process phosphoric acid solutions. The uranium extraction coefficients and the organic phase iron loading were not affected by the MOPPA/DOPPA ratio within the range of 0.7 to 1.5 or by differences in the concentrations of impurities in the two OPAP batches tested. Extraction of uranium from simulated wet-process phosphoric acid solutions was slightly lower than OPAP solutions prepared from commercial OPAP than with solutions prepared from purified MOPPA and DOPPA fractions. This effect was due to the presence in the commercial OPAP of impurities that depressed uranium extraction. Octyl phenol was possibly involved but other impurities were more important, since the effect of octyl phenol on uranium extraction was shown to be small in earlier studies (Arnold et al., 1980). The increased uranium extraction power with purified fractions was not large enough to justify separation and purification of the fractions for process use. In addition, the combination of purified fractions also extracted iron more strongly and was more susceptible to precipitation as a complex with ferric iron.

Uranium extraction was more strongly affected by changes in aqueous phase composition. The extraction coefficient was strongly depressed by increasing the concentrations of $\sum PO_4$, Fe^{III}, or HF, and moderately depressed by increasing the concentration of $\sum SO_4$. The uranium extraction coefficient increased when the concentrations of Si, Al, or Mg were increased. The increases were probably due to complexing of part of the fluoride by Si or Al, and to reduction of the aqueous phase acidity by the addition of Al or Mg. The extraction of Fe^{III} with OPAP was reduced slightly by increasing the $\sum PO_4$ concentration, and was not affected appreciably by changing the concentration of other aqueous phase components.

Based on these results, the most important impurities in wet-process phosphoric acid with regard to uranium extraction with OPAP are Fe^{III} and $\sum F$. Iron(III) is particularly important because it is extracted by OPAP and competes with uranium for the extractant. Uranium extraction could be improved by minimizing the concentrations of these impurities in the acid. Both of these impurities would have a smaller effect on uranium extraction from hemihydrate process acids (high $\sum PO_4$) than from dihydrate process acids (low $\sum PO_4$).

Literature Cited

Arnold, W. D.; McKamey, D. R.; Baes, C. F. "Progress Report on Uranium Recovery from Wet-Process Phosphoric Acid with Octylphenyl Acid Phosphate"; ORNL/TM-7182 (Jan 1980).
Goers, W. E. Proceedings of 28th Annual Meeting, Fertilizer Industry Roundtable, Atlanta, GA, 1978, p 99.
Hurst, F. J.; Crouse, D. J. *Ind. Eng. Chem. Process Des. Dev.* **1974**, *13*, 286.
McGinley, F. E. Paper presented at Uranium Industry Seminar, Grand Junction, CO, 1972.
Orê, F. Proceedings of 28th Annual Meeting, Fertilizer Industry Roundtable, Atlanta, GA, 1978, p 58.
Schwer, E. W.; Crozier, B. T. *Proceedings of 28th Annual Meeting, Fertilizer Industry Roundtable*, Atlanta, GA, 1978, p 69.

Received for review March 19, 1981
Accepted December 28, 1981

This research was sponsored by the Division of Chemical Sciences, Office of Basic Energy Sciences, U.S. Department of Energy, under Contract W-7405-eng-26 with the Union Carbibe Corporation.

Internal Model Control. 1. A Unifying Review and Some New Results

Carlos E. Garcia and Manfred Morari*

Chemical Engineering Department, University of Wisconsin, Madison, Wisconsin 53706

Internal Model Control (IMC) is defined for single input–single output, discrete-time systems and its relationships with other control schemes (Optimal Control, Smith Predictor, Inferential Control, Model Algorithmic Control, Dynamic Matrix Control) are established. Several new stability theorems for IMC are proven which provide practical tuning guidelines. It is concluded that the IMC structure allows a rational controller design procedure where control quality and robustness can be influenced in a direct manner. Finally, the transparency and intuitive appeal of IMC make it attractive for industrial applications.

1. Introduction

Spawned at least partly by the introduction of the state space approach, we witnessed a surge in process control research during the 60's. This period of activity was followed by a period of reflexion during the 70's (Foss, 1973; Lee and Weekman, 1976; Kestenbaum et al., 1976) which was precipitated by the realization that the wealth of new theory had only a minimal impact on the process control practice.

One repeatedly addressed issue was the performance of optimal controllers for single input–single output (SISO) systems in comparison with that of standard controllers of the PID type (Kestenbaum et al., 1976; Bohl and McAvoy, 1976; Palmor and Shinnar, 1979). Though these are quite straightforward applications of optimal control theory which unfolds its power more impressively in the multivariable case, it is nevertheless an issue of significant practical interest.

Before proceeding with any more detail we would like to emphasize that in this paper we are not concerned with the 90+% of the control problems encountered in industry where essentially any control scheme will work satisfactorily but with the other situations where high performance loops are of critical importance for safety or economic reasons but are difficult to design. In these cases we can assume that increased manpower for the design and in-

0196-4305/82/1121-0308$01.25/0

creased hardware expenditures for the implementation are economically justifiable.

For the purpose of comparison we postulate that with regard to chemical processes the quality of a controller should be judged according to the following criteria: (1) Regulatory behavior. The output variable is to be kept at its setpoint despite unmeasured disturbances affecting the process. (2) Servo behavior. Changes in the setpoint should be tracked fast and smoothly. (3) Robustness. Stability and acceptable control performance should be maintained in the face of structural and parametric changes in the underlying process model. (Note that this is entirely equivalent to requiring that a controller should be designable with a minimum of information about the process.) (4) Ability to deal with constraints on the inputs and states.

It is a well-accepted fact (e.g., Lee and Weekman, 1976) that the major economic return from process control arises from the optimization of the operating conditions, i.e., determination of the optimal setpoints, rather than from regulation. A well-functioning regulatory layer is necessary, however, to implement the actions dictated by the optimizing layer. Optimal operating points generally lie at the intersection of constraints (Arkun and Stephanopoulos, 1980). Therefore the ability of the regulatory controller to deal with constraints on both the inputs and the states is very important.

All other criteria mentioned in the literature (Kestenbaum et al., 1976; Palmor and Shinnar, 1979) are either implied by the four above or considered to be of minor importance. In particular, *closed-loop stability* of the system in the *absence* of plant variations (compare "robustness" above) is almost never an issue because the majority of chemical processes is open-loop stable. Sometimes a trade-off situation is imagined between stability and control quality: For example, for a system of order higher than 2 or a system with time delay the gain of a P-controller and thus the achievable control quality is limited by stability considerations. This, however, is only an artifact of the assumed simple control structure. With the proper dynamic compensator any gain is possible for any system.

Controller complexity is an important issue in the sense that the control structure and the effects of the tuning parameters should be transparent to the operator. This is not to be misunderstood as a shortsighted endorsement of controllers of the PID type. We have to realize that the classic three-mode controller is a consequence of hardware limitations. Any kind of dynamic compensation can easily be added when computer control is used. Extremely speaking we could sprinkle poles and zeroes like salt and pepper over the complex plane if this should turn out to be desirable and as long as it does not impair on the overall transparency.

The *response* of the controller *to specific disturbances* (e.g., use of stochastic noise model for controller design, avoiding resonance frequency in certain range, etc.) is of secondary importance. A filter inserted in the completed control system can correct any undesirable features to the maximum possible extent.

The results of the extensive comparisons between optimal control and conventional control can be summarized as follows. Optimal control yields improved servo- and regulator behavior but the crucial robustness issue cannot be addressed directly. Weighting matrices and/or noise models have to be varied in a roundabout obscure fashion in the hope of achieving some robustness which then has to be checked through simulation. Actually these conclusions are not surprising. Though the quadratic performance index is often criticized for just being a mathematical convenience, it yields a controller where the gain can be increased arbitrarily and thus a "perfect" response can be approached asymptotically. On the other hand, it has been stated repeatedly (Horowitz and Shaked, 1975) that the state space description which forms the framework for most of today's optimal control theory is unsuitable for robustness investigations. Sensitivity theory based on the assumption of infinitesimal perturbations is no viable substitute here.

We have to add that *all* methods, modern as well as conventional, are equally incapable of handling constraints. We disregard here indirect tools like penalty functions added to the objective or ad hoc solutions like switching the controller to "manual".

While the academic discussions about the relative merits of the different schemes were in progress, apparently new techniques were developed in industry independently in France (Richalet et al., 1978) and in the U.S. by Shell (Cutler and Ramaker, 1979). They are not the result of a new theory but have a heuristic basis. As an attractive feature the four criteria stated above are addressed quite directly in the design. Moreover, simulations and industrial applications showed the performance of these heuristic methods to be vastly superior in particular with respect to robustness.

Because of the empirical origin of these techniques, reservations about its general qualities and limitations are only appropriate. In this paper we put them in the proper place in the overall control theory framework. A rigorous formulation allowed us to discover similarities and differences between them and many other control schemes, among them the optimal controller and the Smith predictor. This comparison provided new insight into the new methods as well as the old ones. This insight yielded not only a set of new theoretical results which significantly expand the scope of the techniques but also pointed out a large number of new promising research areas.

2. Model Formulation

The strategies considered in this paper provide discrete time control for the single input, single output continuous process

$$\dot{\boldsymbol{x}} = \mathbf{A}\boldsymbol{x} + \boldsymbol{B}m(t - \theta) + \boldsymbol{w} \qquad (1)$$
$$y = \boldsymbol{C}\boldsymbol{x}$$

where $\boldsymbol{x} \in R^n$ is the state, θ is a delay on the input, and $\boldsymbol{w}$ takes into account all unmeasurable disturbances affecting the process. For a given sampling time T the controller prescribes constant input values $m(k)$ to be applied during the intervals $kT < t \leq (k + 1)T$, $k = 0, 1, \ldots$. These inputs are based on output measurements at times kT, $k = 0, 1, \ldots$. Making the simplifying but not essential assumption that θ is an integer multiple of T ($\theta = \tau T$), the description of system 1 at the discrete intervals is

$$\boldsymbol{x}(k + 1) = \boldsymbol{\Phi}\boldsymbol{x}(k) + \Gamma m(k - \tau) + \boldsymbol{w}(k) \qquad (2)$$
$$y(k) = \boldsymbol{C}\boldsymbol{x}(k)$$

where

$$\boldsymbol{\Phi} = e^{\mathbf{A}T}; \qquad \boldsymbol{\Gamma} = \left(\int_0^T e^{\mathbf{A}q}\, \mathrm{d}q\right)\mathbf{B}$$

and $\boldsymbol{w}(k)$ is the discretized disturbance vector $\boldsymbol{w}(t)$.

Analogous to the Laplace domain representation for continuous systems, we find it convenient to express system 2 in the z domain as follows (Jury, 1964)

$$z\mathbf{x}(z) = \Phi\mathbf{x}(z) + z^{-\tau}\Gamma m(z) + \mathbf{w}(z) \tag{3}$$
$$y(z) = \boldsymbol{C}\mathbf{x}(z)$$

where $Z\{\mathbf{x}(k + a)\} = z^a\mathbf{x}(z)$ or in the input/output form as a difference equation

$$y(z) = \boldsymbol{C}(z\boldsymbol{I} - \Phi)^{-1}\Gamma z^{-\tau}m(z) + d(z) \tag{4}$$

where $d(z) = \boldsymbol{C}(z\mathbf{I} - \Phi)^{-1}\mathbf{w}(z)$. For a stable system (i.e., $\|\Phi\| < 1$), one can further expand the transfer function as an infinite series in z

$$\boldsymbol{C}(z\mathbf{I} - \Phi)^{-1}\Gamma = z^{-1}\sum_{l=1}^{\infty}(\boldsymbol{C}\Phi^{l-1}\Gamma)z^{-(l-1)} = z^{-1}\sum_{l=1}^{\infty}h_l z^{-(l-1)} = z^{-1}H(z) \tag{5}$$

finally obtaining the "impulse response model"

$$y(z) = z^{-\tau}z^{-1}H(z)m(z) + d(z) \tag{6}$$

Note that the coefficients h_l in the convolution sum represent the discrete unit impulse response of the system and satisfy

$$\lim_{l\to\infty} h_l = 0 \tag{7}$$

Consequently, for any $\epsilon > 0$, there exists an N such that

$$|\mathbf{C}\Phi^{l-1}\Gamma| < \epsilon \qquad \text{for } l > N$$

such that (6) can be approximated to any desired degree of accuracy by

$$y(z) = z^{-\tau}z^{-1}\sum_{l=1}^{N}h_l z^{-(l-1)}m(z) + d(z) \tag{8}$$

In addition, one can obtain the equivalent "step response model" from (8) by making the substitution

$$h_l = a_l - a_{l-1} \tag{9}$$

where

$$a_l = \sum_{i=1}^{l}h_i$$

are the discrete unit step response coefficients. Therefore (8) becomes

$$y(z) = z^{-\tau}\{z^{-1}\sum_{l=1}^{N}a_l z^{-(l-1)}\}(1 - z^{-1})m(z) + d(z) \tag{10}$$

where now the model relates $y(k)$ to changes in the input

$$\Delta m(k) = m(k) - m(k - 1)$$

Most of the results derived in this paper are independent of the particular process representation. Very often in process control a "natural" fundamental state space process model is not available or the order is too high for it to be useful for control system design. Then the impulse (and equivalently the step) response model offers several advantages: no assumptions have to be made about the number of poles and zeroes present in the system; therefore we call the model structure free. As an apparent drawback many more parameters have to be determined than in a structured model, but because of this nonminimal description the robustness is improved. The use of a state space model would require in addition the selection of a coordinate system. The closed-loop robustness properties are influenced by the selected coordinate system in a nontrivial way and therefore this choice is very difficult. Finally, we repeat that the frequency domain is the appropriate framework for a robustness analysis. Because the transfer function is simply the Z-transform of the impulse response, the impulse response model is very convenient for a theoretical investigation of robustness.

