This paper describes a control technique for regulating the waiting times of jobs in a discrete manufacturing process.

The technique is based on the second method of Lyapunov, which has been extensively used for deterministic processes. Two illustrations of the method are included.

Experimental evidence of the effectiveness of the technique is indicated.

A technique to control waiting time in a queue

by S. Shapiro

In a discrete manufacturing process, it is important to control the waiting times of jobs in the queues associated with the process. For example, jobs arrive randomly at a machine shop and are processed in turn, the processing time also being a random variable. If the average processing time is greater than the average time between arrivals, it is apparent that a queue will build up. Even if, on the average, jobs are processed faster than they arrive, the random nature of the process may cause a queue to form at some time. The time a job spends waiting in the queue, also a random variable, causes difficulty in meeting production schedules.

The associated control problem assumes one of several forms, e.g., the amount of overtime on a particular shift to be authorized on the basis of the number of jobs waiting.

second method of Lyapunov

This paper applies the second method of Lyapunov,[1] extensively used in control problems associated with deterministic processes, to discrete manufacturing processes of the above type.

Let V denote a function of a state variable describing the system, and $\dot{V}$ denote the derivative of V with respect to time. The second method of Lyapunov depends upon the fact that (subject to modest mathematical requirements[2] usually satisfied in practice), the system is stable if $V = \dot{V} = 0$ for values of the control variable corresponding to equilibrium, and $V > 0$ and $\dot{V} < 0$ otherwise. A function V with the latter properties is called a Lyapunov function. Note that the system, if not at equilibrium, approaches equilibrium at a rate proportional to $|\dot{V}|$. In essence,

usually an appropriate non-negative function V is selected, and $\dot{V}$ is kept negative through control action, thus obtaining a Lyapunov function. Control action designed to minimize $\dot{V}$ optimizes the rate at which the system approaches equilibrium.

It should be noted that the technique is applicable to non-linear control problems, since linearity of the system is not required.

control technique

To illustrate the control technique proposed here, suppose it is desired to control a machining operation so that the waiting time of a job is W_d. Consider the quantity $(W - W_d)^2$, the squared difference between the actual time a job waits, W, and the desired waiting time, W_d. Since this quantity is a random variable, a function V may be defined as the expected value of $(W - W_d)^2$,

$$V = E((W - W_d)^2), \tag{1}$$

which satisfies the non-negative requirement of a Lyapunov function.[3] With this definition of V,

$$\dot{V} = \frac{d}{dt} E((W - W_d)^2).$$

Suppose a control technique is desired that gives the number of hours h of overtime to be worked for a given shift. Calculating the value of $\dot{V}$ for each value of h, that value of h is used that minimizes $\dot{V}$. In general, this value of h depends on the state of the process at the time the decision is made.

The function V is non-negative and is zero if, and only if, $W = W_d$. A technique which tends to make V as small as possible, in one sense tends to make W approach W_d. The control technique proposed here essentially involves manipulation of the process parameters so that V is always decreasing at a maximum rate. This is accomplished by focusing attention on $\dot{V}$ and, through appropriate control action, minimizing $\dot{V}$.

In order to obtain a suitable form of $\dot{V}$, first, a more convenient expression for V is found. Suppose the control is to be based upon some statistic n, as, for example, the number of objects in the system, or the average waiting time of the last five objects to depart, etc. It is assumed that the conditional expectation and conditional variance of W, given n, denoted respectively by $E(W|n)$ and $\mathrm{Var}(W|n)$, are known. From standard identities,[4]

$$E((W - W_d)^2) = \mathrm{Var}\,(W) + (W_d - E(W))^2,$$

$$\mathrm{Var}\,(W) = E(\mathrm{Var}\,(W \mid n)) + \mathrm{Var}\,(E(W \mid n)),$$

and

$$(W_d - E(W))^2 = E((W_d - E(W \mid n))^2),$$

so that Equation 1 may be rewritten to obtain

$$V = E(\mathrm{Var}\,(W \mid n)) + \mathrm{Var}\,(E(W \mid n)) + E((W_d - E(W \mid n))^2).$$

A suitable form of $\dot{V}$ may now be obtained from the latter expression:

$$\dot{V} = \frac{\partial}{\partial E(n)} (E(\text{Var}\,(W \mid n)) + \text{Var}\,(E(W \mid n))$$

$$+ E((W_d - E(W \mid n))^2)) \frac{dE(m)}{dt}. \tag{2}$$

Implied in this derivation of $\dot{V}$ is the existence of all expectations taken and their derivatives.

