$$e_r = 1.015 w L_a A k_H v_a \times 10^{-8} \tag{20}$$

Arnold replaces the theoretically required interpole flux shape by a trapezoid, for which he then derives certain constants of shape (Fig. 3).

$$B_{IP_{av}} = 2.54 A k_H \nu \text{ lines per inch}^2 \tag{21}$$

$$B' = 2.54 A \frac{8\pi}{10} \frac{t}{\eta Z}[(\lambda_{slot} + \lambda_{TT_1}) + (\nu - 1)(\lambda_{slot} + \lambda_{TT_2})] \tag{22}$$

$$B'' = 2.54 A \frac{8\pi}{10}(\nu - 1)\lambda_q \tag{23}$$

$$B''' = 2.54 A \frac{8\pi}{10}\nu\lambda e \tag{24}$$

$$B_{IP_{max}} = B' + B'' + B''' \tag{25}$$

$$B_{IP_{av}} = \eta B' + B'' + B''' \tag{26}$$

To calculate the shape constant η, the effective brush width b' is required

$$b' = \frac{D_a}{D_c}\left[(b-j)+\left(1-\frac{a}{p}\right)\tau_c\right] \tag{27}$$

Now, the proper condition must be chosen.

a. $|u\delta|=0 \quad t>b' \qquad \eta=\dfrac{t}{Z} \tag{28}$

b. $|u\delta|=0 \quad t<b' \qquad \eta=\dfrac{b'}{Z} \tag{29}$

c. $|u\delta|\neq0 \quad t>(b'+\delta t) \qquad \eta=\dfrac{t}{Z} \tag{30}$

$$(b'+\delta t)>t>(b'-\delta t) \qquad \eta=\frac{2tb'}{Z(t+b'-\delta t)} \tag{31}$$

$$(b'-\delta t)>t \qquad \eta=\frac{b'}{Z} \tag{32}$$

Required interpole MMF

$$MMF_{IP} = MMF_a + \frac{B_{IP_{max}}d_{IP}k_{cIP}}{\frac{4\pi}{10}2.54} + MMF_{IP}'' - MMF_c \tag{33}$$

with MMF_{IP} an estimated addendum for ampere turns required within the normally unsaturated interpole circuit excluding the interpole air gap.

References

1. COMMUTATION OF LARGE DC MACHINES (book), L. Dreyfus. Springer-Verlag, Berlin-Charlottenburg, Germany, 1929.

2. THE DC MACHINE (book), Arnold-la-Cour. Springer-Verlag, Berlin-Charlottenburg, Germany, 1923.

3. DESIGN CONSIDERATIONS FOR D-C AIRCRAFT GENERATORS, J. D. Miner. AIEE Transactions, vol. 63, 1944, pp. 1234–40; discussion, pp. 1459–60.

4. RESISTANCE COMMUTATION, Frank D. Olney. AIEE Transactions, vol. 69, pt. II, 1950, pp. 1207–18.

5. ELECTRICAL MACHINERY (book), Richter. Springer-Verlag, Berlin-Charlottenburg, Germany, 1924.

————————◆————————

No Discussion

Optimization of Nonlinear Control Systems by Means of Nonlinear Feedbacks

R. S. NEISWANDER
ASSOCIATE MEMBER AIEE

R. H. MacNEAL
NONMEMBER AIEE

Synopsis: The first portion of this paper is devoted to the logical selection of a category of nonlinear systems—systems having: 1. controllers with any number of time lags or leads, and usually operating in their saturated region (velocity saturation, force saturation, etc.), and 2. loads consisting of combinations of nonlinear masses, nonlinear dampers, and nonlinear springs. (Nonlinearities which involve explicit functions of time, such as mass and inertia of a controlled rocket motor missile, introduce special considerations not included here.) In the second part of the paper, characteristics of an intentionally nonlinear element added to the feedback are considered. In the final part, the nonlinear element is shaped to provide optimum response in the saturated region of operation.

THE USE of intentionally nonlinear elements to improve automatic control performance is not a new idea in servomechanism history; it has been recognized for a number of years that the overshooting tendency of saturated control systems could be reduced by increasing damping at saturated velocities. However, the problem of shaping the nonlinear element for best over-all performance has been considered only recently, and in many respects this can be regarded as a completely new branch of the servomechanism field. The work of Tustin[1,2] and others in England has led Porter[3] to predict that servo system analysis of the future may be directed principally toward introducing nonlinear elements as means of optimizing control systems, in contrast to the present emphasis upon linearization by mechanical and electrical design.

The growing interest both here and abroad in nonlinear control systems is probably mostly the result of necessity. McDonald[4] in his survey of the state of the art, points out that the improvement of linear systems by engineering design, reducing time delays and energy storages, and improving acceleration characteristics of servo motors is rapidly reaching a point of diminishing return. As a next step, nonlinear techniques come into being.

There are at present at least three methods of treating nonlinear servomechanisms and, as is common to all nonlinear techniques, each method is applicable to a certain restricted category of control systems. Each of these methods is an outgrowth of an on-off servomechanism analysis technique, although the original techniques did not include the concept of performance optimization. Hazen[5] in his classic (1934) paper proposed analyzing, by operational techniques, on-off controls which were essentially linear except for discontinuities at switching instants. MacColl[6] in 1945, and Weiss[7] in 1946 proposed the use of phase-plane analysis to determine responses of certain second-order on-off servomechanisms. More recently, Kochenburger[8] in 1950 introduced a frequency response method applicable to pseudo-linear on-off controls, that is, although actually nonlinear, several linearizing assumptions may be made regarding certain control systems.

By the introduction of an intentionally nonlinear element into the control system, each of the on-off servo techniques has been extended to provide a means of improving control system performance. The sectionally linear technique has been adapted by Schwartz[9] to improve a linear

servomechanism by means of a superimposed discontinuous control. Sherrard[10] has suggested a means, based upon the more general frequency analysis technique, of determining a nonlinear filter which improves the response of a saturated linear servomechanism. The phase-plane technique has been used by Hopkin[11] for the optimization of certain second-order control systems in which the force applied to the load is assumed to be saturated. In Hopkin's method the resultant error signal includes an intentionally nonlinear function of the static error and a linear function of the output velocity.

The intent of the present paper is to outline a method, in many respects an extension of Hopkin's proposal, which can be applied to complex control systems. The method accounts for various kinds of saturations (velocity, force-rate, force, etc.) occuring in control systems having nonlinear loads, and a number of time lags and leads. Although optimization is accomplished by nonlinearizing rate feedback, it should be noted that most of the features of the method are applicable to optimization by nonlinearizing static error signal. For other than simplified control systems, automatic computers are substituted for graphical phase-plane constructions.

The method to be developed is applicable to the saturated region of operation and is therefore not directly concerned with operation near the control system null. Excluding on-off systems, a large number of control systems are reasonably linear in their null regions and can be designed for optimum null region performance by means of well-known linear techniques. If the resulting parameters (loop gain, and correction networks) are assumed to be invariable and necessary elements of the basic control, the present method can be applied to the resultant system. In other words, the goal of null tightness does not conflict with the goal of optimum saturated performance; both can be attained in the same system.

In considering the class of control systems termed nonlinear, one should keep in mind that they can only be defined negatively: all which are not linear. They constitute an infinitely large and diverse class which defies generalization or general analysis. Consequently, any method of treating nonlinear systems must necessarily be restricted. The restrictions of the method to be presented are not precisely known, for this would require the nearly impossible task of investigating a great number of nonlinear control systems; therefore the
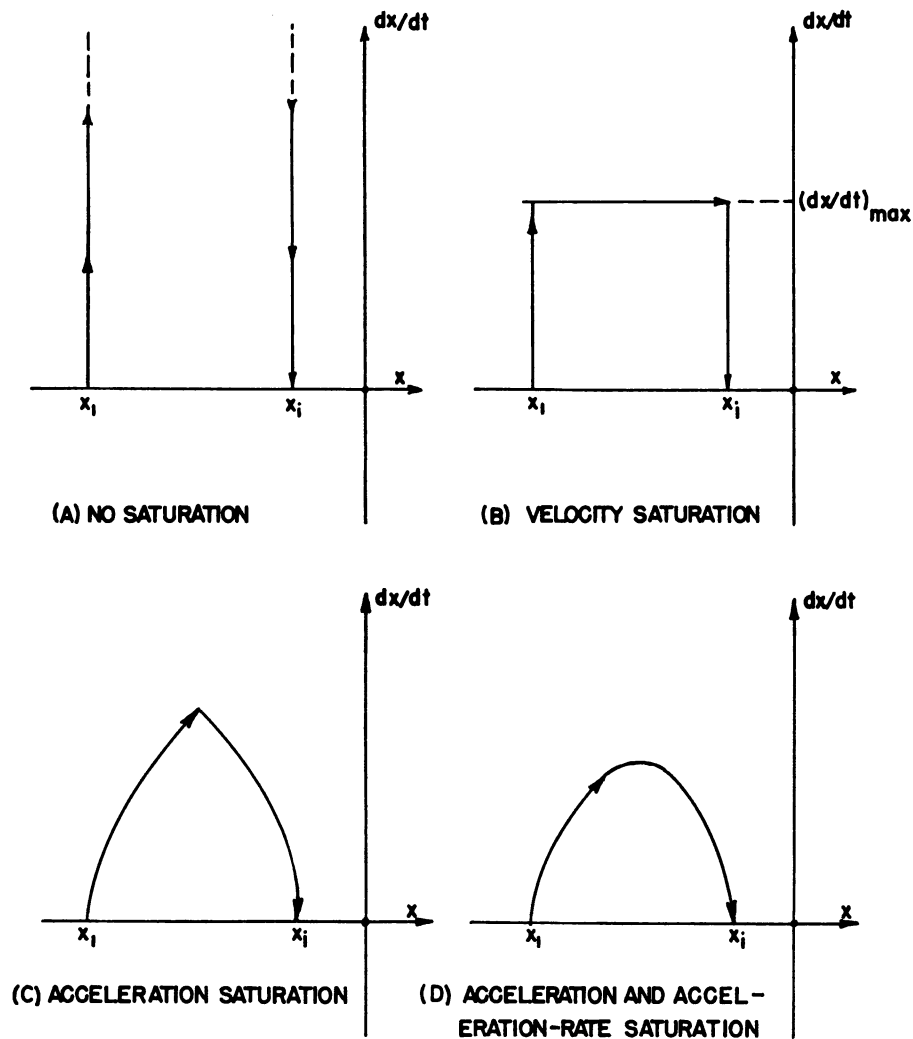