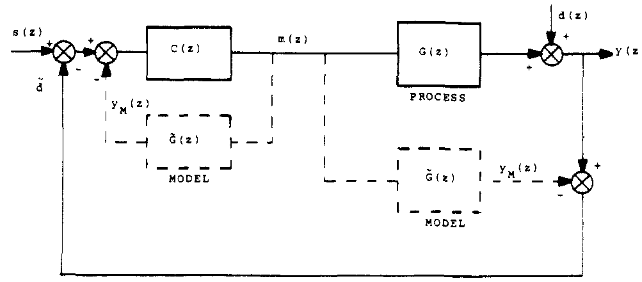

Figure 1. The discrete-time feedback controller with the model modifications (dashed lines).

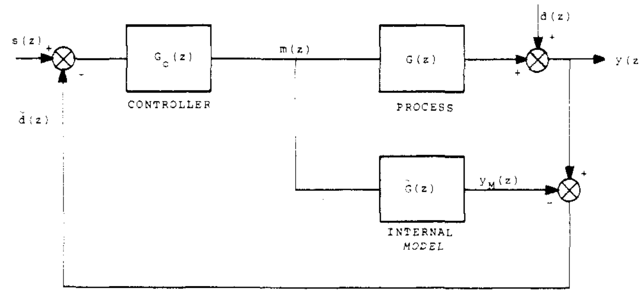

Figure 2. Basic IMC structure.

3. Structural Analogies

Motivated by Brosilow's (1979) discussion, we develop the structure of Internal Model Control (IMC) and establish analogies with conventional control schemes.

3.1. The IMC Structure. 3.1.1. Evolution of the IMC Structure. Consider the familiar feedback control arrangement shown in Figure 1, where the process is represented by a discrete transfer function $G(z)$ with setpoint $s(z)$ and disturbance $d(z)$. Denoting the model transfer function by $\tilde{G}(z)$ one can subtract and add the effect of the input m on the measurement signal ($y_M(z)$) (dashed lines in Figure 1) yielding an entirely equivalent setup. If we consider the control block $C(z)$ with model feedback as our new controller we obtain the basic IMC structure shown in Figure 2. Hence any conventional feedback controller can be restructured to yield IMC. Furthermore, any IMC can be put into the conventional feedback form by defining

$$C(z) = \frac{G_c(z)}{1 - G_c(z)\tilde{G}(z)}$$

If the two structures are interchangeable, what is the advantage of using IMC? We claim that the controller $G_c(z)$ is much easier to design than $C(z)$ and that the IMC structure allows us to include robustness as a design objective in a very explicit manner. This is partly due to the special form of the feedback signal $\tilde{d}(z)$

$$\tilde{d}(z) = (1 + (G(z) - \tilde{G}(z))G_c(z))^{-1}d(z) \tag{11}$$

We note that for a perfect model the feedback signal is just the disturbance $d(z)$. Thus the system is effectively open-loop and stability is not an issue. When there are differences between the plant and the model, $\tilde{d}(z)$ contains some information on them and by modifying $\tilde{d}(z)$ appropriately we can obtain robustness. We will elaborate on these points in more detail in the following.

3.1.2. General Properties of IMC and the Perfect Controller. After some manipulations the following transfer functions are obtained from the block diagram in Figure 2.

$$m(z) = \frac{G_c(z)}{1 + G_c(z)(G(z) - \tilde{G}(z))}(s(z) - d(z)) \quad (12)$$

$$y(z) = d(z) + \frac{G(z)G_c(z)}{1 + G_c(z)(G(z) - \tilde{G}(z))}(s(z) - d(z)) \quad (13)$$

For stability it is necessary and sufficient that *both* of the following characteristic equations have their roots strictly inside the unit circle.

$$\frac{1}{G_c(z)} + (G(z) - \tilde{G}(z)) = 0 \quad (14)$$

$$\frac{1}{G(z)G_c(z)} + \frac{1}{G(z)}(G(z) - \tilde{G}(z)) = 0 \quad (15)$$

Property 1: Dual Stability Criterion. When the model is exact, stability of both controller and plant is sufficient for overall system stability.

Proof. When $G(z) = \tilde{G}(z)$ the characteristic equations become

$$\frac{1}{G_c(z)} = 0 \quad (16)$$

$$\frac{1}{G(z)G_c(z)} = 0 \quad (17)$$

This implies that both the system and the controller poles have to be inside the unit circle. Q.E.D.

Property 1 confirms our statement above that for an open-loop stable plant closed-loop stability (but not necessarily robustness) is obtained automatically for *any* stable control block $G_c(z)$. This is also trivially true even when the control block is nonlinear as long as it is input-output stable. Thus any constraints on the control action, for example, would not affect the stability. For conventional feedback structures with nonlinear controllers the stability issue is much more complex (Popov, 1963). We should also note, however, that in general IMC does not allow the stabilization of an open-loop unstable plant. This could only be achieved when the controller cancels exactly the unstable poles of the system which is impossible in practice. An unstable plant has to be stabilized first by conventional feedback before IMC is used. Throughout the remaining part of this paper we assume the plant to be open-loop stable.

After the resolution of the stability issue the question of the best choice for $G_c(z)$ arises. Equation 13 shows that "perfect" control could be achieved by selecting $G_c(z) = 1/\tilde{G}(z)$. Because $G_c(z)$ has to be stable and realizable we introduce the following factorization

$$\tilde{G}(z) = \tilde{G}_+(z)\tilde{G}_-(z) \quad (18)$$

where $\tilde{G}_+(z)$ contains all the time delays (their inversion would require a prediction) and all the zeroes outside the unit circle (their inversion would yield an unstable controller). These system characteristics are referred to as "non-minimum phase" (NMP) in classical frequency response theory.

The factorization (18) is non-unique, but the following form of $\tilde{G}_+(z)$ is advantageous (see "Property 2")

$$\tilde{G}_+(z) = z^{-(\tau+1)} \prod_{i=1}^{p} \left(\frac{z - \nu_i}{z - \hat{\nu}_i}\right)\left(\frac{1 - \hat{\nu}_i}{1 - \nu_i}\right)$$

where ν_i are the p zeroes of $\tilde{G}(z)$ and $\hat{\nu}_i$ are their images inside the unit circle

$$\hat{\nu}_i = \nu_i \quad |\nu_i| \leq 1$$
$$= 1/\nu_i \quad |\nu_i| > 1$$

In terms of the sum of the squared errors the best performance achievable by any control system can be obtained through the "perfect controller".

Property 2: Perfect Controller. Under the assumption that $G(z) = \tilde{G}(z)$ and that $G(z)$ is stable, the sum of the squared errors is minimized for both the regulator and the servo-controller when

$$G_c(z) = 1/G_-(z) \quad (19)$$

This property can be derived from optimal control theory. Rather than doing that we will discuss the consequences of (19). For the perfect controller (13) becomes

$$y(z) = G_+(z)s(z) + (1 - G_+(z))d(z) \quad (20)$$

We note that $G_+(z)$ contains the part of the process which limits the achievable control quality. These limitations are inherent in the physical system and cannot be removed by *any* control system. In the absence of time delays and nonminimum phase behavior truly perfect control would be possible. It is well known that the system is very sensitive to modelling errors when the perfect controller is used.

Property 3: Zero Offset. A controller G_c which satisfies $G_c(1) = \tilde{G}(1)^{-1}$ yields zero offset.

Proof. From (13) we have for a unit step change in $s(z)$

$$\lim_{t\to\infty} y(t) = \frac{G(1)\tilde{G}(1)^{-1}}{1 + \tilde{G}(1)^{-1}(G(1) - \tilde{G}(1))}(1 - d(1)) + d(1) = 1$$

for any asymptotically constant disturbance d. Q.E.D.

3.1.3. Controller Design Procedure. The properties derived above indicate that IMC allows a two step controller design. Step 1: select $G_c(z) = 1/\tilde{G}_-(z)$ or a suitable approximation thereof. Step 2: add a filter to make the system robust. We will discuss Step 2 in the next section.

As a result of Property 2 it is desirable to perform the factorization (18) and use controller (19). However, if an impulse response model (6) is used this may not be possible. Hence, our contention is that we may conveniently approximate the process inverse and still obtain a controller with desirable properties. This may be done in several ways. One may factor out the delays and then approximate the remaining part of $G(z)$. Alternatively, the complete model inverse may be approximated. For example, proportional control ($C(z) = K$) in Figure 1 yields the IMC control block

$$G_c(z) = \frac{K}{1 + K\tilde{G}(z)}$$

which can be interpreted as an approximation to $\tilde{G}(z)^{-1}$, not necessarily a very good one.

Assuming exact model description (6) and that $H(z)$ has all roots inside the unit circle, the perfect controller is

$$m(z) = H(z)^{-1}(s(z) - d(z)) \quad (21)$$

yielding an overall response

$$y(z) = z^{-\tau-1}s(z) + (1 - z^{-\tau-1})d(z)$$

Note that besides the dead times there is an additional sampling interval inherent in all discrete-time systems which makes the complete inversion impossible even when $\tau = 0$. The resulting delay in the setpoint can be compensated by specifying the required value $\tau + 1$ sampling times ahead, thus yielding

$$y(z) = s(z) + (1 - z^{-\tau-1})d(z) \quad (22)$$

which means perfect compensation after $\tau + 1$ invervals.

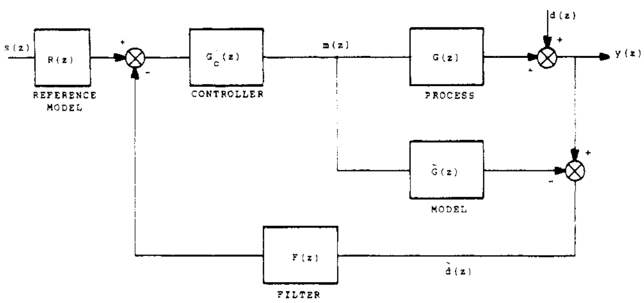

Figure 3. Complete IMC structure.

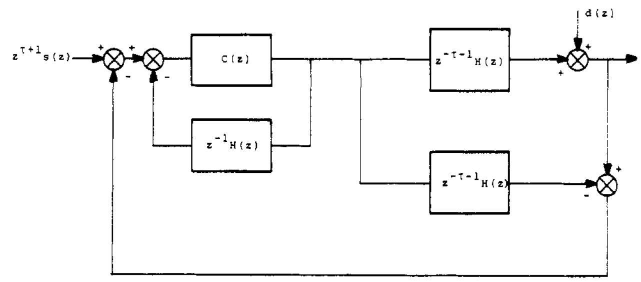

Figure 4. The discrete-time Smith predictor.

In case $H(z)$ has roots outside the unit circle, $G_c(z) = H(z)^{-1}$ also gives the closed-loop response (22); however, setpoint satisfaction is achieved at the expense of an unstable input sequence, as generated by

$$m(z) = \frac{1}{H(z)} (z^{\tau+1}s(z) - d(z)) \tag{23}$$

This result stresses the importance of the dual stability criterion (Property 1). Hence, since the unstable roots cannot be factored conveniently, $H(z)^{-1}$ has to be approximated. Section 4 is devoted solely to the design of stable approximations of $H(z)^{-1}$.

3.1.4. Sensitivity/Reference Trajectory. Although during the rest of the paper exact model description will indeed be assumed, in this section we discuss the possibility of desensitizing IMC with respect to modelling inaccuracies. As proposed by Brosilow (1979), one can include a filter block $F(z)$ in the feedback signal to account for the model and process mismatch reflected through the estimated disturbance $\tilde{d}(z)$ (see Figure 3). (In the case of an exact model such a filter could also compensate for certain types of disturbance dynamics.) As a result, the characteristic equation (14) becomes

$$G_c(z)^{-1} + F(z)(G(z) - \tilde{G}(z)) = 0 \tag{24}$$

For a given model-plant mismatch $F(z)$ can be selected such that (24) has all roots inside the unit circle. A detailed design procedure for $F(z)$ will not be attempted at this point. Nevertheless, we will demonstrate the main idea by means of the following example.

Let us assume that the delay-free part of the model ($H(z)$) has all roots inside the unit circle and is known precisely, while the process has one unit of delay not accounted for by the model, that is

$$G(z) = z^{-2}H(z)$$

but

$$\tilde{G}(z) = z^{-1}H(z)$$

Choosing the stable controller $G_c(z) = H(z)^{-1}$, (24) becomes

$$H(z)(1 + F(z)(z^{-2} - z^{-1})) = 0 \tag{25}$$

For $F(z) = 1$ this characteristic equation has two roots on the unit circle. However, by introducing the stable exponential filter

$$F(z) = \frac{1-\alpha}{1-\alpha z^{-1}} \qquad 0 < \alpha < 1 \tag{26}$$

those roots become

$$z_1,z_2 = \frac{1}{2} \bullet \frac{1}{2}\sqrt{4\alpha - 3}$$

which are stable over the range of α given.