Now, we find the value of the control variable μ that minimizes $\dot{V}$. For each value of the observed statistic n there is a value of μ that minimizes $\dot{V}$ (assuming that $\dot{V}$ is always negative).

This control technique is completely general—no assumptions have been made as to the nature of the arrival or service distributions. The technique could be extended for the case of a multi-server system with several control variables.

single server, Poisson input, and exponential service

An application of the technique to a single-server system with Poisson input and exponential service is now described. Let the observed statistic n in a single-server queuing system be the number of objects in the system, and let the control variable μ be the mean service rate. It also is assumed that the control action is continuous, i.e., the number of objects is continuously observed, and the service rate is varied according to the method previously outlined.

Since n is exactly known at any instant of time, $E(W \mid n)$ is some function of n which is also exactly known, and consequently $\text{Var}(E(W \mid n)) = 0$.

The well-known formulas for a single queue with Poisson input and exponential service,[5]

$$E(W \mid n) = \frac{n+1}{\mu}, \quad \text{Var}\,(W \mid n) = \frac{n+1}{\mu^2}, \quad \frac{dE(n)}{dt} = \lambda - \mu + \mu P_0(t),$$

where μ denotes the mean service rate, λ the mean arrival rate and $P_0(t)$ the probability of an empty system at time t, are used with Equation 2 to obtain

$$\dot{V} = \frac{\partial}{\partial E(n)} \left(E\left(\frac{n+1}{\mu^2}\right) + E\left(\left(W_d - \frac{n+1}{\mu}\right)^2\right)\right)(\lambda - \mu + \mu P_0(t))$$

$$= \left(\frac{3}{\mu^2} - \frac{2W_d}{\mu} + \frac{2E(n)}{\mu^2}\right)(\lambda - \mu + \mu P_0(t)).$$

But n is known at any instant of time and is equal to, say, n_t. Thus,

$$E(n) = E(n_t) = n_t$$

and

$$P_0(t) = \begin{cases} 1 & \text{if} \quad n_t = 0 \\ 0 & \text{if} \quad n_t \neq 0 \end{cases}$$

so that

$$\dot{V} = \left(\frac{3}{\mu^2} - \frac{2W_d}{\mu} + \frac{2n_t}{\mu^2}\right)(\lambda - \mu - \mu P_0(t)).$$

Since W_d, λ, and n_t are known, the value of μ that minimizes $\dot{V}$ is easily found. If this μ is denoted by μ_m, then

$$\mu_m = \begin{cases} \dfrac{2(3 + 2n_t)\lambda}{3 + 2n_t + 2\lambda W_d} & \text{if } n_t \neq 0 \\ \dfrac{3}{W_d} & \text{if } n_t = 0, \end{cases} \tag{3}$$

and for these values of μ_m, $\dot{V}$ is always negative.

To illustrate with a numerical example: Let $W_d = 20$, $\lambda = 0.2$, and μ be constrained to be within the interval 0.16 to 0.30. Using Equation 3, a table is now constructed that gives μ_m for each value of n_t. If the calculated μ_m falls outside the interval, then the closest value of μ within the interval is chosen and Table 1 results.

single servers, constrained total service

We now consider a second example comprised of three completely independent single-server queuing systems with the total service supplied by the three servers constrained to equal some constant, C. Poisson input and exponential service is assumed as in the previous example. The notation λ_i, μ_i, and n_i is used to denote the arrival rate, service rate, and the number of jobs in each of the three subsystems ($i = 1, 2, 3$). The condition of constrained total service is expressed by $\mu_1 + \mu_2 + \mu_3 = C$.

The technique is now used to control the total system, so that the waiting time in each individual queue is made to approach some desired waiting time.

