THE USE OF AUTOMATIC EXPERIMENTATION COMBINED WITH MATHEMATICAL MODELS IN OPTIMALIZING CONTROL OF CONTINUOUS PROCESSES

JENS G. BALCHEN

Norwegian Institute of Technology, Trondheim, Norway

THE PROBLEM

A large percentage of the processes used in the chemical industry, the paper and pulp industry and the petroleum industry can be classified as being continuous processes. A continuous process in contrast to a batch process is characterized mainly by a continuous flow of raw materials to be converted going into the process and that the objectives and the structure of the process are maintained fixed over a long period of time. The necessity of automatic control in such processes is due to the existence of undesirable disturbances acting upon the process so as to change the working conditions away from the desired state. Another reason for applying automatic control is in many cases that the process in itself is unstable or at least behaves as an integrator so that important process variables would tend to increase or decrease indefinitely if no feedback control was employed. The disturbances encountered in industrial processes can certainly be of quite a different nature. Changes in steam pressure in a system supplying heat energy to a process is an example of an external disturbance which is rather easy to measure. Changes in the chemical composition of a raw material entering into a reactor is another example of an external disturbance as seen from the reactor itself. In this case the measurement of the disturbance is more complicated, but not impossible. A well-known example of an internal disturbance is changes in the activity of a catalyst enclosed in a chemical reactor. Such changes are usually rather smooth, decreasing time-functions if the working conditions of the process are kept relatively constant. The actual activity of the catalyst, however, is rather difficult to measure directly. Therefore, one is forced to make estimates of the activity of the catalyst based on the past and present values of the most significant process variables which contribute to reduction of the activity. The control problem is usually greatly simplified if measures of the most significant disturbances are available and if the static and dynamic relationships of the process are known in some detail. Feedforward techniques combined with feedback controls can then be employed so as to achieve the highest possible speed of recovery after the influence of the disturbances.

A most important part of the planning and design of a control system is certainly to specify a realistic objective of the whole system. Until recently it has been customary in process control to device systems that maintain constant conditions in the process. The effects upon the process of external disturbances (which are easily controlled) are usually reduced by local controls in feed rates, pressures and so on. Secondary variables such as reactor temperature, production rate or chemical composition of the product are usually controlled at fixed set-points by means of cascade controls of the input quantities. The objectives of the control system are then fixed and only intermittently and manually adjusted by the operator based on information obtained from an analysis of the quality and amount of product and the amount of energy and raw materials used. The reasons for using essentially fixed set-point control systems are quite obvious. For instance, the designer of the process might know that by maintaining the temperature in a reactor at say 100°C his product will have a certain characteristic quality which is desired. By lowering the temperature to say 90°C the production rate might be reduced appreciably without too much increase in quality. And by increasing the temperature to say 110°C the production rate is only slightly increased but the quality is reduced. Therefore he decides that 100°C is a good temperature for the process. However, when changes in the composition of the raw material occurs and when environment conditions change he also knows that the optimum temperature will change. In this case a more elaborated objective of the control system should evidently have been specified. Keeping the temperature in the reactor constant is certainly not the final objective of the control system, but merely an aid in reaching a result of more direct importants as for instance maximum production at a given quality of the end product or maximum profit taking into account both prices and amounts of raw materials used, energy used and products produced.

If it is possible to develop an expression for a performance index (performance criterion) which is to be kept at a maximum (or minimum) to assure that the process is working at its best, and if it is possible to device a control system which automatically forces the process to work under such optimal conditions, then we talk about an optimalizing control system. Applying an optimalizing control system to a single process or a small group of processes in a large plant means that we are only performing what is termed a "suboptimization" because from a management and production planning point of view it is always possible to give a more general objective for the plant as, for instance, maximizing the profit of the whole plant over a certain period of time. This, of course, is also a kind of sub-optimization because if the plant is only one of many in a company, the company's objective is to maximize the whole profit. From a control engineering point of view, however, it is certainly necessary to set limits to the size of the system subject to an automatic optimalizing control.

THE DYNAMIC OPTIMAL SYSTEM

Acting upon the process are in general three kinds of variables (external and internal):

1. Variables that are both measurable and controllable.
2. Variables that are measurable but non-controllable.
3. Variables that are neither measurable nor controllable.