In addition, one can also make the scheme follow a reference trajectory by feeding the setpoint to a reference model block $R(z)$. The setpoint following characteristics of perfect control will make the output track the desired dynamics. The complete IMC setup including all filter blocks is shown in Figure 3.

In summary, the structure of IMC allows a rational design procedure where in the first step the controller is selected to give perfect control. In the second step we back away from the ideal of perfect control by introducing a filter which makes the system robust to a specified model-plant mismatch. This distinguishes IMC from all other feedback control schemes where a compromise between good response, stability, and robustness is accomplished in a very indirect fashion. Let us explore further how other control schemes are related to IMC.

3.2. The (Inferential) Smith-Predictor (SP). A discrete analog of the celebrated time delay compensator devised by Smith (1957) is given in Figure 4 (Alevisakis and Seborg, 1973). The main idea is to eliminate the time delay from the characteristic equation so that higher controller gains can be achieved. One may note that in the light of the discussion in the previous section the SP "factors out" the pure delays τ, then the rest of the model inverse is approximated, including the intrinsic sampling delay.

Brosilow (1979) has interpreted the SP from an inferential point of view: the effect of disturbances on the outputs is estimated and handled directly by the controller. This reasoning has been adapted and exploited in the present paper.

3.3. The Linear Quadratic Optimal Controller (LQOC). 3.3.1. Delay-Free Plant. Most modern control theory is centered around the combined problem of linear quadratic optimal control and state reconstruction. Invoking the separation theorem (Åström, 1970) the following minimization problem is solved

$$\min_{m(k)} E\{\sum_{k=0}^{\infty} \mathbf{x}(k+1)^T\mathbf{Q}_1\mathbf{x}(k+1) + Q_2 m(k)^2\} \tag{27}$$

subject to

$$\mathbf{x}(k+1) = \tilde{\mathbf{\Phi}}\mathbf{x}(k) + \tilde{\mathbf{\Gamma}}m(k) + \mathbf{w}_1(k)$$
$$y(k) = \tilde{\mathbf{C}}\mathbf{x}(k) + w_2(k)$$

where $\mathbf{Q}_1$ is a symmetric positive semidefinite matrix, $Q_2 \geq 0$, E is the expected value operator, and $\mathbf{w}_1(k)$ and $w_2(k)$ are mutually independent Gaussian processes with zero means. The tildes above system equation matrices $\mathbf{\Phi}$, $\mathbf{\Gamma}$, and $\mathbf{C}$ have been included to emphasize optimization with respect to a model of the plant. The control law is known to have the form

$$m(k) = -\mathbf{K}\hat{\mathbf{x}}(k) + G_{CL}^{-1}(1)s(k) \tag{28}$$
$$\hat{\mathbf{x}}(k+1) = \tilde{\mathbf{\Phi}}\hat{\mathbf{x}}(k) + \tilde{\mathbf{\Gamma}}m(k) + \bar{\mathbf{K}}[y(k) - \tilde{\mathbf{C}}\hat{\mathbf{x}}(k)]$$

where $y(k)$ is the noise corrupted measured output, $\mathbf{K}$ is the proportional gain vector, $\bar{\mathbf{K}}$ is the filter gain vector, $\hat{\mathbf{x}}(k)$ is the estimate of the state $\mathbf{x}(k)$, and $G_{CL}(1)$ is the

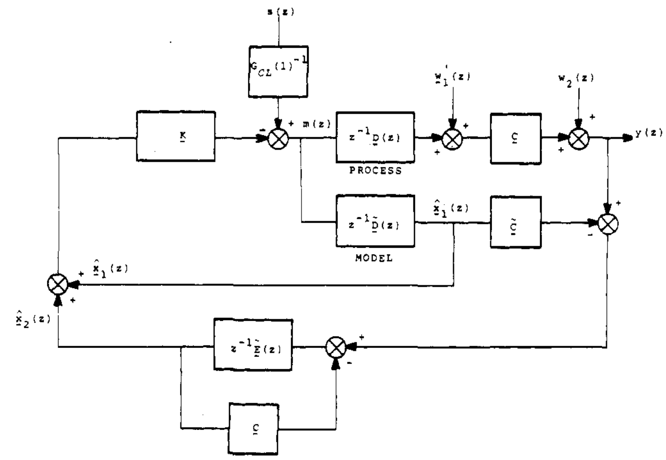

Figure 5. The LQOC structure for a delay-free SISO system.

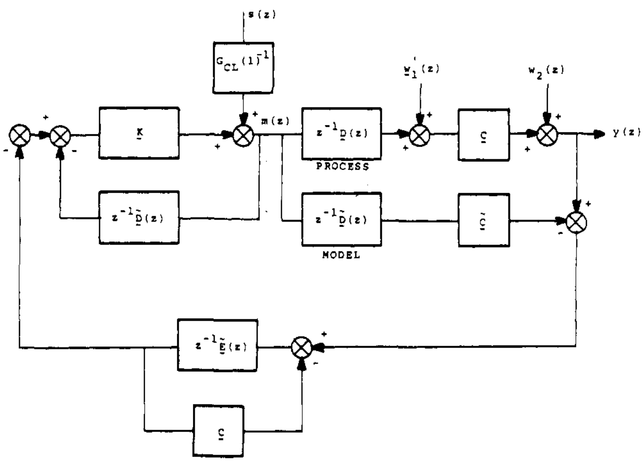

Figure 6. Almost exact resemblance of the LQOC structure with IMC.

steady-state closed-loop gain included to achieve zero offset for setpoint changes (Kwakernaak and Sivan; 1972)

$$G_{\mathrm{CL}}(z) = \tilde{\boldsymbol{C}}(z\mathbf{I} - \tilde{\boldsymbol{\Phi}} + \tilde{\boldsymbol{\Gamma}}\boldsymbol{K})^{-1}\tilde{\boldsymbol{\Gamma}}$$

In transfer function form, eq 28 become

$$m(z) = -\boldsymbol{K}\hat{\boldsymbol{x}}(z) + G_{\mathrm{CL}}^{-1}(1)s(z) \qquad (29)$$

$$\hat{\boldsymbol{x}}(z) = z^{-1}\tilde{\boldsymbol{D}}(z)m(z) + z^{-1}\tilde{\boldsymbol{E}}(z)[y(z) - \tilde{\boldsymbol{C}}\hat{\boldsymbol{x}}(z)]$$

where

$$\tilde{\boldsymbol{D}}(z) = (\mathbf{I} - z^{-1}\tilde{\boldsymbol{\Phi}})^{-1}\tilde{\boldsymbol{\Gamma}}$$

$$\tilde{\boldsymbol{E}}(z) = (\mathbf{I} - z^{-1}\tilde{\boldsymbol{\Phi}})^{-1}\bar{\boldsymbol{K}}$$

Let us now break up the state $\hat{\boldsymbol{x}}(z)$

$$\hat{\boldsymbol{x}}(z) = \hat{\boldsymbol{x}}_1(z) + \hat{\boldsymbol{x}}_2(z)$$

such that

$$\hat{\boldsymbol{x}}_1(z) = z^{-1}\tilde{\boldsymbol{D}}(z)m(z)$$

$$\hat{\boldsymbol{x}}_2(z) = z^{-1}\tilde{\boldsymbol{E}}(z)[y(z) - \tilde{\boldsymbol{C}}\hat{\boldsymbol{x}}(z)]$$

With these definitions the block diagram representation of the optimal controller is shown in Figure 5.

After minor rearrangements one discovers an almost exact resemblance with the IMC structure in Figure 6. It features the same approximate inverse controller as the SP but with the setpoint introduced somewhat differently. However, for a first-order system the setpoint can be shifted to the same position as in IMC, making the analogy complete.

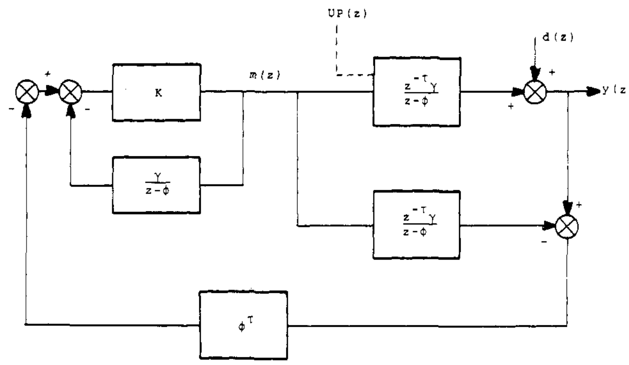

Figure 7. LQOC for a deterministic first order system with delay.

One remarkable result is the natural occurrence of the filter $\boldsymbol{F}(z)$ introduced above somewhat heuristically

$$\boldsymbol{F}(z) = (\mathbf{I} + z^{-1}\tilde{\boldsymbol{E}}(z)\tilde{\boldsymbol{C}})^{-1}z^{-1}\tilde{\boldsymbol{E}}(z)$$

(which is now a vector due to the state reconstruction). Its function is to account for uncertainties in the measurement (w_2) and model. Not surprisingly, the filter was found to play a similar role in the IMC scheme and thus this analogy provides a fundamental justification for its inclusion. The crucial difference is that the filter here is designed by assuming certain intensities for the white noise processes $\boldsymbol{w}_1$ and w_2. Robustness is related in a very complex manner to these choices as well as the selected values for $\mathbf{Q}_1$ and Q_2 and the chosen state space coordinate system (Doyle, 1978; Safanov and Athans, 1978). The IMC structure, on the contrary, allows a filter design to achieve robustness in a direct manner for a specified model-plant mismatch.

3.3.2. Delayed Process. The structural similarity among the SP and the optimal control law has been pointed out in the work of Ogunnaike and Ray (1979). Basically, the scheme "compensates" for τ delays and approximates the rest of the inverse by feedback as was shown above. However, as we presently demonstrate using a simple first-order deterministic system as an example, the optimal control law includes an additional factor which makes it perform sub-optimally as a regulator.

According to standard optimal control theory (Fuller, 1968; Koppel, 1968) the control law for a system with delays in the input is

$$m(k) = -Ky(k + \tau)$$

where $y(k + 1) = \phi y(k) + \gamma m(k - \tau)$, and K is the same proportional gain which solves the delay-free problem. By successively substituting for $y(k)$, one can express $y(k + \tau)$ as follows

$$y(k + \tau) = \phi^{\tau}y(k) + \phi^{\tau-1}\gamma m(k - \tau) + \ldots + \phi\gamma m(k - 2) + \gamma m(k - 1)$$

or in Z-transforms as

$$z^{\tau}y(z) = \phi^{\tau}y(z) + X(z)$$

where

$$X(z) = (\phi^{\tau-1}z^{-\tau} + \phi^{\tau-2}z^{-\tau+1} + \ldots + \phi z^{-2} + z^{-1})\gamma m(z)$$

But using the identity

$$X(z) - z^{-1}\phi X(z) = z^{-1}\gamma m(z) - \phi^{\tau}z^{-\tau-1}\gamma m(z)$$

the control law becomes

$$m(z) = -K\{\phi^{\tau}(y(z) - z^{-\tau-1}H(z)m(z)) + z^{-1}H(z)m(z)\} \qquad (30)$$

Figure 7 shows the resulting law in block diagram form. As expected, we have obtained the SP for a first-order system except for the additional constant filter block ϕ^τ in the feedback path. The optimal controller is designed with the objective to minimize a quadratic performance index when the system returns from a nonzero state to the origin. It is known that for systems without time delay the resulting optimal feedback control law yields also good regulator properties, for example asymptotically vanishing offset for a decreasing penalty on the control effort.

Let us investigate both the servo- and the regulator properties of the optimal controller for time delay systems. For this purpose a unit pulse (UP(z)) will be introduced into the process (this is equivalent to a nonzero state at time τ) and also a constant disturbance d. The response of the closed loop system is

$$y(z) = \frac{z^{-\tau}\gamma}{z-\phi}\left\{1 - \frac{z^{-\tau}\phi^\tau\gamma K}{z-\phi+\gamma K}\right\}\mathrm{UP}(z) + \left\{1 - \frac{z^{-\tau}\phi^\tau\gamma K}{z-\phi+\gamma K}\right\}d \qquad (31)$$

For the minimum variance control problem ($Q_2 = 0$ in (27)) the optimal proportional gain can be shown to be $K = \phi/\gamma$, which puts the root of the characteristic equation precisely at the origin. The response to the unit pulse alone is from (31)

$$y(z) = z^{-(\tau+1)}\gamma(1 + \phi z^{-1} + \ldots \phi^\tau z^{-\tau})\mathrm{UP}(z)$$

which shows dead-beat convergence to zero after $\tau + 1$ sampling times. No other controller, including the SP of Figure 4, can do better in driving the state to the origin.

If we look at the asymptotic behavior of the response (31)

$$\lim_{t\to\infty} y(t) = (1 - \phi^{\tau+1})d$$

we notice, however, that even without any limits on the controller power we will *always* have offset, which increases with the time delay. Note that this offset occurs even when $\tau = 0$ and also in the continuous-time counterpart of the minimum variance problem (Fuller, 1968).

These results may be summarized as follows. For systems with time delay the optimal controller cannot be regarded as an optimal regulator because it is unable to remove offset. Therefore it is inappropriate to use it in a comparison with standard regulators and to draw conclusions about optimal control on this basis (cf. Kestenbaum et al., 1976). The perfect controller (19) has not only equally good servo-behavior as the optimal controller, but also zero offset after $\tau + 1$ intervals. This provides further justification for using the IMC philosophy.