Fig. 1. Optimum saturated responses

limitations noted throughout this paper should not be considered to be exact.

The Optimum System

Contemporary control system analysis evolved from feedback amplifier design and electrical circuit theory, and it deviates with reluctance from the original concept of the completely linear system. In fact, linear control systems with their neat, analytic solutions are so easily treated that there is a tencency to regard the linear system as the optimum system. Nonlinearities are considered undesirable and are neglected, eradicated, or occasionally admitted by approximate extrapolation from linear theory.

On the other hand, there is an interesting and quite different interpretation of system optimization, which perhaps may best be described by a series of examples. Let it be assumed that the objective of a control system is to move its output, x, from an initial point, x_1, to a final point $x_i \pm \delta$, δ being the tolerated error, in the

shortest possible time. It will be advantageous to present the response of such a system by means of a displacement-velocity plot, such as those illustrated in Fig. 1. For a second-order control system, this plot is the phase-plane representation, and for a higher order system this is a phase-plane cross-section of the corresponding multidimensional phase space. However, in either case, the position-velocity plot presents sufficient information, and will be subsequently referred to here as the phase plane.

If the system has almost unlimited available force and perfect control intelligence, its quickest route to the final point is by means of a nearly infinite acceleration to some midpoint and then a nearly infinite deceleration to the final point. The corresponding response trajectory, Fig. 1(A), rises vertically from the initial point, x_1, to an almost infinite velocity, and then descends vertically down to x_i. The relation between the phase-plane trajectory and the response time is apparent
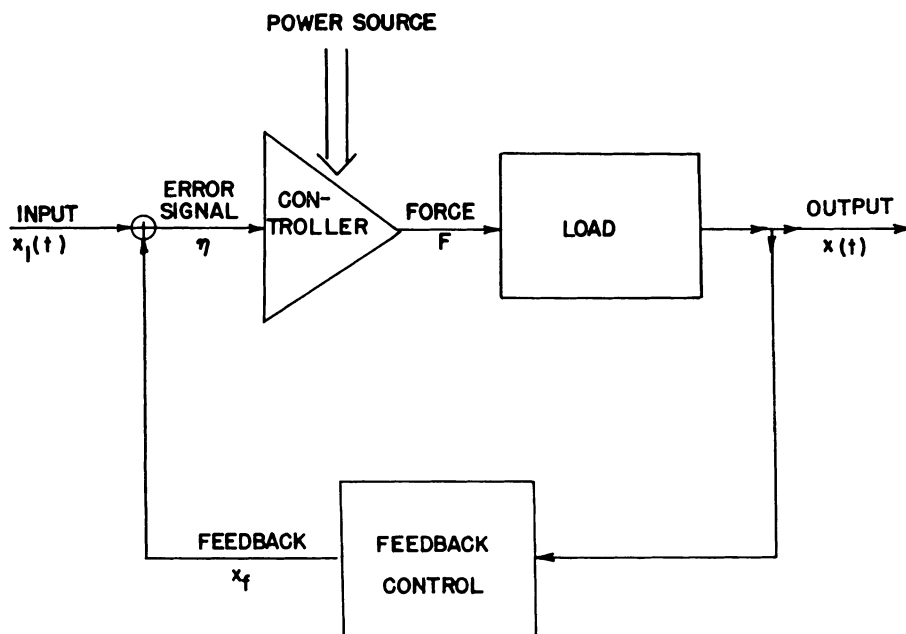
Fig. 2. The central system

$$\text{Response time} = T = \int_0^T dt = \int_{x_1}^{x_i} \frac{dx}{(dx/dt)}$$

Thus, the response time is the area under $1/(dx/dt)$, the reciprocal of the phase-plane trajectory. For the ideal response of Fig. 1(A), the travel time is practically zero.

Using this phase-plane representation, the effects of physical limitations upon the system's responses can readily be seen. Velocity saturation prevents the trajectories from rising above a velocity limit, $(dx/dt)_{max}$. Thus, a control having no limitations except this velocity limit best performs the task of moving from x_1 to x_i by means of an infinite acceleration up to the bounding velocity, and then a constant velocity up to the end point where an infinite deceleration stops the motion. The resultant trajcetory, Fig. 1(B), now has a finite response time.

In all practical systems the output acceleration must be finite, limited by the inertia of the load and the available force. For example, a system having a pure inertia load and limited available force but no other restrictions has, as its best response, maximum acceleration (maximum available force) to the midpoint, and maximum deceleration to the end point. This produces the trajectory of Fig. 1(C).

Furthermore, if the control force is not able instantaneously to reverse itself, the slope of the trajectory cannot abruptly change, and the resultant trajectory has a rounded apex, Fig. 1(D).

Each of these four examples utilizes its absolute maximum physical capabili-

ties in performing the required operation. With these assumed capabilities, no means exist for improving the responses. An absolute limit to a given system's performance has been established, providing an evaluation criterion for the system's actual performance.

Implicit in all of the examples discussed is the assumption of unlimited control intelligence, which provides the perfect saturated approach to the final point. In actual fact, real systems here will be allowed to become unsaturated and possibly linearly controlled in the null region. This provision results in a less than perfect response.

As a basis for the optimization procedure, the foregoing can be summarized: An absolute optimum control system is one which has physical limitations only, having sufficient control intelligence to perform all requested operations at saturated levels.

For all of the foregoing examples, the performance requirement to move from position x_1 to position $x_i \pm \delta$ within the shortest possible time, was employed, and will continue to be used throughout the remainder of this paper. Other performance demands, such as following a sinusoidal input, $A \sin \omega t$ within tolerated error δ, maximizing ω; following a constant velocity input v within tolerated error δ, maximizing the velocity; or minimizing output error for large variations of the load, will in general require different optimizing techniques. Since superposition of solutions is not permissible in nonlinear analysis, correlation between optimization for any two require-

ments, such as the first two mentioned, is difficult.

The Method

The task is to devise an intentional feedback which sponsors the optimum response defined in the previous section. Fig. 2 represents the typical control system to be optimized. The controller, which may have various linear and nonlinear elements, imposes a force F on the load, which may be a combination of nonlinear masses, nonlinear dampers, and nonlinear springs (excluding explicit functions of time). Saturated system optimization assumes complete saturation of the controller and thus does not, in general, provide useful information for the final portions of the response trajectories. A technique for determining the limit, the boundary of the unsaturated region surrounding the end point, will be presented later. Within the completely saturated region, the error signal must be large, and has no influence on the output except to initiate reversal when the error reverses polarity.

To discover just what intelligence is required of the optimum feedback element, it is worth while to re-examine the optimum response trajectories of the system. Consider, for example, a system which has a mass plus damping load and a controller which is force-saturated but which can instantaneously reverse the direction of the force. The perfect responses of this system are shown in Fig. 3(A), for final position x_i and various initial values $x_1, x_2 \ldots$ (A convenient rule to remember is that all trajectories in the upper half-plane must progress to the right since $\dot{x}$ is positive; and all in the lower half-plane to the left.) It is convenient to shift the x co-ordinates so that the final position, x_i, always lies at the origin, that is, $x' = x - x_i$. Along the early portions of the trajectories, from x_1' to s_1, x_2' to s_2, etc., the system is saturated, accelerating positively. At $s_1, s_2 \ldots$ the control intelligence reverses the signal, and the force is reversed to become oppositely saturated. Near the origin (the null) the method is not valid and trajectories must be determined by linear analysis or other means. Therefore, for optimum saturated response the only duty of the control intelligence is to select the initiation of reversal, the optimum switching line $s_1 s_2 s_3 \ldots$ Any saturated trajectory which is switched, upon crossing this line, is properly directed towards the origin.

