \quad x^* < x < 0$$

$$\alpha = -\frac{\lambda_{HL}}{\Delta^H} - \frac{\lambda_{LH}}{\Delta^L}. \tag{18}$$

Still assuming $z = 0$, the balance equation at $(0, L)$ yields

$$c = -\lambda_{LH} p_0 / \Delta^H$$

and at (x^*, H) yields

$$p^* = -c\Delta^H e^{\alpha x^*}/\lambda_{HL} = p_0 \frac{\lambda_{LH}}{\lambda_{HL}} e^{\alpha x^*}.$$

The normalization equation that determines c is

$$p^* + p_0 + \int_{x^*}^{0} [f_H(x) + f_L(x)]dx = 1. \tag{19}$$

Consider two cases. If $\alpha \neq 0$, (19) becomes

$$p^* + p_0 - p_0\kappa(1 - e^{\alpha x^*}) = 1$$

or

$$x^* = \frac{1}{\alpha} \ln \left(\frac{1 - p_0 + p_0\kappa}{p_0(\lambda_{LH}/\lambda_{HL} + \kappa)} \right), \tag{20}$$

where

$$\kappa = \frac{\lambda_{LH}}{\alpha} \left(\frac{\Delta^L - \Delta^H}{\Delta^L \Delta^H} \right).$$

If $\alpha = 0$, (19) becomes

$$p^* + p_0 + p_0\Gamma x^* = 1$$

or

$$x^* = \frac{1 - p_0(1 + \lambda_{LH}/\lambda_{HL})}{p_0\Gamma}, \tag{21}$$

where

$$\Gamma = \lambda_{LH} \left(\frac{\Delta^L - \Delta^H}{\Delta^L \Delta^H} \right).$$

If we use p_0 from the original model in (20) or (21), the resulting x^* gives an equivalent limited backlog model.

Equations (14), (20), and (21) assume that the original model has $x^* < 0$. If $x^* = 0$ in the original model, then the equivalent balking model has $b = 1$ and the equivalent limited backlog model has $x^* = 0$.

5 Optimization

The decomposition of Section 3 can be used to simplify the calculation of the optimal hedging point. The method in Gershwin et al. (2004) uses search, computing the stationary distribution for some z at each iteration. Instead, we compute p_0 once using the method in Gershwin et al. (2004) with $z = 0$. Combined with (13) for τ_-, this fully describes the behavior in the backlog states. Then, for each z, we analyze the surplus behavior using an equivalent model where $x^* = 0$ and state $(0, H)$ has a transition rate $1/\tau_-$ into state $(0, L)$. This *surplus model* has the same average cost as the original model for a given z.

Equations (15)-(19) apply to the surplus model, except that the densities are on $0 < x < z$. For simplicity we use the same notation as before, namely $f_H(x)$, $f_L(x)$, p_z, p^*, and c, to refer to the surplus model. The balance equation at (z, L) is

$$p_z = -c\Delta^H e^{\alpha z}/\lambda_{LH} \tag{22}$$

and at $(0, H)$ is

$$p^* = P(x = 0) = -c\Delta^H \tau_-.$$

Using (19) but integrating over $[0, z]$,

$$c = \left[-\Delta^H(\tau_- + e^{\alpha z}/\lambda_{LH}) + \left(\frac{\Delta^L - \Delta^H}{\Delta^L} \right) \chi \right]^{-1}, \tag{23}$$

where

$$\chi = \begin{cases} (e^{\alpha z} - 1)/\alpha, & \alpha \neq 0 \\ z, & \alpha = 0. \end{cases}$$

Finally,

$$E(X^+) = \int_0^z x[f_H(x) + f_L(x)]dx + zp_z = c \left(\frac{\Delta^L - \Delta^H}{\Delta^L} \right) Q + zp_z, \tag{24}$$

where

$$Q = \begin{cases} [(\alpha z - 1)e^{\alpha z} + 1]/\alpha^2, & \alpha \neq 0 \\ z^2/2, & \alpha = 0. \end{cases}$$

Thus, (10), (22), (23), and (24) give average profit. Differentiation with respect to z yields an equation that can be solved numerically for the hedging point.