4. Approximate Inverses

According to the IMC strategy one design problem is to find a stable approximation to the inverse of $H(z)$, the other to design the filter F to preserve stability for the specified plant-model mismatch. The latter problem is not dealt with in any depth in this paper, but a discussion is included in the conclusion section. The need for an approximation of $H(z)^{-1}$ can arise for two reasons. (1) $H(z)$ has roots outside the unit circle and therefore the exact inverse would be unstable. (2) The use of an exact inverse, even when stable, often requires excessive control action and can lead to an undesirable oscillatory response of the continuous process between the sampling times.

Although this problem can be partly alleviated by using reference model tracking (cf. section 3.1.4), by defining suitable approximations to the inverse the designer has

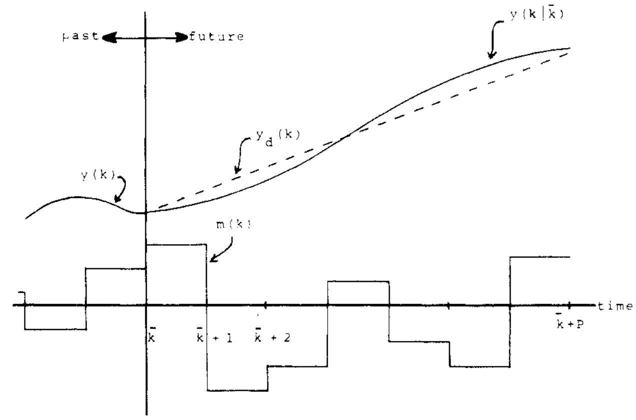

Figure 8. General predictive control scheme.

more freedom to impose desirable dynamic characteristics on the closed-loop system.

In order to derive approximate inverses a general predictive-type control problem is formulated and solved in this section. Tuning parameters of physical significance to the designer are introduced. As a key result, we prove several stability theorems which allow us to make reasonable decisions on those parameters and hence to obtain a control law which is extremely simple to implement and has the desirable characteristics. Finally, we demonstrate that some recently developed algorithms which have been shown to perform well in industrial applications are just special cases of IMC.

4.1. Predictive Control Problem. 4.1.1. Problem Formulation. In many control applications one is interested in predicting the output trajectory of a process on the basis of a model as, for example, in order to prevent the violation of operating constraints. A suitable predictive control strategy would consider the desired trajectory $y_d(k)$ over a horizon of P sampling times into the future. Then the sequence of control actions $m(\bar{k})$, $m(\bar{k}+1) \ldots m(\bar{k} + P - 1)$ is calculated so that the predicted output $y(\bar{k} + l|\bar{k})$, $l = 1, \ldots, P$ follows $y_d(k)$ as closely as possible. This prediction is defined as the model prediction using the inputs up to $\bar{k} + l - \tau - 1$ plus the available information on the outputs up to $\bar{k}$. If the plant model, which is used for the prediction and the control computation, is incorrect the plant response in this open loop control law could deviate significantly from the desired trajectory. Thus it is preferable to implement only the present input $m(\bar{k})$ and to resolve the problem again at $\bar{k} + 1$ with the measured output $y(\bar{k} + 1)$ as the new starting point and so on for $\bar{k} + 2$, $\bar{k} + 3$ etc. Figure 8 illustrates one step in this moving horizon problem.

At each time $\bar{k}$ the following general problem is solved

$$\min_{\substack{m(\bar{k}),\\ m(\bar{k}+1),\ldots,\\ m(\bar{k}+M-1)}} \sum_{l=1}^{P} \gamma_l^2[y_d(\bar{k}+\tau+l) - y(\bar{k}+\tau+l|\bar{k})]^2 + \beta_l^2 m(\bar{k}+l-1)^2 \qquad (32)$$

subject to

$$y(k+\tau|\bar{k}) = y_M(k+\tau) + d(k+\tau|\bar{k}) = h_1 m(k-1) + h_2 m(k-2) + \ldots + h_N m(k-N) + d(k+\tau|\bar{k})$$

$$m(\bar{k}+M-1) = m(\bar{k}+M) = \ldots = m(\bar{k}+P-1)$$

$$\beta_l^2 = 0; \quad l > M$$

where P is the horizon ($P \geq 1$), $y_d(\bar{k}+\tau+l)$ is the desired trajectory, γ_l^2 are time-varying weights on the output error, β_l^2 are time varying weights on the input, M is the input

$$
\begin{aligned}
\gamma_1\epsilon(\bar{k}+1) &= [h_1 m(\bar{k}) + h_2 m(\bar{k}-1) + \ldots + h_N m(\bar{k}-N+1)]\gamma_1 \\
\gamma_2\epsilon(\bar{k}+2) &= [h_2 m(\bar{k}) + h_1 m(\bar{k}+1) + h_3 m(\bar{k}-1) + \ldots + h_N m(\bar{k}-N+2)]\gamma_2 \\
&\vdots \\
\gamma_N\epsilon(\bar{k}+N) &= [h_N m(\bar{k}) + h_{N-1} m(\bar{k}+1) + \ldots + h_1 m(\bar{k}+N-1)]\gamma_N \\
\gamma_{N+1}\epsilon(\bar{k}+N+1) &= [h_N m(\bar{k}+1) + \ldots + (h_1+h_2) m(\bar{k}+N-1)]\gamma_{N+1} \\
&\vdots \\
\gamma_P\epsilon(\bar{k}+P) &= [(\sum_{i=1}^{N} h_i) m(\bar{k}+N-1)]\gamma_P \\
0 &= \beta_1 m(\bar{k}) \\
0 &= \beta_2 m(\bar{k}+1) \\
&\vdots \\
0 &= \beta_N m(\bar{k}+N-1)
\end{aligned}
\qquad (34)
$$

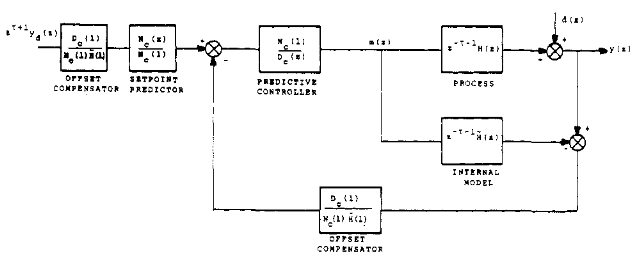

Figure 9. IMC structure with predictive controller and offset compensator.

suppression parameter which specifies the number of intervals into the future during which $m(k)$ is allowed to vary; it is kept constant afterwards, $y(k+\tau|\bar{k})$ is the predicted output, $y_M(k+\tau)$ is the output of the internal model, and $d(k+\tau|\bar{k})$ is the predicted disturbance.

The best prediction of the disturbance is to set it equal to the disturbance at the present time

$$d(k+\tau|\bar{k}) = d(\bar{k}) \qquad \text{for all } k > \bar{k} \qquad (33a)$$

$d(\bar{k})$ is the feedback signal in the internal model controller (Figure 2) which is obtained from the current measurement and the current output of the internal model

$$d(\bar{k}) = y(\bar{k}) - y_M(\bar{k}) \qquad (33b)$$

If only model prediction is used and disturbances are omitted from the above problem one obtains the linear quadratic feedback control problem. Its inability to handle persistent disturbances is a direct consequence of this omission.

M, P, β_l, and γ_l are tuning parameters of the algorithm which have a direct influence on stability and dynamic response. The theoretical basis for their selection will be given in section 4.2.

4.1.2. Control Law Computation. For simplicity we will restrict $M \le N$ so that inputs have an effect over a total horizon $P = N + M - 1$. This selection has no influence on the general results and extension to other values is straightforward. Defining the error as $\epsilon(\bar{k}+l) = y_d(\bar{k}+\tau+l) - d(\bar{k}+\tau+l|\bar{k})$, we may observe that the least-squares solution of eq 34 is equivalent to the solution of (32) for $M = N$.

In eq 34 the contribution of past inputs (over which we have no control) lies to the right of the dashed line. The input suppression constraints ($M < N$) can be included in the formulation by collecting the corresponding input terms, so that the equations can be written concisely in matrix form as

$$\begin{bmatrix} \Gamma_P\epsilon_P(\bar{k}+1) \\ \text{------} \\ 0 \end{bmatrix} = \begin{bmatrix} \Gamma_P R_P \Lambda T_M \\ \text{------} \\ B_M \end{bmatrix} U_M(k) + \begin{bmatrix} \Gamma_P\psi \\ \text{------} \\ 0 \end{bmatrix} V(k-1) \qquad (35)$$

where

$$\epsilon_P(\bar{k}+1) = \begin{bmatrix} \epsilon(\bar{k}+1) \\ \epsilon(\bar{k}+2) \\ \vdots \\ \epsilon(\bar{k}+P) \end{bmatrix};$$

$$U_M(\bar{k}) = \begin{bmatrix} m(\bar{k}) \\ m(\bar{k}+1) \\ \vdots \\ m(\bar{k}+M-1) \end{bmatrix}; \quad V(\bar{k}-1) = \begin{bmatrix} m(\bar{k}-1) \\ m(\bar{k}-2) \\ \vdots \\ m(\bar{k}-N+1) \end{bmatrix}$$

$$R_P = \left[\, I_P \;\vdots\; 0 \,\right]; \quad T_M = \begin{bmatrix} I_M \\ \text{------} \\ \begin{array}{c:c} & 1 \\ & 1 \\ 0 & \vdots \\ & 1 \end{array} \end{bmatrix}; \quad B_M = \operatorname{diag}(\beta_1 \ldots \beta_M); \quad \Gamma_P = \operatorname{diag}(\gamma_1 \ldots \gamma_P);$$

$$\Lambda = \begin{bmatrix} h_1 & & & 0 \\ h_2 & h_1 & & \\ \vdots & & \ddots & \\ h_N & h_{N-1} & \ldots & h_1 \\ 0 & h_N & \ldots & (h_1+h_2) \\ 0 & 0 & h_N & \vdots \\ & & \ddots & \vdots \\ & 0 & & \sum_{i=1}^{N} h_i \end{bmatrix}; \quad \psi = [I_P \;\vdots\; 0] \begin{bmatrix} h_2 & h_3 & \cdots & h_N \\ h_3 & h_4 & \cdots h_N & 0 \\ \vdots & & & \\ h_N & & & 0 \end{bmatrix}$$

and $\mathbf{I}_j$ is an identity matrix of dimension $j \times j$.

The matrix T_M accounts for the effect of M on the optimization. For example, for $M = 1$ ($\gamma_i = 1, \beta_i = 0$), which means that the input applied is kept constant throughout, the equations become simply

$$\begin{bmatrix} \epsilon(\bar{k}+1) \\ \epsilon(\bar{k}+2) \\ \cdot \\ \cdot \\ \cdot \\ \epsilon(\bar{k}+N) \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_N \end{bmatrix} m(\bar{k}) + \Psi V(\bar{k}-1)$$

where a_j are the step response coefficients defined in section 2. We stress the fact that since the problem is solved at each time, the actual inputs applied to the process are not kept constant after M intervals of time. The factor M is an artifice of the problem statement included to produce a more cautious controller.

The input to be implemented is obtained by selecting the first element of the least-squares solution of (35) yielding the control law

$$m(\bar{k}) = \boldsymbol{b}^T \boldsymbol{U}_M(\bar{k})$$
$$= \boldsymbol{b}^T(\mathbf{T}_M{}^T\Lambda^T\mathbf{R}_P{}^T\Gamma_P{}^2\boldsymbol{R}_P\Lambda\mathbf{T}_M + \mathbf{B}_M{}^2)^{-1}\,\mathbf{T}_M{}^T\Lambda^T\mathbf{R}_P{}^T\Gamma_P{}^T(\epsilon_P(\bar{k}+1) - \Gamma_P\Psi V(\bar{k}-1))$$
$$\boldsymbol{b}^T = (10 \ldots 0) \quad (36)$$

which has the closed form

$$m(\bar{k}) + \sum_{l=1}^{N-1}\delta_l m(\bar{k}-l) = \sum_{l=1}^{P}\nu_l\{y_d(\bar{k}+\tau+l) - (y(\bar{k}) - y_M(\bar{k}))\} \quad (37)$$

The general form of the control law in Z-transforms is therefore

$$D_c(z)m(z) = z^{\tau+1}N_c(z)y_d(z) - N_c(1)[y(z) - y_M(z)] \quad (38)$$

where

$$D_c(z) = 1 + \delta_1 z^{-1} + \ldots + \delta_{N-1}z^{-N+1}$$
$$N_c(z) = \nu_1 + \nu_2 z + \ldots + \nu_P z^{P-1}$$

Stability of the scheme is determined by the roots of $D_c(z)$. By defining the backward shift operator $q = z^{-1}$ we equivalently require the roots of the characteristic equation

$$C(q) = 1 + \boldsymbol{b}^T(T_M{}^T\Lambda^T R_P{}^T\Gamma_P{}^2 R_P\Lambda T_M + B_M{}^2)^{-1}T_M{}^T\Lambda^T R_P{}^T\Gamma_P{}^2\boldsymbol{\Psi}\begin{bmatrix} q \\ q^2 \\ \cdot \\ \cdot \\ \cdot \\ q^{N-1}\end{bmatrix} = 0 \quad (39)$$

to lie *outside* the unit circle for stability.