The load mass and damping elements

of this example may be nonlinear without changing the character of the perfect response trajectories. If either of these elements is a function of x, or if a spring is added to the load, perfect responses such as those of Fig. 3(A) will be different for different final positions, and a series of phase planes is required, each representing a specific final position and a complete range of initial positions.

If the load is not an explicit function of x, that is, only a function of time derivatives of x, the various phase-plane plots for different final points will be identical. The optimum switching line can either be considered as a function of x', which is in reality the static error signal, or as a function of output velocity. Using static error signal as the independent variable, Hopkin[11] devised a phase-plane method of optimizing second-order control systems by intentionally nonlinearizing the static error. The method presented here utilizes output velocity as the independent variable, with consequent optimization by nonlinearizing the rate feedback. These two viewpoints are very closely related, and many of the features of the nonlinear rate optimization described here are applicable to nonlinear static error optimization.

It would appear that this optimum saturated control is in many respects similar to that of an on-off servo; and, indeed, in some respects the relay appears to be the ideal controller. However, as mentioned previously in discussing required intelligence, is it assumed that the null accuracy of response is improved if the system is approximately linear.

Admitting a maximum rate of change of controller force dF/dt_{max} the family of optimum responses becomes those of Fig. 3(B). Note that the optimum switching line $s_1s_2s_3$. . .has been advanced to allow for reversal time, the intervals s_1t_1,s_2t_2. . .being the portions of the trajectories during reversal. Again, when any saturated trajectory has its signal switched, upon crossing the optimum switching line, reversal is initiated in such a way that the trajectory is properly deflected towards the origin.

If, in addition, the system has an absolute velocity barrier, the optimum switching line may take the odd form shown in Fig. 3(C).

In brief, the optimum switching line can be expressed as a function of dx/dt for certain systems (loads independent of x), and in general as a function of dx/dt and the input, or null, value, x_i. In other words: The sufficient control intelligence for the optimum control system (always saturated) is an optimum switch-
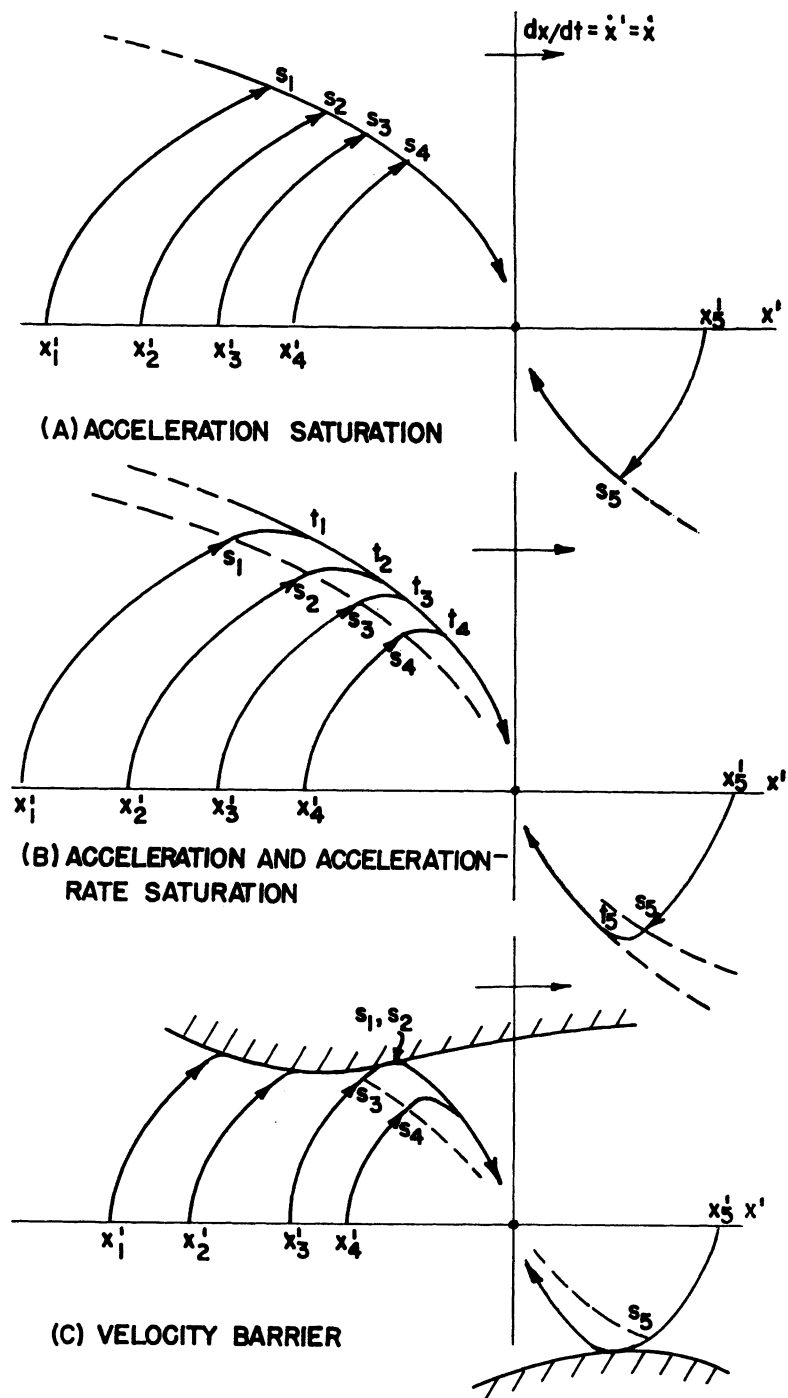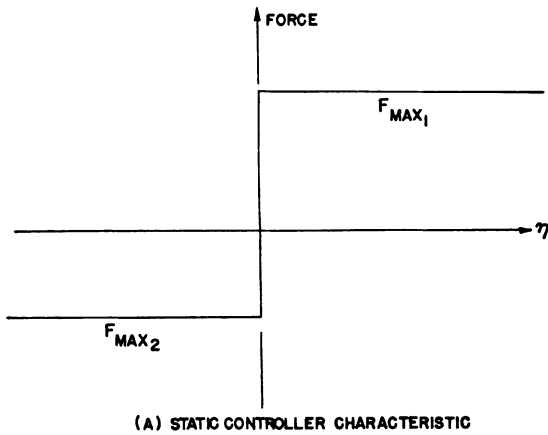
ing line $S\{dx/dt, x_i\}$ determined by the perfect response trajectories of the system. Throughout this paper, the brackets $\{\}$ will mean a function of. . . .

Since the optimum switching line is the entire intelligence required by an optimum control, it must be provided by the feedback. But just what is the nature of influence of the feedback? First, the feedback must transfer $-x$ to the linear summing element, for it is implicitly assumed that the output should follow the input. Second, it must have some sort of optimizing function $A\{\}$ to provide
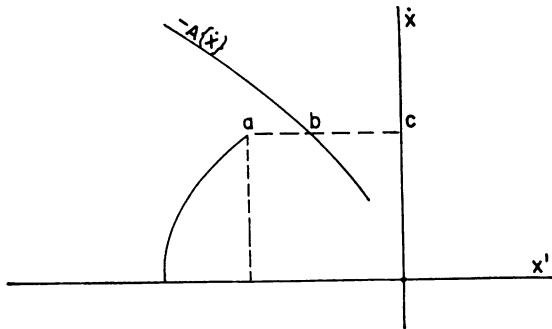
the optimum switching line. The resultant feedback signal to the summing element is therefore $x_f = -x - A\{\}$, and the corresponding error signal is $\eta = x_i - x - A\{\}$. Shifting the x axis to maintain the end point at the origin $(x' = x - x_i)$

$$\eta = -x' - A\{\}$$ (1)

As is usual with nonlinear phenomena, simple examples perhaps best demonstrate the important features of equation 1. Consider first a hypothetical system having a static controller characteristic as shown in Fig. 4(A). The controller can
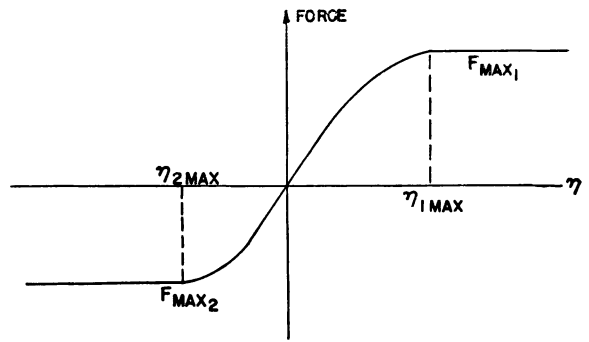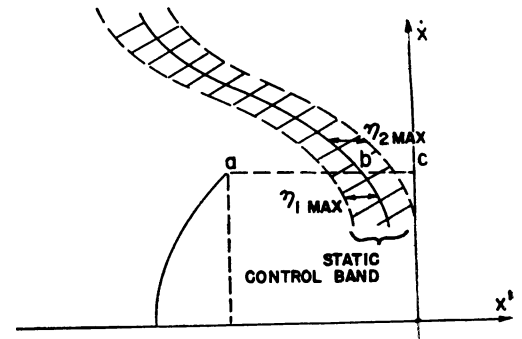