6 Numerical examples

This section explores the sensitivity of the mean sojourn time τ_- to customer behavior. The sensitivity of the optimal hedging point to τ_- is also tested. The parameter values used are listed in Table 1. Case 1, taken from Gershwin et al. (2004), has a traffic intensity $E(d)/\underline{u} = 1.5$ and Case 2 has an intensity of approximately 1.1. As in Tan and Gershwin (2004) and Gershwin et al. (2004), a sigmoid, i.e., a function of the form $1/(1 + e^x)$, is used for the defection function. However, the numerical insights are not sensitive to the shape used.

A useful parameterization of a sigmoid is

$$B(x) = \frac{1}{1 + [(1 - \varepsilon)/\varepsilon]^{(-x/\eta+1)/v}}, \quad x < 0.$$

First, note that $\eta < 0$ is the median backlog tolerance, i.e., $B(\eta) = 0.5$. Setting $\varepsilon = .025$, the interval $[B^{-1}(1 - \varepsilon), B^{-1}(\varepsilon)]$ contains 95% of customer backlog tolerances and can be

Case	A	g^+	$\underline{u}$	μ_H	μ_L	λ_{HL}	λ_{LH}
1	3	0.1	0.6	1.5	0.3	.05	.05
2	6	0.1	0.6	1.0	0.3	.05	.05

Table 1: Parameter values

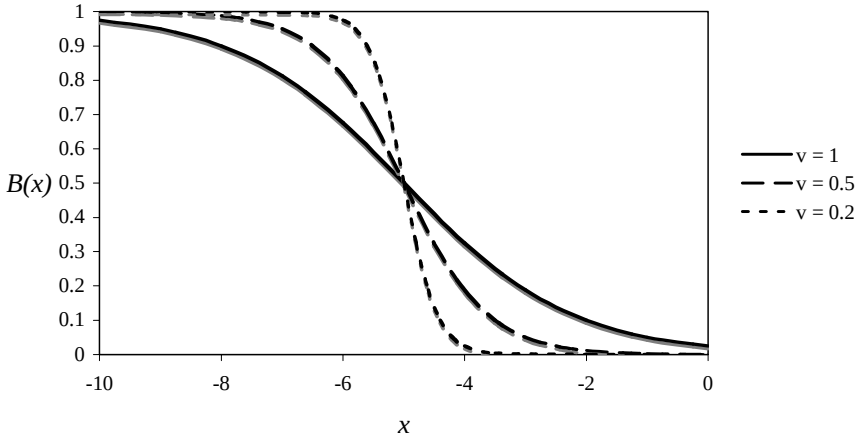

Figure 1: Effect of v on defection function, $\eta = -5$.

used to measure the variability in backlog tolerance. But

$$v = \frac{\eta - B^{-1}(1-\varepsilon)}{-\eta} = \frac{B^{-1}(\varepsilon) - B^{-1}(1-\varepsilon)}{-2\eta}.$$

Thus, v is a natural choice for a unitless measure of variability. Figure 1 illustrates the effect of v. A different parameterization is used in Gershwin et al. (2004), namely

$$B(x) = \frac{1}{1 + e^{\gamma(x-\eta)}}, \quad x < 0$$

where $\gamma = \ln((1-\varepsilon)/\varepsilon)/(-\gamma v)$.

Figures 2 and 3 show mean sojourn time as a function of η and v for Case 2. As $\eta \to 0$, $\tau_- \to 1/\lambda_{HL} = 20$. Because the system is overloaded, defection is required for stability. Thus, as $\eta \to -\infty$, $\tau_- \to \infty$. For traffic intensities less than one, τ_- approaches the mean sojourn time for the system without defection. Although τ_- appears linear over the range of η shown, we expect it to be nonlinear and convex. Note that there is little sensitivity to v; variability in customer backlog tolerance has little effect.