The only means of bringing the process to optimal working conditions is to adjust the controllable variables. If the process is in a stationary state, i.e. all disturbances, both measurable and non-measurable, are non-varying quantities or at most slowly varying quantities then optimum conditions can be maintained by at most slow adjustments of the controllable variables. In such a case the dynamic phenomena in the process do not enter into the problem in a significant manner. If, on the other hand, the disturbances are continuously changing at a rate which at least is in the same order of magnitude as the speed of the response of the process, then the time patterns used for the corrective actions in the controllable variables are highly significant.

A system is said to be dynamic optimal if corrective actions are generated which under any circumstances yield maximum value of the performance index.

To illustrate the importance of dynamic optimal systems consider a very simple example. Fig. 1 shows a block diagram of a hypothetical process with only one controllable variable, x_1. The only disturbance (non-controllable) acting upon the process is x_2. The performance index z is related to variables in the process as described previously. Suppose now that it has been possible to find the static (stationary) relationships between the quantities x_1, x_2 and z, so that under non-varying conditions

$$z = f(x_1, x_2). \tag{1}$$

The three-dimensional surface described by (1) is called a static or steady-state performance surface. Such a surface is shown in Fig. 2, where the value of the performance index is indicated by contours. If the disturbance x_2 is changing very slowly, optimum conditions (maximum z) will be maintained by adjusting x_1 in accordance with the solid curve of Fig. 2 which is

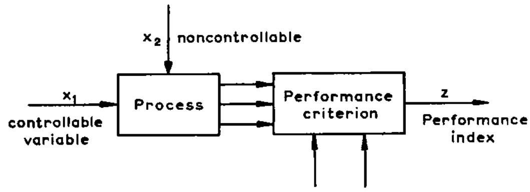

Fig. 1.—Simple two-variable problem.

the solution of

$$\partial z/\partial x_1 = 0 \quad x_{1,\,\mathrm{opt}} = g(x_2). \tag{2}$$

Provided that x_2 is measurable and slowly changing, it is thus easy to make a system to maintain optimum conditions. Now consider a case when x_2 is changed stepwise from one constant value to another, as indicated in Fig. 3*A*. It is assumed that optimum conditions existed before the change in x_2 occurred. If no correction is made in x_1 the performance index will decrease at a rate given by the dynamics of the process and eventually end up at a value as given by the point P_2 in Fig. 2. The time pattern of change in the performance index z might be of the form indicated by curve *a* in Fig. 3*C*. If a correction is made in x_1 based on the relation of (2) (stepwise correction in x_1 as indicated by curve *b* in Fig. 3*B*), then the time pattern of the performance index might be of the form indicated by curve *b* in Fig. 3*C*. The performance index now eventually reaches the correct level which is the steady-state maximum, but it has a large transient of low performance. This transient loss of performance can be greatly reduced by proper attention to the dynamics of the process. By using a corrective action in x_1, as indicated by the curve *c* in Fig. 3*B*, the transient in the performance index is reduced as indicated by curve *c* of Fig. 3*C*. In case the performance index is a measure of profit, the area between the different time patterns of Fig. 3*C* gives an idea of the improvement obtained by proper dynamic optimization. Obviously the same arguments apply in case the disturbance x_2 is varying in a random manner and in cases of multi-dimensional performance surfaces with a number of controllable variables and a number of disturbances. The following questions immediately arise:

1. How much do we have to know about the static and dynamic relationships of the process?
2. Is it necessary to get instruments to measure all the measureable, but non-controllable, disturbances?
3. What is the effect of the non-measurable disturbances?

A straightforward answer to the first question is that the more knowledge has been acquired about the mechanisms of the process, the more accurate is it possible to make the dynamic optimization. However, because accurate static and dynamic relationships are very hard to find for most processes it

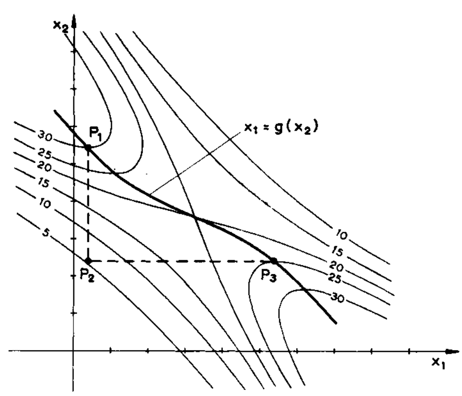

Fig. 2.