Fig. 3. Switching lines for optimum responses

Fig. 4 (left). Perfect relay controller

(A) STATIC CONTROLLER CHARACTERISTIC

(B) NONLINEAR RATE FEEDBACK

(A) STATIC CONTROLLER CHARACTERISTIC

Fig. 5 (right). Proportional controller

(B) NONLINEAR RATE FEEDBACK

be thought of either as an amplifier with infinite gain at null, or as a zero time-delay relay. It is reasonable to suppose that the intentional feedback $A\{\ \}$ supplying the intelligence for switching is a function of the same variables as the optimum switching line, $S\{\dot{x}, x_i\}$ and therefore, for a given end point can be represented on the phase plane as a line $A\{\dot{x}\}_{xi\,=\,\mathrm{const}}$. Fig. 4(B) shows a saturated accelerating trajectory approaching the line $-A[\dot{x}]$. At point a, $ac = x_a'$, $bc = -A[\dot{x}]_a$, and $ab = x_a' + A[\dot{x}]_a$ or

$$ba = -x_a' - A\{\dot{x}\}_a = \eta \tag{2}$$

For the assumed controller, the force is switched when the error signal $\eta = 0$. In other words $-A\{\dot{x}\}$ is the signaled switching line. When the signaled switching line is coincident with the optimum switching line, the feedback sponsors perfect control, and the system optimization has been accomplished. In the selected example, this is extremely simple

$$S\{\dot{x}\}_{xi\,=\,\mathrm{const}} = -A\{\dot{x}\}_{xi\,=\,\mathrm{const}}$$

or

$$S\{\dot{x}, x_i\} = -A\{\dot{x}, x_i\}$$

Now consider the more practical controller static characteristic shown in Fig. 5(A), which has force limits of F_{max_1} and F_{max_2} but which is unsaturated and nearly linear between these limits. From equation 2, the length ba of Fig. 5(B) is

the signal η. Upon entering the band between $-A\{\dot{x}\}$ displaced η_{max_1} to the left and displaced η_{max_2} to the right, all trajectories become statically unsaturated. Upon leaving the band, the trajectories have opposite static saturation (F_{max_2}). Reversal is initiated at the instant the accelerating trajectory enters the band, and therefore the signaled switching line is $-A\{\dot{x}\}_{xi\,=\,\mathrm{const}} - \eta_{\mathrm{max}_1}$.

Within the band of Fig. 5(B), the controller is not statically saturated and

consequently unsaturated control of the system will result unless other saturations occur. Accordingly, this zone is labeled the static control band. It presents an important problem in connection with the method of optimization being described. If no other saturations occur, the system does not perform saturated reversals, and the present method is not strictly applicable. The saturated reversal and the unsaturated reversal must be treated differently.
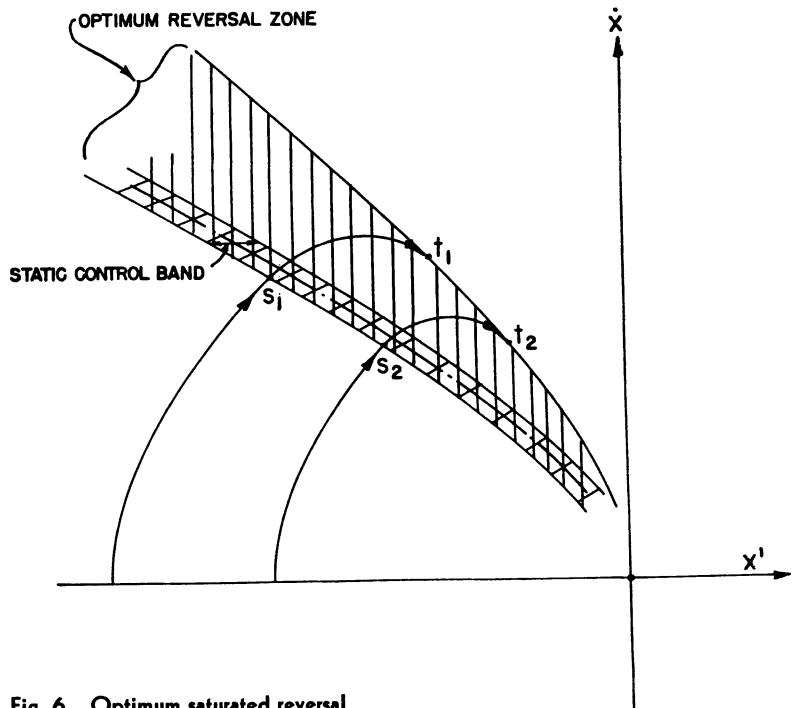

Fig. 6. Optimum saturated reversal

1. *Saturated reversal.* Nonstatic saturations produce optimum trajectories such as those of Figs. 3(B) and 3(C); as an example, the force-rate saturated trajectories of 3(B) are reproduced in Fig. 6. When a controller is force-rate saturated, its output force is changing at the maximum rate permitted by the physical limitations of the controller. The region between $s_1s_2s_3$...and $t_1t_2t_3$..., where saturated reversal is effected, is termed here the optimum reversal zone. Trajectories reversed within this zone have maximum curvature, for it is physically impossible to reverse any trajectory more quickly than its force-rate saturation permits.

Reversal within a static control band narrower than the optimum reversal zone would imply a greater than possible curvature of the reversing trajectories, therefore in this case saturated reversal is assured. The controlled trajectories will thus be identical with the optimum trajectories, and the signaled switching line will be identical with the optimum switching line

$$S\{\dot{x}, x_i\} = -A\{\dot{x}, x_i\} - \eta_{max_1} \qquad (3)$$

Graphically, this amounts to matching the leading edges of the static control band and the optimum reversal band.

2. *Unsaturated reversal.* By similar reasoning, one would expect that, when the static control band is wider than the optimum reversal zone, the trajectories would be unsaturated, at least for the latter portion of the reversal. This prohibits unrestricted use of the present method. If, as in the case of many simple systems, once captured, the trajectories forever remain within the static control

band, optimization is accomplished by matching the exit boundaries of the two zones, Fig. 7

$$T\{\dot{x}, x_i\} = -A\{\dot{x}, x_i\} + \eta_{max_2} \qquad (4)$$

However, one must be very careful in applying the method to this region, for it may not produce accurate response information. Effects such as secondary oscillations, which might not only be undesirable but might swing the trajectory outside the static control band, are not revealed except by special techniques, such as linear perturbation analysis.

In all of these examples, time lags and leads in the nonlinear feedback and, for that matter, in the portion of the controller ahead of the point where static saturation occurs, were not included. Now consider the more general problem of control systems with several time lags or leads. Necessarily, here, the operating region optimized will be confined to that of total saturation—saturated accelerating trajectories, saturated reversals, and saturated decelerations to the null region. An optimum switching line still exists, and the feedback, as before, sponsors a signaled switching line. When the arbitrary feedback is shaped to make the two switching lines coincident, perfect saturated control is attained.

Optimization Procedure for Complex Control Systems

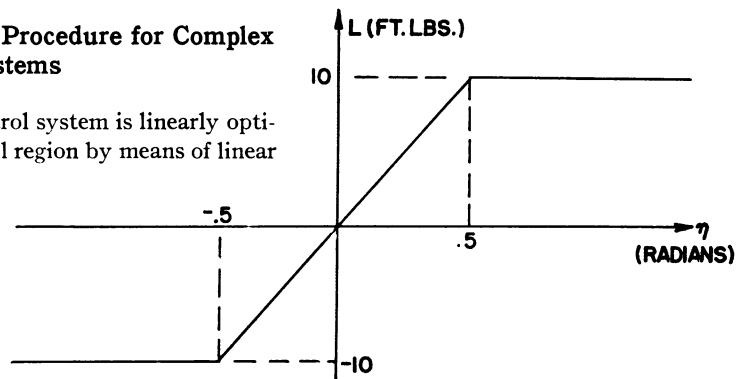1. The control system is linearly optimized in its null region by means of linear

feedback elements and error networks in the controller. The resultant system may have both intentional and unavoidable time lags and leads in the feedback and in the controller and, as mentioned before, may have a load consisting of a combination of nonlinear masses, dampers, and springs (not explicit functions of time).