For the balking model, mean sojourn time is more sensitive to the balking rate (Figure 4). The vertical asymptote represents the stability condition (2). For the limited backlog model, the sensitivity to x^* (Figure 5) is comparable to the sensitivity to η.

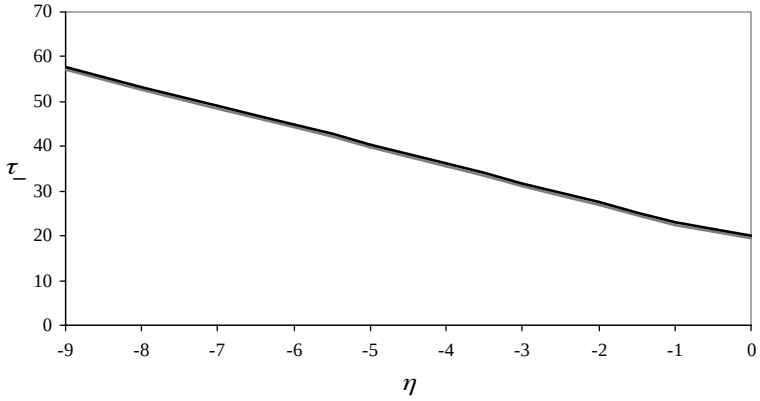

Figure 2: Mean sojourn time vs. η, Case 2 with $v = 1$.

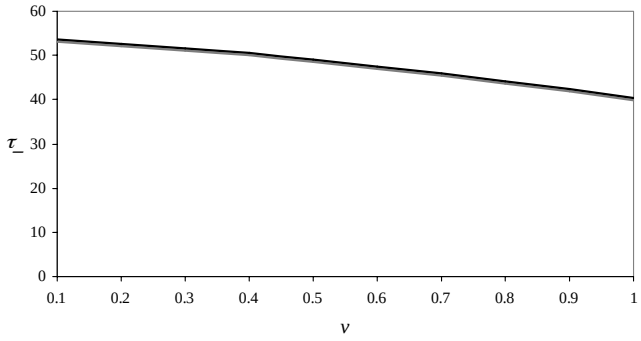

Figure 3: Mean sojourn time vs. v, Case 2 with $\eta = -5$.

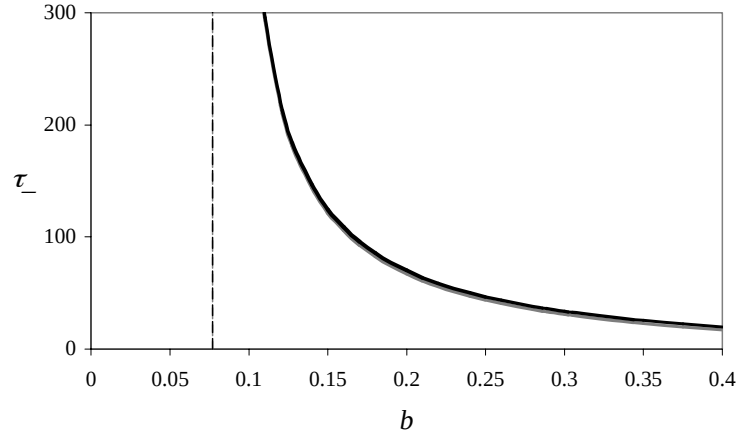

Figure 4: Mean sojourn time for balking model, Case 2.