The block diagram representation of the control law is shown in Figure 9. The offset compensator is introduced in accordance with Property 3. Note the explicit inclusion of future setpoint values because of the predictive nature of the polynomial $N(z)$. The following theorems show the effect of M, P, β_i, and γ_i on the stability of $D_c(z)$ and thus provide guidelines for their selection. Once chosen, matrices in eq 36 are completely specified allowing the computation of the control law.

4.2. Stability Theorems. The perfect controller is a special case of (36) as shown by the following theorem.

Theorem 1. For $\gamma_i \neq 0$, $\beta_i = 0$, selecting $M = P \leq N$ yields the model inverse control $G_c(z) = H(z)^{-1}$.

Proof. When $M = P \leq N$, $\gamma_i \neq 0$, $\beta_i = 0$ the set of eq 34 can be solved exactly for $m(k)$, $m(k+1)$, ..., $m(k+M-1)$. Since the matrix to be inverted is triangular, the first equation gives the desired control law

$$m(k) = \frac{1}{h_1}\{\epsilon_1(k+1) - \sum_{l=2}^{N}h_l m(k-l+1)\}$$

or $H(z)m(z) = z^{\tau+1}y_d(z) - (y(z) - y_M(z))$. Thus, $G_c(z) = N_c(1)/D_c(z) = H(z)^{-1}$ and $N_c(z) = 1/h_1$. Q.E.D. The following obvious corollary is included to emphasize the effect of NMP characteristics on stability.

Corollary 1. If the zeroes of $H(z)$ lie outside the unit circle, for $\gamma_i \neq 0$, $\beta_i = 0$, $M = P \leq N$ the control law (36) is unstable.

We next prove several theorems which demonstrate the stabilizing effect of the different tuning parameters in the presence of NMP characteristics of the plant.

Theorem 2. Assume that the system has a discrete monotonic step response and that $\gamma_i = 1$, $\beta_i = 0$, $P = N$. For M chosen sufficiently small the control law (36) is stable. (Note: a discrete monotonic step response does *not* imply that all the roots of $H(z)$ are inside the unit circle.)

Proof. A process with discrete monotonic step response has $h_i > 0$ (or < 0), $i = 1, \ldots, N$. We will establish the result by showing stability for $M = 1$.

If $\gamma_i = 1$, $\beta_i = 0$, $M = 1$, $P = N$, the characteristic eq 39 reduces to

$$C(q) = (\sum_{i=1}^{N}a_i{}^2) + (\sum_{i=1}^{N-1}a_i h_{i+1})q + \ldots + (\sum_{i=1}^{2}a_i h_{N+i-2})q^{N-2} + a_1 h_N q^{N-1} = 0 \quad (40)$$

where $a_k = \sum_{i=1}^{k} h_i$ are the step response coefficients. Starting with the last ($a_1 h_N$) one can verify that all the terms in each coefficient are included in the next. Consequently, if $h_i > 0$, $a_i > 0$ the monotonic conditions in lemma 1 of Appendix A are satisfied, and thus $C(q)$ has all roots outside the unit circle. If $h_i < 0$ then all $a_i < 0$ and the terms in the coefficients of (40) are all positive, again implying the result. Q.E.D.

Monotonic step response systems are not uncommon in chemical processes, which are characterized by high order, over-damped responses. Besides, by using a sufficiently large sampling time all h_i's can be made positive for NMP systems which exhibit inverse response. Consequently, theorem 2 ensures the existence of a stable approximate inverse of $H(z)$ for a large class of systems.

In case a discrete monotonic response cannot be obtained, the following theorem guarantees stability for any system when a sufficient penalty on the input is introduced.

Theorem 3. There exists a finite $\beta^* > 0$ such that for $\beta_i \geq \beta^*$ ($i = 1, \ldots, M$) the control law (36) is stable for all $M \geq 1$, $P \geq 1$, and $\gamma_i \geq 0$.

Proof. We will show the result for $\beta_i = \beta$, $i = 1, \ldots, M$. Then by selecting $\beta_i \geq \beta$ the control law will remain stable.

The characteristic eq 39 is

$$1 + \delta_1 q + \ldots + \delta_{N-1}q^{N-1} = C(q)$$

where

$$\begin{bmatrix} \delta_1 \\ \delta_2 \\ \cdot \\ \cdot \\ \cdot \\ \delta_{N-1}\end{bmatrix}^T = b^T(\Delta^T\Delta + B_M{}^2)^{-1}\Delta^T\bar{\boldsymbol{\Psi}};$$

$$\Delta = \Gamma_P R_P \Lambda T_M; \quad \bar{\boldsymbol{\Psi}} = \Gamma_P \Psi$$

Defining $\epsilon = 1/(N-1)$ and $\boldsymbol{d}_i = [\bar{\delta}_{i1}, \bar{\delta}_{i2}, ..., \bar{\delta}_{iN-1}]^T$

$$\bar{\delta}_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

then

$$|\delta_i| = |\boldsymbol{b}^T(\Delta^T\Delta + \beta^2\mathbf{I})^{-1}\Delta^T\bar{\Psi}\boldsymbol{d}_i| = \frac{1}{\beta^2}\left|\boldsymbol{b}^T\left(\frac{1}{\beta^2}\Delta^T\Delta + \mathbf{I}\right)^{-1}\Delta^T\bar{\Psi}\boldsymbol{d}_i\right| \leq \frac{1}{\beta^2}\|\boldsymbol{b}^T\|\,\|\Delta^T\bar{\Psi}\boldsymbol{d}_i\|\,\left\|\mathbf{I} - \frac{1}{\beta^2}\Delta^T\Delta - \frac{1}{\beta^4}(\Delta^T\Delta)^2 - ...\right\|$$

where we have expanded the inverse in a Newmann series. Thus

$$|\delta_i| \leq \frac{1}{\beta^2}\|\boldsymbol{b}^T\|\,\|\Delta^T\bar{\Psi}d_i\|\left\{1 + \frac{1}{\beta^2}\|\Delta^T\Delta\| + \frac{1}{\beta^4}\|\Delta^T\Delta\|^2 + ...\right\}$$

Assuming β^2 is chosen sufficiently large such that $\beta^2 > \|\Delta^T\Delta\|$ the above series will converge. Then there exists a $\beta^2 > \|\Delta^T\Delta\|$ such that $|\delta_i| < \epsilon$ for any $\Delta,\bar{\Psi}$ and consequently

$$\sum_{i=1}^{N-1}|\delta_i| < (N-1)\epsilon = 1$$

It follows then from lemma 2 of Appendix A that the polynomial

$$1 + \delta_1 q + ... + \delta_{N-1}q^{N-1} = C(q)$$

has all $N-1$ roots outside the unit circle. Q.E.D.

We finally give a result which demonstrates the stabilizing effects of the horizon limit $P > N + M - 1$.

Theorem 4. Assume $\gamma_i = 1$, $\beta_i = 0$. Then for a sufficiently small M and a sufficiently large $P > N + M - 1$ the controller (36) is stable.

Proof. The result is established by showing stability for $M = 1$ and $P > N$ sufficiently large. For any $P > N$, and $M = 1$ the characteristic eq 39 becomes

$$C(q) = \left(\sum_{i=1}^{N}(\gamma_i a_i)^2 + \left(\sum_{N+1}^{P}\gamma_i^2\right)a_N^2\right) + \left(\sum_{i=1}^{N-1}a_i\gamma_i^2 h_{i+1}\right)q + ... + \left(\sum_{i=1}^{2}a_i\gamma_i^2 h_{N+i-2}\right)q^{N-2} + a_1\gamma_1^2 h_N q^{N-1} = 0$$

Hence, since $\gamma_i^2 = 1$, we can find a $P > N$ large enough such that

$$\left|\sum_{i=1}^{N}(\gamma_i a_i)^2 + (P-N)a_N^2\right| > \left|\sum_{i=1}^{N-1}a_i h_{i+1}\right| + ... + \left|\sum_{i=1}^{2}a_i h_{N+i-2}\right| + |a_1 h_N|$$

and consequently by lemma 2 (Appendix A) all roots of $C(q)$ lie outside the unit circle. Q.E.D.

One may look at the effect of a $P > N$ as equivalent to introducing a penalty on the inputs as may be seen from the problem formulation in eq 35. Thus the results of theorem 4 can be interpreted in the light of theorem 3. Moreover, it is evident from the proof that the restriction $\gamma_i = 1$ can be eased. For stability it is only required that the sum $\sum^{P}_{N+1}\gamma_i^2$ be sufficiently large.

4.3. Implementation of IMC. 4.3.1. Tuning Procedures. We should distinguish two cases: Case A. The plant model $H(z)$ is nonminimum phase, some zeroes of $H(z)$ are outside the unit circle, and $H(z)^{-1}$ is unstable. Case B. $H(z)$ is minimum phase and $H(z)^{-1}$ is stable. From theorem 1 we know that for $\beta_i = 0$, $\gamma_i \neq 0$, $M = P \leq N$ the control law (32) is equal to $H(z)^{-1}$. By changing the tuning parameters we obtain different approximations of $H(z)^{-1}$. In case A the first goal is to obtain a stable approximation; in the second step we modify it to obtain a more desirable response. In case B stability is not an issue and only the quality of the response is affected.

In view of this it is advisable to start any tuning procedure by checking the roots of the equation $H(z) = 0$ employing a polynomial root finder. Actually, since all we need is the modulus of the largest (smallest) root of $D_c(z)$ ($C(q)$) to establish stability, an algorithm to find the spectral radius of its associated companion matrix as, for example, by successive substitutions, is sufficient (Faddeeva, 1959). Due to encountered convergence difficulties this approach was not used here.

(1) Sampling Time (T). Case A: Frequent sampling of a continuous process with RHP zeroes produces an $H(z)$ with roots outside the unit circle. It is straightforward to show, however, that if the sampling time is sufficiently increased those roots are made stable.

Another option is to increase T just enough to ensure $h_i > 0$ ($i = 1, ..., N$) and to obtain a stable approximation by decreasing M according to theorem 2. Case B: Contrary to conventional feedback control the stability of IMC is not affected by T. Larger T's generally lead to less extreme excursions of the manipulated variable, but they have a detrimental affect on the ability of the system to handle frequent disturbance changes.

(2) Input Suppression Parameter (M). Case A: According to theorem 2 any system with $h_i > 0$ ($i = 1, ..., N$) can be stabilized by choosing M sufficiently small. Case B: Perfect control ($M = P = N$) requires severe variations in the input variable and usually leads to strong oscillations of the output between the sampling times. Reducing M reduces the extreme excursion of the manipulated variables and thus leads to a more desirable response of the plant.

(3) Input Penalty Parameter (β_i). Case A: According to theorem 3 we can always obtain a stable approximation of $H(z)^{-1}$ by making β_i sufficiently large. Case B: Increasing β_i will decrease the actions taken by the manipulated variables and make the system more sluggish. Furthermore, $\beta_i \neq 0$ will in general lead to offset since in that case $N_c(1)/D_c(1) \neq 1/\tilde{H}(1)$. This is corrected by the offset compensators shown in Figure 9.

An alternate procedure is to formulate the problem by penalizing control action increments thus minimizing the objective

$$\sum_{l=1}^{P}\gamma_l^2(y_d(k+\tau+l) - y(k+\tau+l|k))^2 + \beta_l^2(m(k+l-1) - m(k+l-2))^2 \quad (41)$$

then it is straightforward to show that $N_c(1)/D_c(1) = 1/\tilde{H}(1)$ for any β_l. However, excessive weighting would yield a sluggish controller since it puts the largest root of $D_c(z)$ on the unit circle. It is important to note that by penalizing control action increments only it is generally *not* possible to obtain a stable approximation of $H(z)^{-1}$ for NMP systems. An example is presented in Appendix B.

(4) Output Penalty Parameter (γ_i). We first note that the use of γ_i is only meaningful when these parameters are time varying. Otherwise one should choose $\gamma_i = 1$. The motivation for introducing γ_i arises from the work of Reid et al. (1979). As we remarked repeatedly, $G_c(z) = H(z)^{-1}$ leads to large excursions of the input and strong oscillations of the output between the sampling intervals. The reason for this undesirable behavior is that the controller drives only the *output* to zero without taking care of the other system states (e.g., derivatives of the output). It is well known that any nth order system can be brought precisely to rest in no more than n discrete steps. This suggests to postulate the following predictive control law: "Let the inputs be variable over the next n steps and zero afterwards. Determine the inputs such that the outputs at steps $n, n+1, ..., 2n-1$ are zero". This implies the following choice of tuning parameters for a system of order n: $P = 2n-1$; $M = n+1$; $\beta_1 = \beta_2 = ... = \beta_n = 0$; β_{n+1} = large; $\gamma_1 = \gamma_2 = ... = \gamma_{n-1} = 0$; $\gamma_n = \gamma_{n+1} = ... = \gamma_{2n-1} = 1$. These values resulted in superb performance in our test simulations. If just the impulse response of a system is identified the order n is unknown. Nevertheless, we have reduced the tuning procedure to the choice of this one parameter n from which all other parameters can be determined.

(5) Optimization Horizon (P). Case A: Any system can be stabilized by selecting M sufficiently small and P sufficiently large. Case B: For minimum phase systems the length of the optimization horizon has virtually no effect on the performance as soon as P exceeds about twice the system order.