2. The optimum saturated responses of the system are computed. These are initial condition problems concerned only with the controller and the load, which permit straightforward solutions by analogue or digital computers. All time lags and leads occurring in the controller prior to the points where saturations occur should be excluded from this analysis. The complexity of the load (number of degrees of freedom, backlash couplings, etc.) is limited only by the computing facilities available.
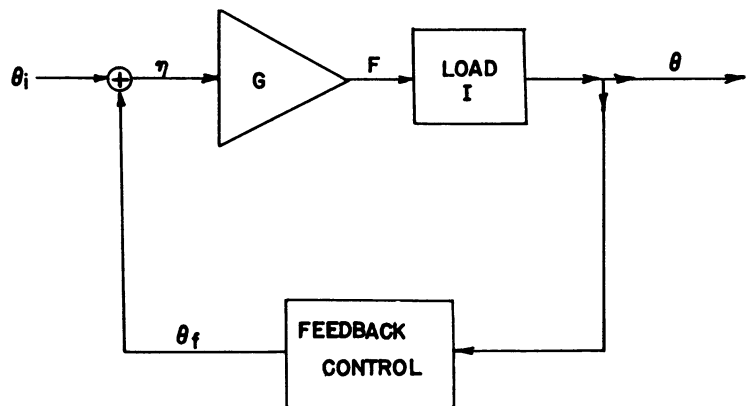
As a procedural example, assume that the given system is at initial point, x_1. It is usually permissible to allow the saturated accelerating force to be applied immediately at the initiation of the input signal, neglecting the initial hook in the trajectory resulting from force build-up time or other nonstatic saturations, for once static force saturation is at-

(A) STATIC CONTROLLER CHARACTERISTIC

(B) BLOCK DIAGRAM

Fig. 8. Simplified control system

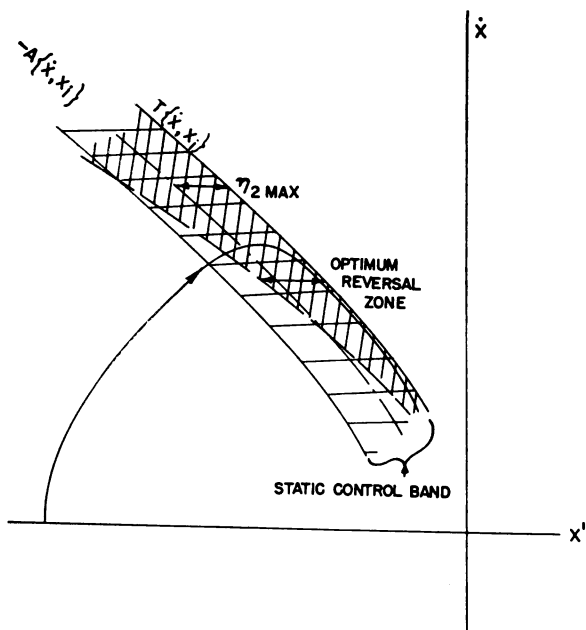
Fig. 7. Optimum unsaturated reversal

tained, the problem is unconcerned with initial conditions. During the acceleration phase, the solution is that of a known force acting upon a known load. At an arbitrary time, τ_1, reversal is initiated by reversing the signal, and a new initial value problem is created with known initial conditions, forces, and load. This solution is continued until the output velocity is zero, producing final position x_{t1}. Using various switch initiation times and various initial points x_1, x_2, etc. the complete set of saturated responses can be obtained. As an accuracy refinement, the final portions of the trajectories, where, in the actual system, linear operation occurs, can be computed by linear analysis. Fortunately, all final trajectories for any one final position are identical.

From the optimum responses, the optimum switching line, $S\{\dot{x}, x_i\}$ is immediately obtained.

3. Assuming no time lags (or leads) in the feedback or in the controller prior to static saturation point, the first approximation of the feedback function is simply

$$A_0\{\dot{x}, x_i\} = -S\{\dot{x}, x_i\} +\eta_{max1}$$

The value of η_{max1} results directly from the controller static force limit and the null slope determined in step 1.

4. In general, the nonlinear feedback function $A\{\dot{x}, x_i\}$ must be modified to account for multiple time lags and leads and other feedback and controller charac-

teristics neglected in the foregoing analysis. The basic procedure proposed is this: The desired responses, x versus time are known from step 2. If the control system loop is broken just prior to the feedback and the perfect responses (along with proper control system inputs) applied as forcing functions, then the actual output responses of the system should coincide with the applied perfect responses. If outputs do not correspond, the nonlinear feedback $A\{ \}$ must be adjusted to produce proper responses.

A suggested procedure for accomplishing this is as follows: The first estimate of the nonlinear feedback has already been found. This function can be tried in the form of $kA_0\{\dot{x}, x_{in}\}$ for some specific perfect response applied as a forcing function to the feedback, and the factor k adjusted until proper switching is obtained. This provides a point on the improved $A\{\dot{x}, x_{in}\}$ curve. Repeating with the balance of the perfect responses, the second approximation $A\{\dot{x}, x_i\}$ results.

5. It is clear that this method is applicable, strictly only to the operating region which is continuously saturated, and it is therefore very important to determine the limits of this region. Fortunately, the unsaturated boundaries are the immediate results of step 4. A limiting optimum response exists which when applied as a forcing function produces a very slightly unsaturated reversal of the load; that is, the resultant trajectory deviates slightly from the optimum trajec-

tory (by displaying less curvature). In an exact sense, this trajectory is the limit of usefulness of the method.

6. The final step in the design of the control system is the reconciliation of the linear and saturated results. The control system has been optimized for its linear null region operation and for its saturated operation. Furthermore, the limits of the saturated operation are known. An unsaturated but very nonlinear zone of operation may exist between the two solved regions. When this gap is relatively small and the system's nonlinear functions are well-behaved, interpolation of $A\{\dot{x}, x_i\}$ across the zone is fairly simple. If the nonlinear, unsaturated zone is large compared to the other two operating regions, optimization can in many instances be accomplished (see previous discussion of unsaturated reversal) but requires considerable caution. In complex systems the innocent-appearing static control band of the position-velocity plot is in reality an n-dimensional space tube with the associated freedoms for instabilities.

A Simple Example

Optimization of a simplified control system will be carried out, first to demonstrate the problems of reconciliation in dual optimization, and second to show by comparison of dual with linear optimization alone, the marked improvement in system response. The control system chosen has the following inherent features:
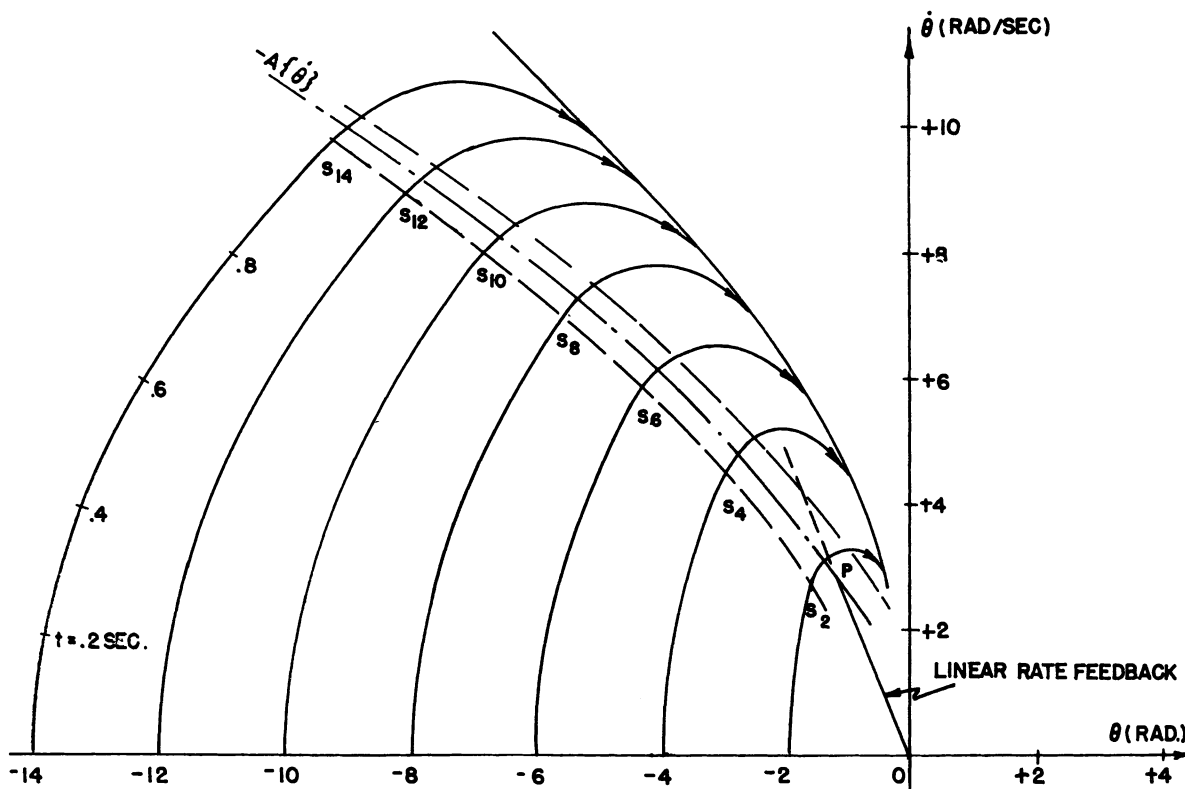

Fig. 9. Determination of $A\{\dot{\theta}\}$

1. Controller, linear up to the static torque saturation, which occurs when $|L| = 10$ foot-pounds. Rate of reversal of torque cannot exceed 50 foot-pounds per second.