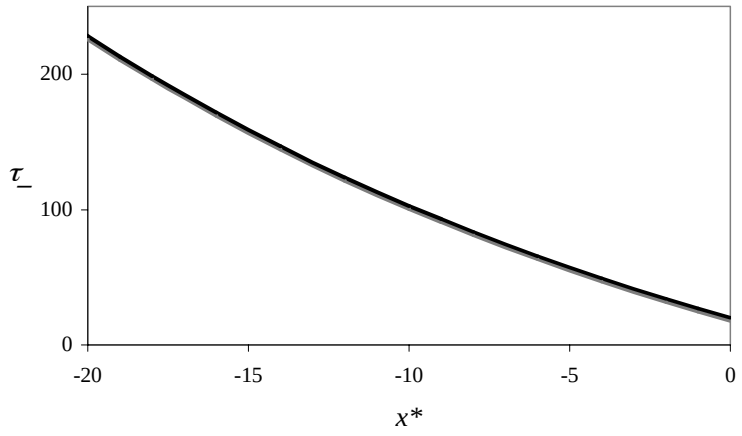

Figure 5: Mean sojourn time for limited backlog model, Case 2.

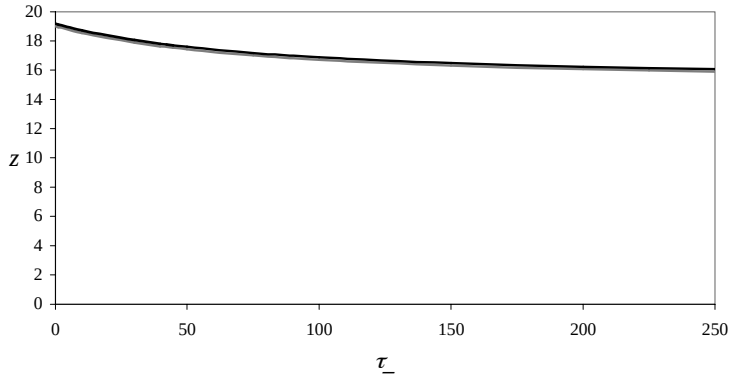

Figure 6: Optimal hedging point vs. mean sojourn time, Case 2.

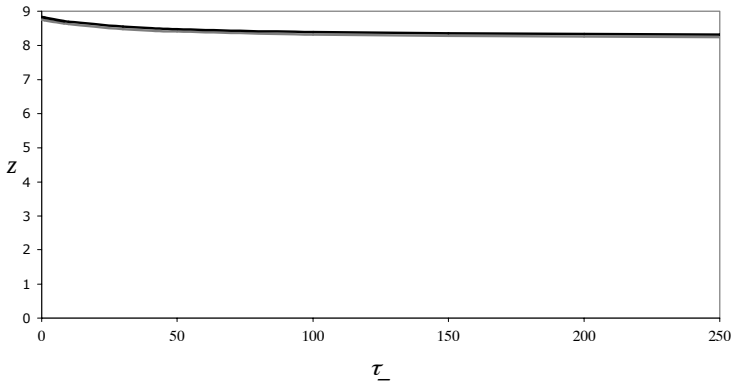

Figure 7: Optimal hedging point vs. mean sojourn time, Case 1.

The sensitivity of the optimal hedging point to τ_- is shown in Figure 6. The optimal hedging point makes a tradeoff between throughput and hedging costs, both of which are increased by hedging. From (10), the throughput effect on profit is $Ap_z(\underline{u}-\mu_L)$. For systems with $\rho > 1$, p_z will tend to be small (it goes to zero as $z \to \infty$), making the optimal hedging point fairly insensitive to τ_-. In Case 2, when $\tau_- = 50$ and the optimal hedging point is used p_z is only .07. Case 1, with a traffic intensity of 1.5, shows less sensitivity (Figure 7). Optimal profit is somewhat more sensitive to τ_-; over the range shown in Figure 6 it increases 36%.