4.3.2. On-Line Computations. The control law is of the form

$$m(k) = -\sum_{l=1}^{N-1} \delta_l m(k-l) + \sum_{l=1}^{P} \nu_l \{y_d(k+\tau+l) - (y(k) - y_M(k))\} \quad (37)$$

Once the tuning parameters are selected the coefficients δ_i, ν_i could be calculated off-line from (36). Alternatively, if on-line tuning is desired (36) can be solved on-line, requiring the inversion of a matrix of dimension M each time the tuning parameters are modified. If only some of the tuning parameters are changed the inversion of a matrix of lower order can be sufficient after proper partitioning. We can also express the control law (37) in velocity form to provide bumpless transfer from manual to computer control

$$\Delta m(k) + \sum_{i=1}^{N-1} \delta_i \Delta m(k-i) = \sum_{i=1}^{P} \nu_i \{\Delta y_d(k+\tau+i) - \Delta(y(k) - y_M(k))\}$$

where $\Delta x(k) = x(k) - x(k-1)$. Observe that past input changes need to be stored and the future desired trajectory specified for implementation.

For a reduction of the computational burden it is further possible not to move the horizon after each step but to implement the first k control moves before redoing the calculations. This, however, makes the performance more sensitive to modelling errors.

It was also suggested to "block" the inputs for the computations, i.e., to allow them to vary only each kth step. With proper weighting excellent results were achieved (Reid et al., 1980).

4.4. Special Cases. Two recently developed algorithms which have been reported to perform well in chemical process applications are discussed here. In fact, we demonstrate both schemes to be special cases of IMC for specific choices of tuning parameters. Their excellent performance is explained in the light of the exposed properties of IMC.

4.4.1. Model Algorithmic Control (MAC). MAC was developed in France in the late 60's inside the chemical process industry. Mainly a heuristic algorithm (then called Model Predictive Heuristic Control), it was interpreted as a dual of the least-squares identification method and it was not until the past few years, after several successful applications, that special attention has been given to its theoretical properties. The original concept is expounded by Richalet et al. (1978) and the theory studied by Mehra et al. (1980), with special attention to nonminimum phase systems given by Mehra and Rouhani (1980). Applications range from steam generator control (Lecrique et al., 1978), electric power plants (Mehra and Eterno, 1980), to flight control (Mehra et al., 1978). A review of the developments is presented in Mehra et al. (1981).

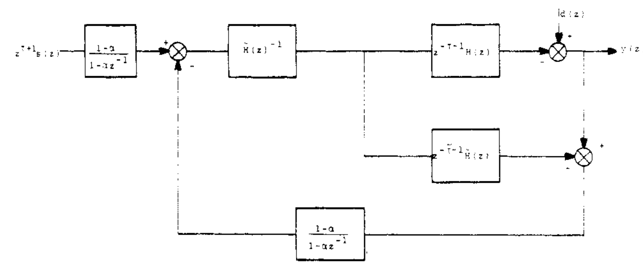

Figure 10. MAC with reference trajectory.

The control law is obtained from the optimization problem

$$\min_{m(k)} [y_d(k+\tilde{\tau}+1) - y(k+\tilde{\tau}+1|k)]^2 \quad (42)$$

subject to

$$y_M(k+\tilde{\tau}+1) = \sum_{l=1}^{N} \tilde{h}_l m(k+1-l)$$

which is a one-step ahead predictive controller. A reference trajectory is defined as

$$y_d(k+\tilde{\tau}+1) = (1-\alpha)s(k+\tilde{\tau}+1) + \alpha y(k+\tilde{\tau}|k)$$

where $0 \leq \alpha < 1$ specifies an exponential approach from the predicted output $\tilde{\tau}$ time intervals ahead to the setpoint $s(k+\tau+1)$. The prediction is just

$$y(l+\tilde{\tau}|k) = \sum_{j=1}^{N} \tilde{h}_j m(l-j) + [y(k) - y_M(k)]$$

and we note that $y(k) - y_M(k)$ is the estimated disturbance at time k (eq 33).

Solution of (42) yields the control law

$$m(z) = \tilde{H}(z)^{-1}[(1-\alpha)z^{\tilde{\tau}+1}s(z) + \alpha\{z^{-1}\tilde{H}(z)m(z) + (y(z) - y_M(z))\} - (y(z) - y_M(z))] \quad (43)$$

which is IMC's model inverse controller with an exponential trajectory $y_d(k+\tilde{\tau}+1)$ (see theorem 1). The resemblance with the general IMC structure is more obvious if we rearrange (43) in the form

$$m(z) = \tilde{H}(z)^{-1}\left(\frac{1-\alpha}{1-\alpha z^{-1}}\right)\{z^{\tilde{\tau}+1}s(z) - (y(z) - y_M(z))\} \quad (44)$$

with block diagram shown in Figure 10. Observe that this is precisely the structure in Figure 3 with

$$R(z) = F(z) = \frac{1-\alpha}{1-\alpha z^{-1}}$$

$$G_c(z) = \tilde{H}(z)^{-1}$$

Thus MAC's reference trajectory is equivalent to using an exponential filter in the feedback path of IMC. By adjusting the time constant of the reference trajectory the operator can affect the robustness of the controller in a direct manner. As would be expected, a slow reference trajectory makes the scheme more robust.

As we already know, the perfect controller obtained in MAC is unstable for NMP systems and thus modifications have to be introduced. From eq 24 the characteristic equation for this scheme is

$$(1 - \alpha z^{-1})\tilde{H}(z) + (1 - \alpha)[z^{-\tau-1}H(z) - z^{-\tilde{\tau}-1}\tilde{H}(z)] = 0 \quad (45)$$

and consequently when $\tilde{\tau} \neq \tau$, $\tilde{H}(z) \neq H(z)$, α could be manipulated in the attempt to put the roots of (45) inside the unit circle. However, for exact model description (45) becomes $(z - \alpha)H(z) = 0$ and thus α affects the dynamic characteristics of the system but has no influence on the "unstable" zeroes of $H(z)$.

Mehra and Rouhani (1980) propose two methods for handling NMP systems. Their pole placement method factors $H(z)$ as

$$H(z) = H_s(z)H_u(z)$$

where $H_s(z)$ and $H_u(z)$ collect the stable and unstable roots of $H(z)$, respectively. By choosing a suitable polynomial factor $P_M(z)$ for the internal model, $\tilde{H}(z) = H_s(z)P_M(z)$, one can make the roots of $H_s(z)\{(1 - \alpha)H_u(z) + (z - 1)P_M(z)\} = 0$ stable. This method has two disadvantages. One is that there are no guidelines for choosing P_M to obtain a good dynamic response. The second one is more serious: the internal model used for output prediction is changed—it barely resembles the plant model—and thus the main attractive feature of IMC, i.e., subtracting the effect of the inputs from the measurements, is destroyed.

The other method they suggest also changes the internal model. The internal model $\tilde{H}(z)$ (and the corresponding controller $\tilde{H}(z)^{-1}$) are selected as "close" as possible to the plant $H(z)$ but such that the roots of (45) lie inside the unit circle. Mathematically, the problem can be stated as

$$\min_{\tilde{h}_1,\tilde{h}_2 \ldots \tilde{h}_N} \frac{1}{2}\sum_{l=1}^{\infty}[\sum_{j=1}^{N} h_j m(l - j) - y_d(l + \tau)]^2 \quad (46)$$

where $m(l - j)$ is the system input obtained from the MAC control law (44). Problem (46) is solved off-line using optimal control theory requiring the solution of a Riccati equation. Although "optimal", solving a Riccati equation of order N (number of terms used in the impulse response $\approx$ 40–80) presents quite a computational burden. This is specially inconvenient because the Riccati equation has to be solved each time the on-line tuning parameter α is modified.

In terms of IMC, the approximation to the inverse we suggest either by reducing M or increasing β_l provides no additional complications and conserves the inferential structure. In addition, the approach is also optimal for a given optimization horizon.

Before concluding this discussion we should mention a recent paper by Åström (1980) where robustness of MAC is demonstrated with a simplified first-order truncation model

$$\tilde{H}(z) = z^{-1}bm(z)$$

Stability is proven for continuous monotonic response systems although by proper selection of the sampling time it should be possible to handle inverse response systems as well (see discussion on sampling time T above).

4.4.2. Dynamic Matrix Control (DMC). Using a step response model, Cutler and Ramaker (1979) of Shell Oil Co. developed an algorithm which they applied successfully to the control of chemical processes such as catalytic-cracking units. The scheme is a predictive-type controller capable of following a time varying trajectory and exhibiting zero offset to persistent disturbances. It was used by Prett and Gillette (1979) also of Shell in a constraint control optimizer with excellent results.

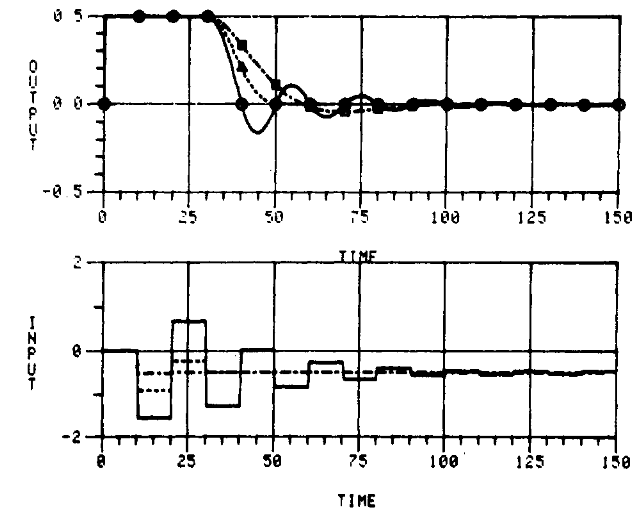

Figure 11. Effect of $M < N$ for the MP system (§5.1): (○), $M = 10$, $\rho = 0.77$; (△), $M = 2$, $\rho = 0.54$; (□), $M = 1$, $\rho = 0.41$ ($P = N = 10$, $\beta_l = 0$, $\gamma_l = 1$, $\alpha = 0$).

The scheme is precisely the IMC predictive controller with $P = N$ and $M < N$. Due to the model description employed, input *changes* are penalized as in eq 41.

Our new results provide convenient tuning guidelines for DMC and also point out many generalizations for performance improvement.

5. Examples

In order to demonstrate the features of IMC several simulation examples were run on a digital computer. In all cases the systems start at steady state ($y_s = m_s = 0$) and a disturbance $d(1)$ of magnitude 0.5 is introduced. The sampling time (T) is chosen as 10 units with a delay of $\tau = 2$ ($\theta = 20$). We include the magnitude of the largest root of $D_c(z)$ (ρ) in all results for comparison.

5.1. Minimum Phase System. A second-order process with transfer function

$$G(s) = e^{-20s}/(100s^2 + 12s + 1)$$

is controlled using IMC. We would like to compare the responses obtained using the model inverse controller, $G_c(z) = H(z)^{-1}$, vs. the proposed inverse approximations. The 10th-order truncation model ($N = 10$)

$$y(k + 1) = 0.322m(k - 2) + 0.461m(k - 3) + 0.255m(k - 4) + 0.056m(k - 5) - 0.034m(k - 6) - 0.043m(k - 7) - 0.023m(k - 8) - 0.004m(k - 9) - 0.003m(k - 10) - 0.002m(k - 11)$$

was used to obtain sufficient accuracy.

Effect of M. Using $\gamma_l = 1$, $\beta_l = 0$, and $P = 10$, the model inverse controller is given by $M = 10$. As shown in Figure 11, it gives perfect setpoint compensation after $\tau + 1 = 3$ discrete time intervals but at the expense of severe input variations. Reducing M to values of 2 and 1 gives better input dynamics and less oscillations in the output although the setpoint is not satisfied exactly at the sampling periods.

Effect of Penalty on Inputs m. For $M = P = 10$ a constant penalty parameter of $\beta_l = 5$ was tried. Without proper compensation, a significant offset is observed (see Figure 12). Introducing the offset compensator (Figure 9) a smooth response is obtained with good input dynamics.

Effect of Penalty on Δm. Using the same weight as in Figure 12 ($\beta_l = 5$), a penalty on input changes $\Delta m(k)$ yields a sluggish controller (Figure 13). A better response

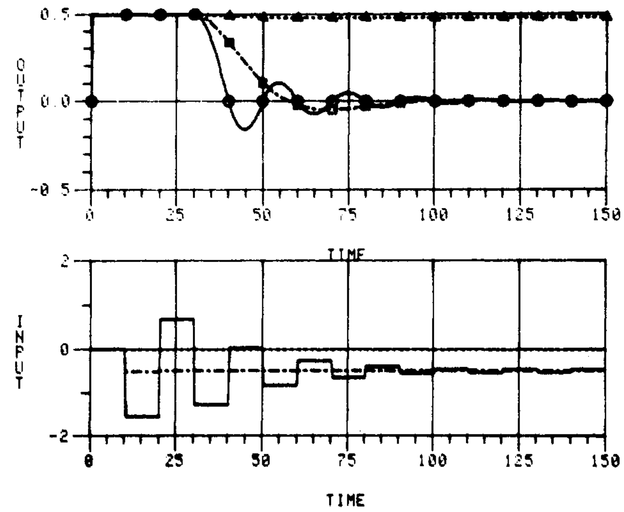

Figure 12. Effect of penalty on $m(k)$ for system (§5.1): (O), $\beta = 0$, $\rho = 0.77$; (△), $\beta = 5$, $\rho = 0.52$ no offset compensation; (□), $\beta = 5$, $\rho = 0.52$ with offset compensation ($P = M = N = 10$, $\gamma_l = 1$, $\alpha = 0$).