2. Load, constant inertia, $I = 1$ slug-foot².

1. *Linear optimization.* The imposed means of linear optimization is a linear rate generator driven by the system output, that is, the feedback $\theta_f = -\theta - a_1\dot{\theta}$. By means of a linear analysis which includes a consideration of system time lags of the order of 0.02 second, the optimum controller gain is selected as $G = 20$ foot pounds per radian; and for a little less than critical damping, the rate feedback coefficient, a_1, becomes 0.4. Therefore, the static controller characteristic is that of Fig. 8(A), and the basic control system, (neglecting the small time lags) is that of Fig. 8(B).

2. *Optimum response trajectories.* The saturated torque, L_{max}, acting upon a constant inertia, produces the parabolic acceleration and deceleration trajectories of Fig. 9. During saturated reversal

$$\left(\frac{dF}{dt}\right)_{max} = -50 = \frac{d}{dt}(\dot{\theta}) = \dot{\theta}\frac{d}{d\theta}\left[\dot{\theta}\frac{d\dot{\theta}}{d\theta}\right]$$

$$= \dot{\theta}^2\frac{d^2\dot{\theta}}{d\theta^2} + \dot{\theta}\left(\frac{d\dot{\theta}}{d\theta}\right)^2$$

Hence

$$\frac{d^2\dot{\theta}}{d\theta^2} = -\frac{50}{\dot{\theta}^2} - \frac{(d\dot{\theta}/d\theta)^2}{\dot{\theta}} \tag{5}$$

At an arbitrary point on the deceleration trajectory, $\dot{\theta}$ and $d\dot{\theta}/d\theta$ are known, which allows equation 5 to be solved for $d^2\dot{\theta}/d\theta^2$, the rate of change of trajectory slope. By stepwise construction, the trajectory during reversal can be determined backwards. As an alternative procedure, the trajectories can be found by standard phase-plane geometric techniques.[12] The resultant optimum response set is shown in Fig. 9.

3. *Zero time-lag optimization.* The width of the static control band is 1.0 radians, Fig. 8(A), and is much narrower than the optimum reversal zone, except near the null. Therefore, saturated reversal is assured except near the null, and the optimum and signaled switching lines should be made coincident, as mentioned earlier. Since the load does not vary with θ, the complete optimum feedback is $A_0\{\dot{\theta}\}$, which can be read directly from Fig. 9.

4. *Optimizing with time lags.* Assume the various feedback and controller time lags can be lumped as a single time lag of 0.02 second. This lag delays the actual
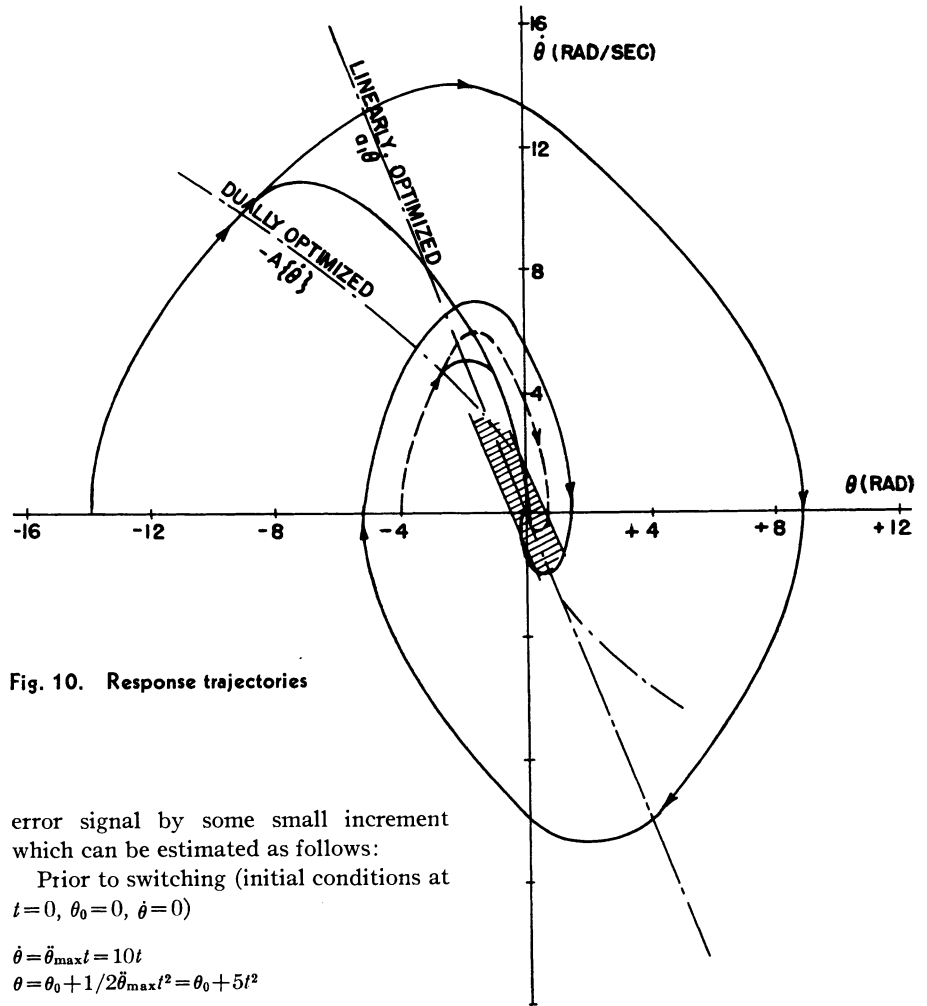

Fig. 10. Response trajectories

error signal by some small increment which can be estimated as follows:

Prior to switching (initial conditions at $t = 0$, $\theta_0 = 0$, $\dot{\theta} = 0$)

$$\dot{\theta} = \ddot{\theta}_{max}t = 10t$$
$$\theta = \theta_0 + 1/2\ddot{\theta}_{max}t^2 = \theta_0 + 5t^2$$

Note that θ_0 is always negative for left half-plane trajectories. From Fig. 9, an approximate expression for $A\{\}$ is

$$A_0\{\dot{\theta}\} = 0.06\dot{\theta}^2 + 0.3\dot{\theta} = 6t^2 + 3t$$

And the zero time lag error signal is

$$\eta_0 = -(\theta + A_0\{\dot{\theta}\}) = -\theta_0 - 11t^2 - 3t$$

Inserting the single time lag network, the incremental change in the signal is (where $s = d/dt$, $\tau = 0.02$ second)

$$\eta - \eta_0 = \Delta\eta = -\tau s\eta$$

Therefore

$$\Delta\eta(s) = \frac{\theta_0}{s+50} + \frac{3}{s(s+50)} + \frac{22}{s^2(s+50)}$$

(By Laplace convention, $s = \frac{d}{dt}$.)

And noting t at switching point is of the order of unity

$$\Delta\eta(t) \sim 0.06 + 0.44t$$

At switch point, s_{14} for example, t is about 1 second, and the increment $\Delta\eta \sim 0.5$ radians. As a first-order approximation, $A_0\{\dot{\theta}\}$ should be advanced to initiate switching 0.5 radian earlier. Obviously, for trajectories nearer the origin the correction is smaller, and for trajectories

more remote the correction becomes quite important.

5. *Unsaturated region.* In this example, since no unsaturated nonlinear gap exists, the limits of the unsaturated region do not have to be precisely determined. Therefore, the control band is assumed to be unsaturated when the static control band is wider than the optimum reversal zone, which is indicated in Fig. 9 by the discontinuation of both zones near the null.

6. *Reconciliation of the linear and the saturated optimizations.* The nonlinear rate feedback $A\{\dot{\theta}\}$ had been determined, and the null region rate feedback $a_1 \cdot \dot{\theta}$ has been selected. Both of these are plotted in Fig. 9. From these two, a composite feedback must be formed. For the purposes of this report, it is chosen as linear, $a_1\dot{\theta}$, up to point P and thence nonlinear, $A\{\dot{\theta}\}$. From the phase-plane plots alone, one might be tempted to follow $A\{\dot{\theta}\}$ closer to the origin, but this would allow less damping than linear analysis recommends. Near the origin, it is generally wise to let linear analysis take precedence.
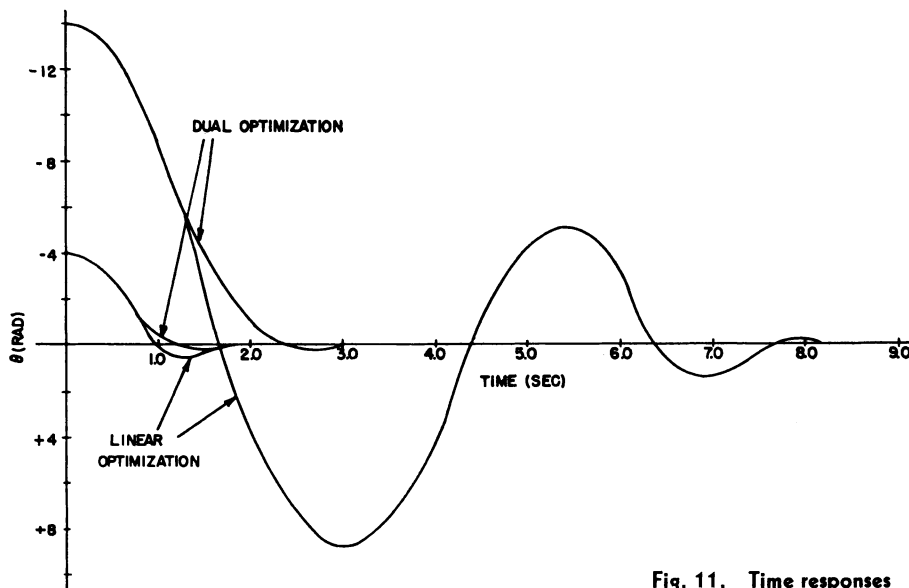
The resultant nonlinear feedback is

Fig. 11. Time responses

plotted in Fig. 10, along with the feedback, $a_1\dot{\theta}$, which would have resulted from linear optimization alone. The comparison of trajectories for the dually and singly optimized systems is shown in Fig. 10, both for large and small requested jumps. Converted to time displays, the responses are again compared in Fig. 11. If the tolerated error is taken to be 0.1 radian, the response time of the linearly optimized system to a 14-radian step is over 8 seconds, while the response time of the dually optimized system is about 2.3 seconds.