7 Conclusion

This paper demonstrates that, for a simple model of production control, the impact of customer impatience on lost sales can be captured by a single parameter, even when customer behavior is a function of the backlog. This key observation leads to more efficient computation of the optimal hedging point. However, perhaps the most significant implication is that customer behavior can be modeled by measuring a single, readily observed quantity: a mean sojourn time. Specifically, the duration of stockout periods could be observed and their mean (τ_-) estimated. This quantity depends on production and demand rates but *not* on the hedging point policy.

Numerical tests show that τ_- is fairly insensitive to the variability in customer backlog tolerance, depending primarily on the mean backlog tolerance. Thus, the hedging point can be set and average profit estimated in many scenarios based solely on median customer behavior. Tests also show that as the system becomes overloaded, the sensitivity of the optimal hedging point to customer impatience decreases, because hedging does little to prevent lost sales in an overloaded system.

These conclusions would be quite different for a model that included lost sales and backlog costs, such as Martinelli and Valigi (2004). A second quantity, average backlog during backlog periods, would be needed to describe the system behavior during backlogs. As a result, there would not be a complete ordering of customer defection functions; which customer behavior allows the most profit would depend on the relative magnitude of the lost sales and backlog costs. However, the analysis based on decomposing system behavior into backlog and surplus would still apply. The decomposition also applies to a single-part-type, single-machine make-to-stock queue, being a consequence of the skip-free property of the underlying Markov chain.

For systems with multiple part types or multiple machines, the beginning of a backlog period is not a renewal of the Markov process and the decomposition of backlog and surplus behavior does not apply. For these systems, the optimal control can be qualitatively different for different customer behavior. In Veatch and Wein (1996), for example, a two-class make-to-stock queue is analyzed under lost sales and complete backordering. The optimal switching curves between serving class 1 and class 2 are shown numerically to have different shapes for the lost sales and complete backordering cases. Other customer behavior models could lead to different optimal policies.

8 Acknowledgements

I would like to thank Stan Gershwin and Baris Tan for their helpful suggestions on this work.

References

Anderson, E., G. J. Fitzsimons, and D. Simester (2003). Mitigating the cost of stockouts. Technical report.

Bielecki, T. and P. R. Kumar (1988). Optimality of zero-inventory policies for unreliable manufacturing systems. *Oper. Res. 36*(4), 532–541.

Bolotin, V. (1994). Telephone circuit holding time distributions. In J. Labertoulle and J. W. Roberts (Eds.), *Proc. International Teletraffic Congress, ITC 14*, Amsterdam, The Netherlands, pp. 125–134. North-Holland.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao (2005). Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association 100*(469), 36–50.

Fleming, W., S. Sethi, and H. Soner (1987). An optimal stochastic production planning problem with randomly fluctuating demand. *SIAM Journal of Control and Optimization 25*, 1495–1502.

Gershwin, S. B., B. Tan, and M. H. Veatch (2004). Production control with backlog-dependent demand. Working paper available at http://faculty.gordon.edu/ns/mc/Mike-Veatch/index.cfm.

Hu, J. (1995). Production rate control for failure prone production with no backlog permitted. *IEEE Trans. Auto. Control 40*(2), 291–295.

Martinelli, F. and P. Valigi (2004). Hedging point policies remain optimal under limited backlog and inventory space. *IEEE Trans. Auto. Control 49*(10), 1863–1869.

Perkins, J. and R. Srikant (2001). Failure-prone production systems with uncertain demand. *IEEE Trans. Auto. Control AC-46*, 441–449.

Tan, B. and S. B. Gershwin (2004). Production and subcontracting strategies for manufacturers with limited capacity and volatile demand. *Annals of Operations Research 125*, 205–232.

Veatch, M. and L. Wein (1996). Scheduling a make-to-stock queue: Index policies and hedging points. *Oper. Res. 44*, 634–647.

Whitt, W. (1999). Improving service by informing customers about anticipated delays. *Management Science 45*(2), 192–207.

Whitt, W. (2005). Engineering solution of a basic call-center model. *Management Science 51*(2), 221–235.