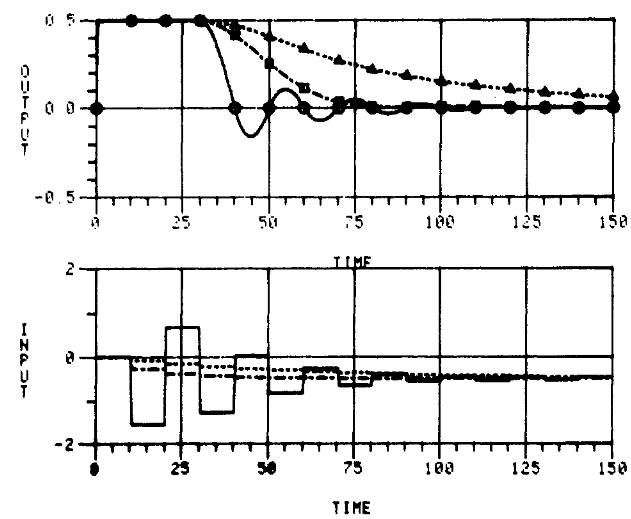

Figure 13. Effect of penalty on $\Delta m(k)$ for system (§5.1): (O), $\beta = 0$, $\rho = 0.77$; (△), $\beta = 5$, $\rho = 0.83$; (□), $\beta = 1.5$, $\rho = 0.58$ ($P = M = N = 10$, $\gamma_l = 1$, $\alpha = 0$).

is obtained by using a $\beta_l = 1.5$. Note the difference in input dynamics as compared to Figure 12.

Effect of Reference Trajectory. Figure 14 shows the response when α is increased. The exponential discrete approach to the setpoint causes decreased input amplitudes; however, a much smoother response is obtained by reducing M or by using input penalties. This is the result of α not being able to move the roots of $H(z)$.

Effect of Penalty on Outputs. According to our discussion above we choose the following tuning parameters for this second-order system. $P = M = 3$, $\beta_1 = \beta_2 = 0$, β_3 = very large, $\gamma_1 = 0$, $\gamma_2 = \gamma_3 = 1$. The excellent performance of this control scheme for different reference trajectories is shown in Figure 15. Note that DMC with $M = 2$ and $P = N$ (Figure 11) is almost as good as dead beat control. This is to be expected due to the similar way the input moves are restricted in both schemes.

5.2. Nonminimum Phase System. 5.2.1. Discrete Monotonic Response System. Stabilizability with M. The following system with a RHP zero

$$G(s) = \frac{1 - 5s}{(10s + 1)^2} e^{-20s}$$

was controlled using the IMC approach. Figure 16 shows

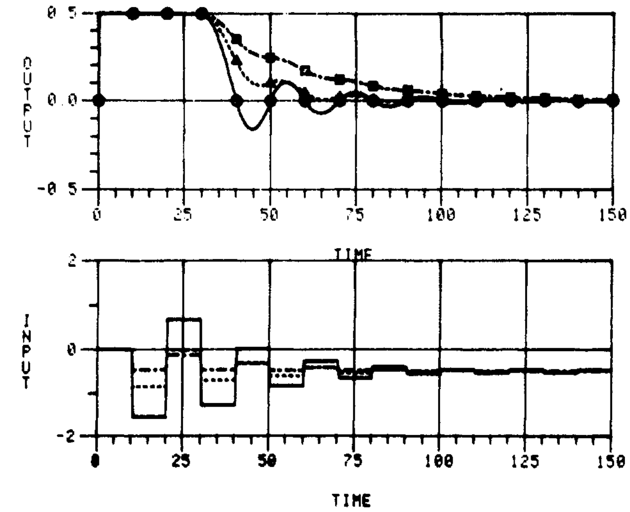

Figure 14. Reference trajectory for system (§5.1): (O), $\alpha = 0$; (△), $\alpha = 0.45$; (□), $\alpha = 0.70$ ($P = M = N = 10$, $\beta_l = 0$, $\gamma_l = 1$, $\rho = 0.77$).

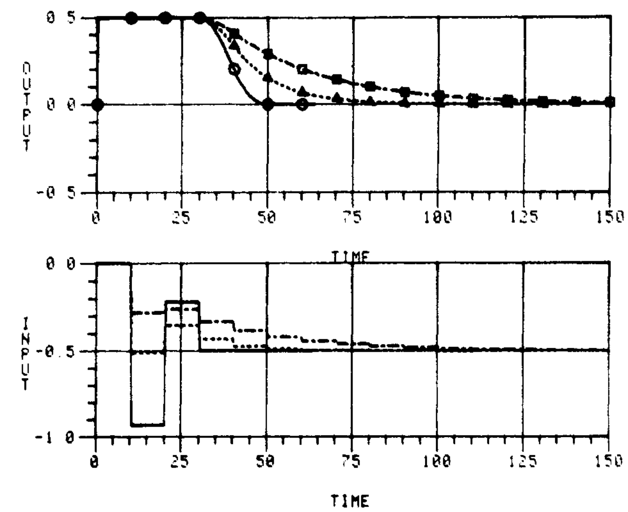

Figure 15. Dead beat controller with reference trajectory for system (§5.1): (O) $\alpha = 0$; (△), $\alpha = 0.45$; (□), $\alpha = 0.70$ ($n = 2$ in §4.3.1, no. 4).

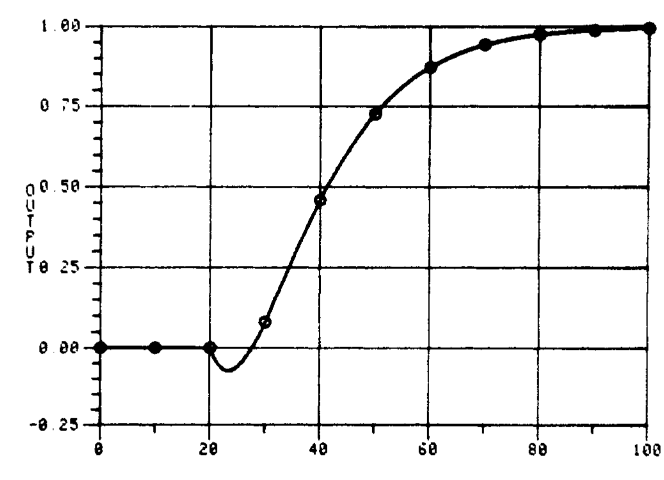

Figure 16. Unit step response for NMP system in §5.2.1: (O), indicate discrete samples for $T = 10$.

the response to a unit step change. In order to guarantee stability for a sufficiently small M a sampling time of 10 units was chosen producing a discrete monotonic response. The largest root of $H(z)$ (using $N = 10$) has a magnitude of 3.97 for this sampling time and hence the inverse con-

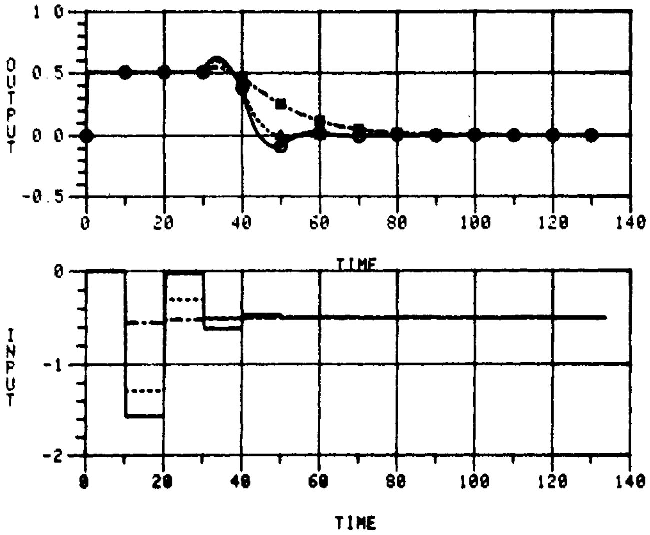

Figure 17. Stable controllers for NMP system for Figure 16 by reducing M: (○), M = 9, ρ = 0.51; (△), M = 2, ρ = 0.42; (□), M = 1, ρ = 0.36 ($P = N = 10$, $\beta_l = 0$, $\gamma_l = 1$, $\alpha = 0$).

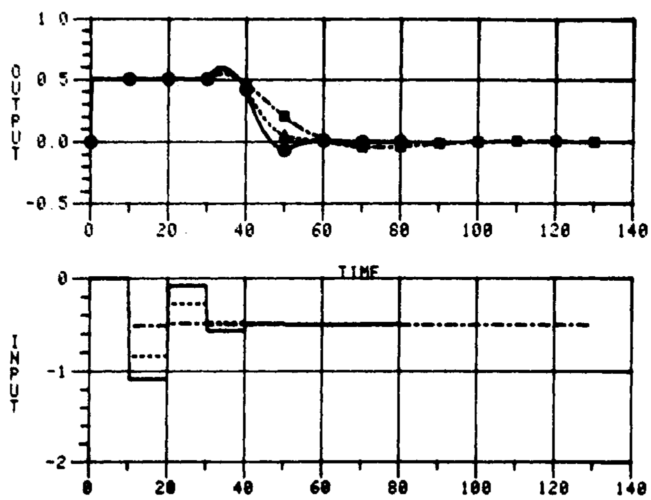

Figure 18. Stable controllers for nonmonotonic NMP system of §5.2.2 by reducing M: (○), M = 9, ρ = 0.62; (△), M = 2, ρ = 0.52; (□), M = 1, ρ = 0.40 ($P = N = 10$, $\beta_l = 0$, $\gamma_l = 1$, $\alpha = 0$).

troller is unstable. Figure 17 shows the stable responses obtained when M is reduced. We should mention that using a sampling time of 5 which yields an $h_1 < 0$, a stable controller is obtained for M = 1. Also, for T = 20 the magnitude of the largest root of $H(z)$ is brought down to 0.62 and thus the exact model inverse controller could be used.

5.2.2. Nonmonotonic Response System. In order to demonstrate the effectiveness of the tuning procedures the process

$$G(s) = \frac{(1 - 5s)e^{-20s}}{100s^2 + 12s + 1}$$

was controlled using IMC, which exhibits inverse response and a nonmonotonic profile. For T = 10, ρ = 6.0 and hence the exact inverse controller cannot be used. We also employed a 10th-order truncation model here.

Although stability by reducing M for $P = N$, $\beta_l = 0$ is ensured only for systems with $h_i > 0$, in Figure 18 we show that M = 9, 2, and 1 provide stable controllers, with the same dynamic characteristics as before, for this nonmonotonic response system.

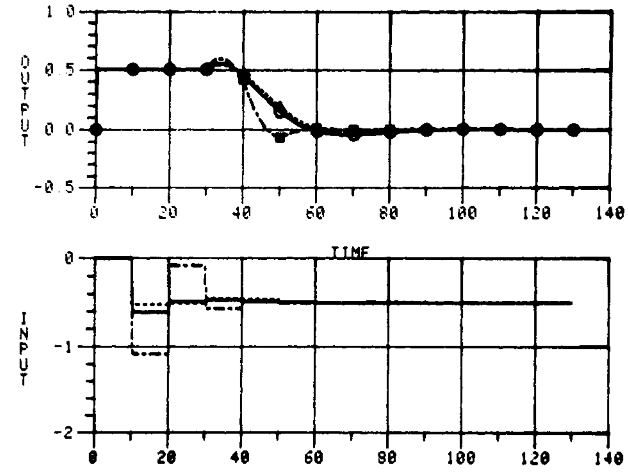

Figure 19. Stable controllers for NMP system of §5.2.2 by introducing input penalties and increasing the horizon: (○), penalty on $m(k)$ with β = 1 and offset correction ($P = M = N = 10$, $\gamma_l = 1$, $\alpha = 0$, ρ = 0.51); (△), penalty on $\Delta m(k)$ with β = 0.5 ($P = M = N = 10$, $\gamma_l = 1$, $\alpha = 0$, ρ = 0.46); (□), P = 11 ($M = N = 10$, $\beta_l = 0$, $\gamma_l = 1$, $\alpha = 0$, ρ = 0.62).

Table I. Normalized Impulse Response of Furnace Outlet Temperature to Fuel Flow (Cutler and Ramaker, 1979)

	time					
	1	2	3	4	5	6
normalized temperature	0.014	0.072	0.128	0.200	0.186	0.136
	time					
	7	8	9	10	11	12
normalized temperature	0.100	0.068	0.045	0.037	0.014	0.0

Table II. Effect of Tuning Procedures on Magnitude of Largest Control-Law Root (ρ) for System of Table I

reduction of M ($P = N = 11$, $\beta_l = 0$)		penalty on $m(k)$ ($P = M = N = 11$)		penalty on $\Delta m(k)$ ($P = M = N = 11$)		increase in horizon P ($M = N = 11$, $\beta_l = 0$)	
M	ρ	β_l	ρ	β_l	ρ	P	ρ
11	3.4	0	3.4	0	3.4	11	3.4
10	1.3	10^{-3}	1.2	10^{-3}	1.2	12	1.3
9	0.95	10^{-2}	0.80	10^{-2}	0.77	13	0.82
8	0.82	10^{-1}	0.73	10^{-1}	0.74	17	0.82
7	0.82	1	0.59	1	0.70	21	0.82
6	0.82	10	0.33	3.5	0.63		
5	0.83			10	0.94		
4	0.81						
3	0.72						
2	0.68						
1	0.51						

The responses when other methods of stabilization are used are shown in Figure 19. As expected, a penalty of β_l = 1 on $m(k)$ produces a stable controller. Moreover, we were also able to stabilize this system by penalizing $\Delta m(k)$ and increasing P.