Discussion

The engineering problem of producing the desired nonlinear function is one which cannot be completely answered at present. It should be noted that the

identical problem, that of arbitrary function generation, has been facing the designers of analogue computers for several years; and a variety of devices is now available: 1. photoformers,[13] slaving a cathode ray to a shaped screen mask; 2. diode switching arrays,[13] forming functions by linear segments; and 3. simple elements which product specific functions such as square-law, three-halves power, logarithmic, etc. In general, these devices produce outputs which are functions of a single variable.

If a function of two independent variables, $A\{\dot{x}, x_t\}$, is desired, the product approximations $A_1\{\dot{x}\}\cdot A_2\{x_t\}$ or $A_1[\dot{x}\cdot A_2\{x_t\}]$ should be considered, noting that multiplication by means of the rate sensor excitation is often possible.

If, as in some servomechanisms, output rate is not conveniently nonlinearized,

the optimization method may be considered from the viewpoint of nonlinearizing the error signal rather than the rate feedback.

References

1. The Effects of Backlash and Speed-Dependent Friction on the Stability of Closed-Cycle Control Systems, A. Tustin. *Journal*, Institution of Electrical Engineers, London, England, vol. 94, pt. IIA, no. 1, 1947, pp. 143–51.

2. Analyzing the Effect of Non-Linearity in Closed-Cycle Systems. *Ibid.*, pp. 152–60.

3. An Introduction to Servo-mechanisms, Menthuen's Monographs on Physical Subjects (book), A. Porter. John Wiley and Sons, Inc., New York, N. Y., 1950.

4. Nonlinear Techniques for Improving Servo Performance, Donald McDonald. *Proceedings*, National Electronics Conference, Chicago, Ill., vol. 6, 1950, pp. 400–21.

5. Theory of Servo-mechanisms, H. L. Hazen. *Journal*, Franklin Institute, Philadelphia, Pa., vol. 218, no. 3, Sept. 1934, pp. 279–331.

6. Fundamental Theory of Servomechanisms (book), L. A. MacColl. D. Van Nostrand Company, New York, N. Y., 1945.

7. Analysis of Relay Servomechanisms, H. K. Weiss. *Journal of Aeronautical Sciences*, New York, N. Y., vol. 13, July 1946, p. 364.

8. A Frequency Response Method for Analyzing and Synthesizing Contactor Servomechanisms, Ralph J. Kochenburger. *AIEE Transactions*, vol. 69, pt. I, 1950, pp. 270–84.

9. Piecewise Linear Servomechanisms, J. W. Schwartz. *AIEE Transactions*, vol. 71, pt. II, Jan. 1953, pp. 401–5.

10. Stabilization of a Servomechanism Subject to Large Amplitude Oscillation, Emile S. Sherrard. *Ibid.*, pp. 312–24.

11. A Phase-plane Approach to the Compensation of Saturating Servomechanisms, Arthur M. Hopkin. *AIEE Transactions*, vol. 70, pt. I, 1951, pp. 631–39.

12. Introduction to Nonlinear Mechanics (book), Nicolas Minorsky. J. W. Edwards, Ann Arbor, Mich., 1947.

13. Electronic Techniques Applied to Analogue Methods of Computations, G. D. McCann, C. H. Wilts, B. N. Locanthi. *Proceedings*, Institute of Radio Engineers, New York, N. Y., vol. 37, no. 8, Aug. 1949.
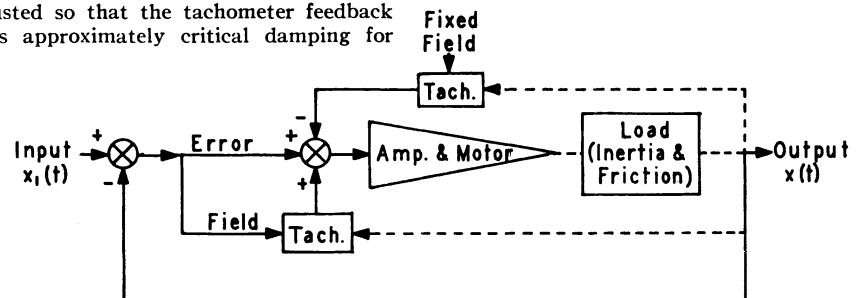

Discussion

D. M. Collier (Oak Ridge National Laboratory, Oak Ridge, Tenn.): As an addendum to the authors' section entitled "Discussion," it should be noted that the nonlinear feedback originally proposed by J. B. Lewis[1] is one practical example of the type of nonlinear rate feedback under discussion. This feedback is rather easily achieved in practice and can be modified to produce the desired nonlinear function $A\{\dot{x}, x_1\}$ certainly for the specific case of a second-order system with a linear load and possibly for other systems. As Caldwell and Rideout[2] point out, the Lewis feedback network, when applied to a torque-limited system, gives performances which approach optimum. The system has a phase-plane plot similar to that shown in Fig. 1(D).

Lewis approached the problem somewhat differently, starting from the well-known second-order linear system with error and derivative feedback. Then by extrapola-

tion from the performance of the linear system with a constant damping coefficient, he devised a modification which makes the damping coefficient a linearly decreasing function of the error. For step inputs his system gives a marked improvement in performance over that obtainable with a linear system.

Briefly, Lewis' feedback network is as follows. A second-order linear system is adjusted so that the tachometer feedback gives approximately critical damping for

the linear system. Then another tachometer feedback has the system error applied to its excitation field, and its shaft connected to the system output. This nonlinear tachometer thus feeds back the product of output derivative times the position error, and this voltage is added to the motor input in the manner shown in Fig. 12. As will be noted, the system has a differential equation of the general form

Fig. 12. Lewis feedback network

$$\ddot{x}+[A-B(x_1-x)]\dot{x}+cx=cx_1$$

It is apparent that this is a decided improvement over the linear system in response to positive-step inputs; but this system is even slower than the linear one for negative-step inputs, as may be seen from considering the equation. For this reason, Lewis added another nonlinearity which takes the absolute value of the error voltage before applying the error to the field of the nonlinear tachometer, with the result that the second term in the foregoing equation becomes

$$[A-B|x_1-x|]\dot{x}$$

In truth, this allows the system to follow both positive and negative input steps, but the system has overshoot for large input steps and is unstable for input steps larger than a certain level determined by the relative values of A and B.

Even in the above form, this nonlinear feedback gives performance approaching optimum for some inputs to a torque-limited motor. It is this writer's opinion that, using the methods outlined in the above papers, the Lewis feedback can be adjusted to give optimum nonlinear performance with a torque-limited motor (the usual case in practice) by removing the absolute value nonlinearity and substituting a polar relay type of device which changes the derivative term in certain combinations at the boundaries of the various phase-plane quadrants. The performance of such an arrangement should be much like that of Hopkin's (see ref. 11 of the paper) "anticipator" system optimized for both step and constant-velocity inputs. The system would be rather like the one described in the authors' section entitled, "A Simple Example." A network of this general type is being investigated by the writer with the objective of obtaining a simple practical feedback network for application in systems which have heretofore largely been investigated only with computers.

REFERENCES

1. THE USE OF NONLINEAR FEEDBACK TO IMPROVE THE RESPONSE OF A SERVOMECHANISM, J. B. Lewis. *AIEE Transactions*, vol. 71, pt. II, Jan. 1953, pp. 449-53.

2. A DIFFERENTIAL-ANALYZER STUDY OF CERTAIN NONLINEARITY DAMPED SERVOMECHANISMS, R. R. Caldwell, V. C. Rideout. *AIEE Transactions*, vol. 72, pt. II, July 1953, pp. 165-70.

Arthur M. Hopkin (Northwestern University, Evanston, Ill.): The authors are to be congratulated on their efforts to extend the concept of control of nonlinear systems by means of controlled reversals of a saturating forcing function. They have extended the concept to permit the use of phase-plane techniques for the design of nonlinear elements to be used in optimizing the response of feedback control systems in which a second-order system is preceded by a switch or relay, which in turn may be preceded by a complicated system involving t me constants and delays. This technique extends the use of certain phase-plane methods so that certain types of higher order saturating feedback systems can be optimized by use of phase-plane methods.