Define

$$V = \sum_{i=1}^{3} E((W_i - W_{id})^2), \tag{4}$$

where W_i and W_{id} denote, respectively, the actual and desired waiting times. The following expression for $\dot{V}$ is derived from Equation 4 in the same manner in which Equation 2 was obtained from Equation 1.

$$\dot{V} = \sum_{i=1}^{3} \frac{\partial}{\partial E(n_i)} (E(\text{Var}\,(W_i \mid n_i)) + \text{Var}\,(E(W_i \mid n_i)) + E((W_{id} - E(W_i \mid n_i))^2)) \frac{dE(n_i)}{dt}.$$

Table 1

n_t	μ_m (*calculated*)	μ_m (*constrained*)
0	0.150	0.160
1	0.154	0.160
2	0.187	0.187
3	0.212	0.212
4	0.232	0.232
5	0.247	0.247
6	0.261	0.261
7	0.272	0.272
8	0.282	0.282
9	0.289	0.289
10	0.297	0.297
11	0.303	0.300
12	0.309	0.300

As in the previous example, the formulas for Poisson input and exponential service are now introduced in the latter equation for $\dot{V}$. In addition, the method of Lagrangian multipliers is used, and the term $\delta(\mu_1 + \mu_2 + \mu_3 - C)$ is subtracted from $\dot{V}$. If the constraint $\mu_1 + \mu_2 + \mu_3 = C$ is satisfied, this term is zero. The Lagrangian multiplier δ is used to ensure that the constraint is satisfied in the solution. The following equation for $\dot{V}$ results:

$$\dot{V} = \sum_{i=1}^{3} \left(\left(\frac{3}{\mu_i^2} - \frac{2W_{id}}{\mu_i} + \frac{2n_i}{\mu_i^2} \right) (\lambda_i - \mu_i + \mu_i P_0(t)) \right) - \delta(\mu_1 + \mu_2 + \mu_3 - C).$$

Assuming that $n_i \neq 0$, the values μ_{im} of μ_i minimizing $\dot{V}$ are found

to be

$$\mu_{im} = \frac{2(3 + 2n_i)\lambda_i}{2\lambda_i W_{id} + (1 + \delta)(3 + 2n_i)}, \qquad i = 1, 2, 3,$$

and δ is chosen so that

$$\mu_{1m} + \mu_{2m} + \mu_{3m} = C.$$

If some of the n_i's are equal to zero, a solution is obtained in a similar manner.

concluding remarks

The technique outlined gives a solution to the problem of optimizing the service rate(s) in a queuing system according to the given criteria. It should be noted that this solution is suboptimal in several respects. Whereas the stated objective of the procedure is to make $W \to W_d$, the procedure actually makes $E((W - W_d)^2) \to 0$. The minimization of $E((W - W_d)^2)$ tends to weight deviations on either side of W_d equally, which implies that the cost of an item finishing ahead of schedule is the same as the cost of an item late by an equal time.

In the examples, the optimal service rate is based on the system state. If this state changes before all the items in the queue are in service (which is almost certain to happen), the effect on the optimal strategy previously determined is not clear. Also, since it is usually impossible to continuously vary the service rate, the performance of the system is reduced.

Even though there are these difficulties, experience with the technique suggests that it will yield a workable solution where other procedures are unavailable. In an experiment simulating a control system incorporating the technique, the mean square deviations of the actual waiting times W from the desired waiting times W_d were markedly reduced. The variance of W was reduced in some cases by a factor greater than 2. This suggests that the technique may effectively be employed to increase the reliability of processing jobs in a manufacturing facility within the desired time. It is conjectured that additional work on reducing the effect of some of the aforementioned limitations might further refine the technique.

CITED REFERENCES AND FOOTNOTES

1. An excellent discussion of Lyapunov's second method can be found in R. E. Kalman and J. E. Bertram, "Control system analysis and design via the second method of Lyapunov," *Journal of Basic Engineering*, June, 1960. This paper discusses the method and gives several applications. There is also a discussion of the concept of stability.
2. For details, see the reference of Footnote 1.
3. Although the variables V and W are functions of time and could be written $V(t)$ and $W(t)$, time dependencies are omitted throughout for the sake of notational simplicity.
4. Emanuel Parzen, *Modern Probability Theory and its Applications*, John Wiley & Sons (1960), New York, Chapter 2.
5. Lajos Takács, *Introduction to the Theory of Queues*, Oxford University Press (1962), New York, Chapter 1.