5.3. Dynamic Matrix Control. In their original article, Cutler and Ramaker (1979) give a furnace outlet temperature response to a step change in fuel flow which exhibits a discrete monotonic response (see Table I). However, using the response coefficients given, an unstable inverse controller would result since for $M = P = N = 11$, ρ = 3.4. In Table II we show how ρ varies with the different tuning parameters.

Observe that the decrease in magnitude of the largest root is nonmonotonous with M and it is not until $M = 8$ that a value significantly less than 1 is obtained. Moreover, by penalizing $\Delta m(k)$ as done in DMC, ρ decreases to a minimum and then approaches 1. Also, observe that an increase in P is able to stabilize this NMP system.

6. Conclusions

IMC consists of three parts: (1) internal model to predict the effect of the manipulated variables on the output; (2) filter to achieve a desired degree of robustness; (3) control algorithm to compute future values of the manipulated variable such that the process follows a desired trajectory closely.

This structure gives IMC a highly intuitive appeal: (1) A model based prediction of the output is attractive in the presence of output constraints. Any constraint violations can be anticipated and corrective action can be taken. (2) Introducing a filter is equivalent to requiring the output to follow a reference trajectory. This interpretation gives the filter constants physical meaning and therefore allows on-line tuning by the operating personnel. (3) The objective of the control law is transparent; input constraints can easily be incorporated without affecting the stability. It was shown that for MP systems the adjustment of only one tuning parameter (the "system order" n) is sufficient to obtain a very good response.

From a theoretical point of view IMC has the following attractive features: (1) The IMC structure allows affecting the quality of the response and the robustness virtually independently through the design of the control block and the filter, respectively. These objectives are hopelessly intertwined in most other design procedures. (2) The stability of the closed loop system is not an issue, only the robustness. (3) The use of an impulse response model is advantageous because a structural model identification is not required and the nonminimal representation adds robustness to the scheme. Furthermore, this model representation is very suitable for a theoretical robustness analysis and thus for the tuning of the filter.

We showed that many conventional control schemes have a structure identical with IMC. The design philosophy, the computation of the inputs, and therefore the robustness are completely different, however.

We have not dealt in any depth with certain other aspects of IMC because no or only preliminary results are available to date. Let us discuss them here and point out some areas for future research along the way.

The control law formulation used in this work does not allow one to account for output or state constraints directly but only through penalty terms in the objective. Inspection of the output predicted by the model allows the operating personnel to anticipate violations and to change the penalty terms on-line to prevent them. An alternate more attractive way is to add inequality constraints to the optimization problem (32) and to solve the resulting quadratic program on-line each time a new value for the input has to be found. A gradient projection method has been developed for this purpose (Mehra et al., 1981), but no results have been reported. If the absolute values of the deviations of the output from the reference are minimized instead of the squares, we obtain a linear program. This might be more attractive for on-line computations.

The currently suggested identification methods for the impulse response coefficients (e.g., Richalet et al., 1978) are unsatisfactory in the presence of low signal/noise ratios. It is probably altogether undesirable to identify an impulse response model because the lack of parsimony is likely to give rise to large errors in the estimates. It might be better to identify a low order parametric model and then use the impulse response from this model in IMC. Furthermore, it would be interesting to employ the IMC structure in an adaptive control law.

We have not discussed how to design the filter (or equivalently how to choose the reference trajectory) to compensate for a specific model-plant mismatch. An excellent summary of the available results is provided by Mehra et al. (1979). It was shown that by making the filter time constant sufficiently large robustness can virtually always be achieved. The use of alternate more complex filters should be explored in order to improve the control quality.

The IMC structure carries over to the multivariable case in a straightforward fashion. The factorization of the transfer matrix into an invertible and a noninvertible part, the construction of approximate inverses, and the filter design are less trivial. Some light has been shed on these questions in a recent report by Zames (1979).

Acknowledgment

Support of this research through NSF Grant ENG-7906353 is gratefully acknowledged. Many of the ideas appearing in this paper were stimulated through discussions with C. Brosilow. K. Åström, P. Belanger, J. J. Haydel, R. K. Mehra, and R. E. Otto provided many interesting comments during the Engineering Foundation Conference "Chemical Process Control II". Finally, we found the criticism by W. H. Ray and B. Ogunnaike here at Wisconsin very useful.

Appendix A

The following lemmas were used in the proofs of stability theorems 2, 3, and 4 of 4.2. Their proofs are found in the book by Jury (1964, p 116).

Lemma 1. Monotonic Conditions. The real polynomial

$$P(x) = \alpha_n x^n + \alpha_{n-1}x^{n-1} + \ldots + \alpha_1 x + \alpha_0 \qquad \text{(A1)}$$

has roots outside the unit circle if $\alpha_0 > \alpha_1 > \ldots > \alpha_n > 0$.

Lemma 2. Dominant Coefficient. Consider the real polynomial (A1). If the kth coefficient is such that

$$|\alpha_k| > |\alpha_n| + |\alpha_{n-1}| + \ldots + |\alpha_1| + |\alpha_0|$$

then $P(x)$ has k roots inside the unit circle, ($(n - k)$ outside).

Appendix B

It can easily be verified that in the limit as $\beta_l \to \infty$, the criterion (41) yields a control law with a root on the unit circle. We show by means of an example that in general an NMP system cannot be stabilized by increasing β_l^2 sufficiently.

Let us assume an impulse response model with $N = 2$

$$y(k+1) = h_1 m(k) + h_2 m(k-1) \qquad (h_1, h_2 \neq 0)$$

and apply criterion (41) for $\gamma_l = 1$, $\beta_l = \beta$ to compute the control law. For $P = M = N = 2$, minimization of (41) yields a control system with the characteristic root

$$z_1(\beta^2) = \frac{(h_1^2 + \beta^2)(\beta^2 - h_1 h_2)}{(h_1^2 + h_2^2 + 2\beta^2)(h_1^2 + \beta^2) - (h_1 h_2 - \beta^2)^2} \qquad \text{(B1)}$$

As a check, $z_1(0) = -h_2/h_1$, the open-loop zero, and $z_1(\infty) = 1$, as expected. If $|h_2/h_1| \geq 1$ the process is NMP and thus the process inverse controller ($\beta = 0$) is unstable. We would like to see how that unstable root reaches the limiting value of unity as β^2 is increased.

Hence, we are interested in calculating $dz_1(\beta^2)/d\beta^2$ and observe its sign as $\beta^2 \to \infty$. We can easily perform the differentiation and obtain

$$\frac{dz_1(\beta^2)}{d\beta^2} = (h_1 + h_2)\frac{\{(2h_1 + h_2)\beta^4 + 2h_1^3\beta^2 + h_1^3(h_1^2 + h_1h_2 + h_2^2)\}}{\{\beta^4 + (3h_1^2 + 2h_1h_2 + h_2^2)\beta^2 + h_1^4\}^2}$$

It is straightforward to verify that for any MP system ($|h_2/h_1| < 1$) this derivative is positive for any finite β^2 and hence, the controller cannot be made unstable by increasing β (as expected). Also, NMP systems with monotonic step responses can be stabilized using β. However, one can easily concoct examples where an unstable root approaches unity from above, e.g. for $h_1 = -0.4$, $h_2 = 0.7$, $z_1(0) = {}^7/_4$, and

$$\frac{dz_1(\beta^2)}{d\beta^2} = (0.3)\frac{\{(-0.1)\beta^4 + (-0.128)\beta^2 + (-0.064)(0.37)\}}{\{\beta^4 + 0.41\beta^2 + 0.0256\}} < 0$$

for all finite values of β^2.

Literature Cited

Alevisakis, G.; Seborg, D. E. *Int. J. Control* **1973**, *3*, 541.
Arkun, Y.; Stephanopoulos, G. *AIChE J.* **1980**, *26*, 975.
Åström, K. J. "Introduction to Stochastic Control Theory"; Academic Press: New York, 1970; p 275.
Åström, K. J. *Automatica* **1980**, *16*, 313.
Bohl, A. H.; McAvoy, T. J. *Ind. Eng. Chem. Process Des. Dev.* **1976**, *15*, 24.
Brosilow, C. B. "The Structure and Design of Smith Predictors from the Viewpoint of Inferential Control"; Joint Automatic Control Conference Proceedings, Denver, CO, 1979.
Cutler, C. R.; Ramaker, B. L. "Dynamic Matrix Control—A Computer Control Algorithm"; AIChE 86th National Meeting, April, 1979; also in Joint Automatic Control Conference Proceedings, San Francisco, CA, 1980.
Doyle, J. C. *IEEE Trans. Autom. Control* **1978**, *AC-23*, 756.
Faddeeva, V. N. "Computational Methods of Linear Algebra"; Dover Publications, Inc.: New York, 1959; p 202.
Foss, A. S. *AIChE J.* **1973**, *19*, 209.
Fuller, A. T. *Int. J. Control* **1968**, *8*, 145.
Horowitz, I. M.; Shaked, U. *IEEE Trans. Autom. Control* **1975**, *AC-20*, 84.
Jury, E. I. "Theory and Application of the Z-transform Method"; R. E. Krieger, Ed.; Huntington: New York, 1964; p 116.
Kestenbaum, A.; Shinnar, R.; Thau, F. E. *Ind. Eng. Chem. Process Des. Dev.* **1976**, *15*, 2.
Koppel, L. B. "Introduction to Control Theory with Applications to Process Control"; Prentice Hall: Englewood Cliffs, NJ, 1968; Chapter 8.
Kwakernaak, H.; Sivan, R. "Linear Optimal Control Systems"; Wiley-Interscience: New York, 1972; Chapter 6.
Lecrique, M.; Rault, A.; Tessier, M.; Testud, J. L. "Multivariable Regulation of a Thermal Power Plant Steam Generator"; Int. Fed. Autom. Control World Congress, Helsinki, 1978.
Lee, W.; Weekman, V. W., Jr. *AIChE J.* **1976**, *22*, 27.
Mehra, R. K.; Eterno, J. S. "Model Algorithmic Control for Electric Power Plants"; IEEE Conf. on Dec. and Control, Albuquerque, NM, 1980.
Mehra, R. K.; Kessel, W. C.; Rault, A.; Richalet, J.; Papon, J. "Alternatives for Linear Multivariable Control"; Sain, et al. Ed.; NEC Inc.: Chicago, 1978; p 317.
Mehra, R. K.; Rouhani, R. "Theoretical Considerations on Model Algorithmic Control for Non-minimum Phase Systems"; Joint Automatic Control Conference Proceedings, San Francisco, CA, 1980.
Mehra, R. K.; Rouhani, R.; Eterno, J.; Richalet, J.; Rault, A. "Model Algorithmic Control: Review and Recent Developments"; Eng. Foundation Conf. on Chemical Process Control II, Sea Island, GA, 1981.
Mehra, R. K.; Rouhani, R.; Praly, L. "New Theoretical Developments in Multivariable Predictive Algorithmic Control"; Joint Automatic Control Conference Proceedings, San Francisco, CA, 1980.
Mehra, R. K.; Rouhani, R.; Rault, A.; Reid, J. G. "Model Algorithmic Control: Theoretical Results on Robustness"; Joint Automatic Control Conference Proceedings, Denver, CO, 1979.
Ogunnaike, B. A.; Ray, W. H. *AIChE J.* **1979**, *25*, 1043.
Palmor, Z. J.; Shinnar, R. *Ind. Eng. Chem. Process Des. Dev.* **1979**, *18*, 8.
Popov, V. M. *Autom. Remote Control* **1963**, *24*, 1.
Prett, D. M.; Gillette, R. D. "Optimization and Constrained Multivariable Control of a Catalytic Cracking Unit"; AIChE 86th National Meeting, April, 1979; also in Joint Automatic Control Conference Proceedings, San Francisco, CA, 1980.
Reid, J. G.; Chaffin, D. E.; Silverthron, J. T. "Output Predictive Algorithmic Control: Prediction Tracking with Application to Terrain Following", Joint Automatic Control Conference Proceedings, San Francisco, CA, 1980.
Reid, J. G.; Mehra, R. K.; Kirkwood, E. "Robustness Properties of Output Predictive Dead-Beat Control: SISO Case", Proc. IEEE Conf. on Dec. and Control, Fort Lauderdale, FL, 1979, p 307.
Richalet, J. A.; Rault, A.; Testud, J. D.; Papon, J. *Automatica* **1978**, *14*, 413.
Safonov, M. G.; Athans, M. *IEEE Trans. Autom. Control* **1978**, *AC-23*, 717.
Smith, O. J. M. *Chem. Eng. Prog.* **1957**, *53*, 217.
Zames, G. "Feedback and Optimal Sensitivity: Model Reference Transformations, Multiplicative Seminorms, and Approximate Inverses"; Report, Electrical Engineering Department, McGill University, Montreal, 1979.

Received for review March 30, 1981
Revised manuscript received September 30, 1981
Accepted November 30, 1981