However, attention should be called to the fact that the type of feedback control system which involves a relay-type element followed by a system whose behavior may be described by a third- or higher order differential equation can not be optimized by the phase-plane approach at the present state of the art. Let us consider a simple Ward-Leonard positioning system with a relay contolling the application of voltage to the field of the generator for the purpose of gaining an insight into the problems involved. Fig. 13 shows a schematic diagram of such a system.

If the reference input r is a sudden step in the forward direction, it is logical to have the controller suddenly close the switch S_1 in the forward direction, and thereby set up a flux in the generator which results in an armature voltage, which drives the motor to reduce the error and hence to make c approach r. If the controller will control the time of reversal of the switch s_1 in such a way that the motor velocity bezomes zero at the same instant that the error becomes zero, and if the controller will then open the switch s_1, it might at first seem that there would be an ideal controller for step inputs. However, when $e=0$ and the speed $\omega=0$, and S_1 is open there is still a flux in the generator
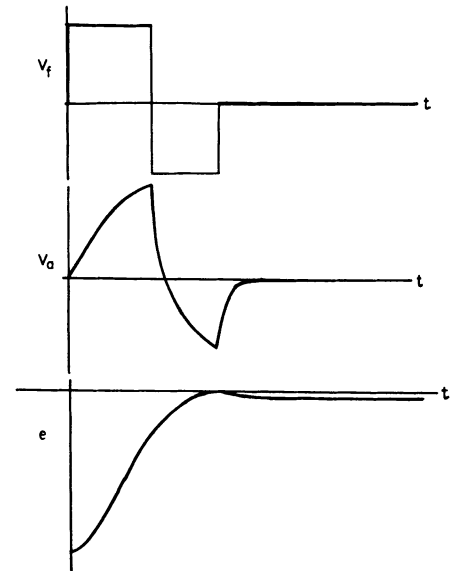

Fig. 14. Performance of the Ward-Leonard system with a single reversal of the control relay

field in such a direction as to produce a negative V_a. This flux would decay at a rate set by the resistance and inductance of the field-winding and the eddy-current paths. Even if the damping resistance R_D were made infinite, and the switch could be made arc free, there would still be a flux-decay time constant because of the parameters of the eddy current circuit. Therefore, the motor would reach zero error at zero velocity, but it would then back up to give a negative error as a result of negative voltage still produced by the decaying field. The controller might go through another control cycle to counteract the remaining error, but it could not be considered an optimum controller. The performance described is delineated graphically in Fig. 14.

This performance could be made to approach an optimum by introducing a second
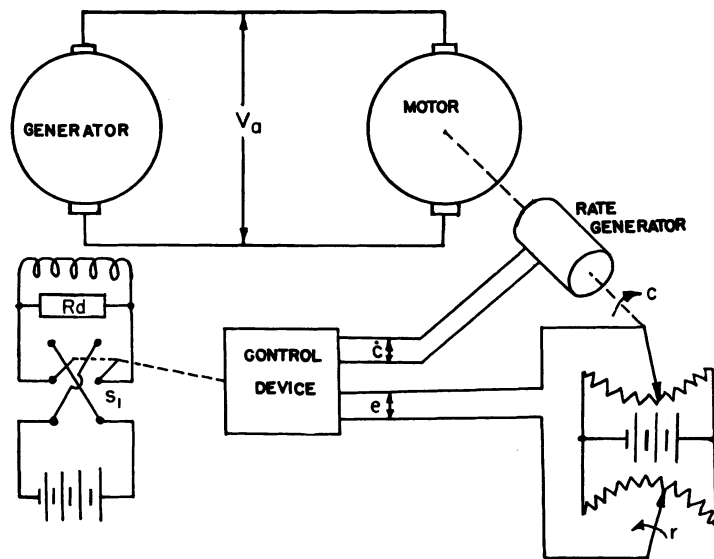

Fig 13 (left). A Ward - Leonard positioning system controlled by a field-reversing relay
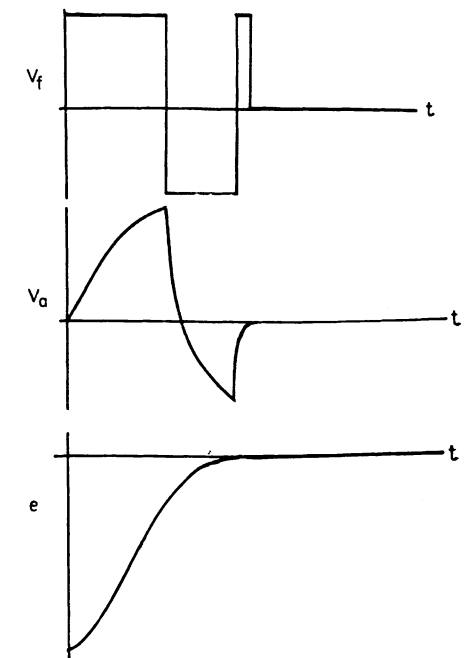

Fig. 15 (right). Performance of the Ward-Leonard system with two reversals of the control relay

reversal into the cycle just before the zero error position was reached. The purpose of this reversal would be to bring the field flux quickly to zero. By proper choice of the first and second reversal times, the field flux could be brought to zero at the same instant that the speed and the error were brought to zero. Opening of the field switch at this point would leave the system in equilibrium with no stored energy. Fig. 15 shows this type of performance. The problem in setting up this type of controller is to find a device which will choose correctly both the first and second reversal times for all boundary conditions and for all amplitudes of stop-reference input change.

Since the effects of armature circuit inductance have been neglected in this discussion, the system following the relay could be approximated by a third-order equation with two energy storage points. One might be tempted to say that a control relay element followed by an nth order system would be required to make $(n-1)$ reversals give optimum response to a step input, based on the discussion involving second- and third-order systems. However, a much more careful study of higher order systems would be required before such a statement could be made. It is evident that, as the system controlled by the relay element becomes of higher order, more reversals would be required for optimum response to a step input.

The authors have helped advance the frontiers of knowledge in this field by their contributions. However, where the problem is that of optimizing the step response of a feedback system which involves a relay device followed by a higher order system, there are still many unsolved problems.

R. S. Neiswander and **R. H. MacNeal:** The authors would like to express their gratitude to Prof. Hopkin and Dr. Collier for their generous comments and contributions concerning the paper.

With the growing enthusiasm in nonlinear control and consequent large number of excellent papers, the engineer is sometimes a little overwhelmed at the apparent diversity of techniques. A general criticism of the paper is that it does not make its intentions sufficiently clear. Therefore a highly simplified classification of techniques is given now to show more precisely the application of the proposed method; the system's region of operation has been used as a basis for classification.

Null Region Methods. The null region is the region of small motions. Linear methods are often satisfactory, and several methods are available for considering nonlinear null region effects such as backlash and coulomb friction.

Unsaturated Region Methods. Excepting on-off controllers, there is a dynamic operating region away from the null which is unsaturated. The most common approach has been to keep this region as linear as possible, allowing linear analysis. Techniques by McDonald (ref. 4 of the paper) and Lewis (ref. 1 of Mr. Collier's discussion) have been proposed recently to improve response of systems otherwise linear in this region by intentionally nonlinearizing elements. A frequent result of such methods is decreasing the rate feedback over that specified by linear criteria. When prominent unintentional nonlinearities exist in the system, no general methods are available.

Saturated Region Methods. A quite common design practice is to designate the point at which saturations occur as the boundary of the system's usefulness. Excepting special situations such as that introduced by severe noise limiting, the reason for imposing the boundary is simply that linear analysis breaks down. Actually, control systems are capable of much wider dynamic regions of operation. The present method is intended to handle a specific performance requirements in this region, and

at the same time be compatible with methods used in the null region. A frequent result of this method is to increase rate feedback over that dictated by linear null criteria.

In general, methods which apply to one region do not necessarily apply to another region. A system required to operate over a wide dynamic range might well be designed by a combination of two or three applicable methods.

Prof. Hopkin's discussion is concerned primarily with on-off control systems in which there is no region of linear or unsaturated response. In the type of system envisaged in our paper there is a small region of linear response near the null within which the system transients are allowed to decay naturally. Hence there will be no steady-state position error regardless of the order of the equation describing the system. The response time of such a system to a step input will be somewhat longer than that obtainable with an ideal on-off control, but it has compensating advantages in stability and simplicity. The method described in our paper is not essentially a phase-plane technique and the phase-plane cross sections should be considered as helpful response "pictures" not actually vital to the method.

Dr. Collier has pointed out a family of useful, nonlinear optimizing elements which were completely overlooked by the authors. As he has suggested, the feedback arrangement utilized by Lewis can be simply extended to supply a large assortment of functions. For example, the feedback function of Fig. 9 was in the paper assumed to be $A\{\dot\theta\}$. It would be just as appropriate to consider the feedback as $a\{\theta\}\dot\theta$ (linear rate feedback with a variable coefficient); or $b\{\dot\theta\}\theta$ (linear static error with a variable coefficient); or any of a multitude of combinations of functions of θ and $\dot\theta$. The real advantage of these rephrasings appears when attempting to physically create nonlinear elements.