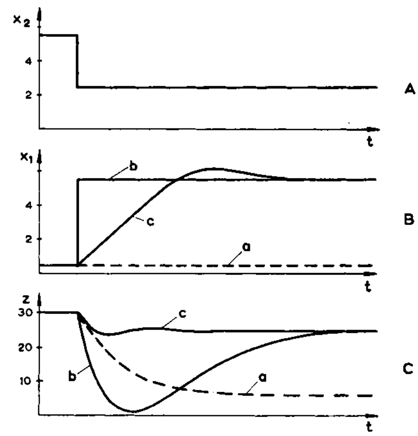

Fig. 3.

Fig. 2.—3-dimensional steady-state performance surface.

Fig. 3.—Time patterns of disturbance, corrective action and performance index.

is important to know that the accuracy required in order to reach almost optimal conditions is not necessarily very high.

For almost linear processes it is conceivable that a steady-state performance surface of quadratic form (or at most piecewise quadratic form) will describe the static behaviour of the process adequately in the neighbourhood of optimum conditions. For example, a reasonable approximation to the performance surface of Fig. 2 in the actual range of variations of x_2 could be a quadratic surface, which in this case would be a hyperbolic paraboloid yielding a straight line for the function $x_1 = g(x_2)$. If a single quadratic surface does not give adequate accuracy different parts of the actual surface can be approximated by different quadratic surfaces which leads to a piecewise linear relation for the expression $x_1 = g(x_2)$. Furthermore, for almost linear processes the dynamic relationships as obtained by linearization techniques assuming relatively small perturbations around average working conditions are expected to give acceptable accuracy. This is so because a slight error in the dynamic model of the process will only give a slight dynamic reduction of the performance index.

The second question has the obvious answer that the sooner the effect of the disturbance can be measured, the sooner it is possible to take corrective action. If measurements on significant and measurable disturbances are neglected and the effect of the disturbances only detected at the process output, then the inherent filtering action of the process will make the problem of restoring information about the disturbances more difficult. Therefore, when disregarding the expense involved in measuring the disturbances directly, the answer to the question is that all possible disturbances should be measured. The information obtained is valuable even in cases when the measurements are obscured by noise not related to the actual disturbance; for instance, due to the use of primitive measuring techniques. The effect of noise in the measurements can usually be greatly reduced by proper filtering techniques.

The non-measurable disturbances certainly represent a great problem. It is conceivable that some of the effects of disturbances, that cannot be measured directly, can be detected indirectly by measurements on dependent output variables in the process which are more or less closely related to these disturbances. Provided that the mechanisms of the process are rather well known, it might then be possible to get at least rough estimates of some of these unknwon disturbances. However, there will always be something left over which is completely unpredictable. The only known way to compensate for such disturbances is to perform an experiment on the process by slowly searching over the performance surface until the optimum point has been found. Experiments on dynamic processes, however, require time, especially if there are many disturbances acting simultaneously and many controllable variables to be corrected. This means that it is impossible to make fast corrections for rapidly changing non-predictable disturbances.

THE FEEDFORWARD COMPENSATOR

Consider again the simple example of the process in Fig. 1. A more detailed block diagram of this process including the performance criterion is drawn in Fig. 4*A* assuming that the process has only two dependent variables y_1 and y_2 entering into the performance criterion. Furthermore, it is assumed that the relations in the steady-state between these two dependent variables and the input variables x_1 and x_2 are given by non-linear functions of the form

$$\left.\begin{aligned} y_1 &= \psi_1(x_1, x_2) \\ y_2 &= \psi_2(x_1, x_2) \end{aligned}\right\} \tag{3}$$

The non-linear mechanisms of the process are also assumed to be such that the dynamic effects of realistic changes in x_1 and x_2 can be linearized and be placed in front of the non-linear, non-dynamic functions ψ_1 and ψ_2 as indicated in Fig. 4*A*. The linearized dynamic relations are given in terms of normalized transfer functions $h_{ij}(s)$ which have the property

$$|h_{ij}(s)|_{s \to 0} \to 1, \tag{4}$$

provided they have no poles or zeros at the origon. The new variables introduced in the block-diagram of Fig. 4*A* (x_{01}, x_{11}, x_{21} and x_{02}, x_{12}, x_{22})

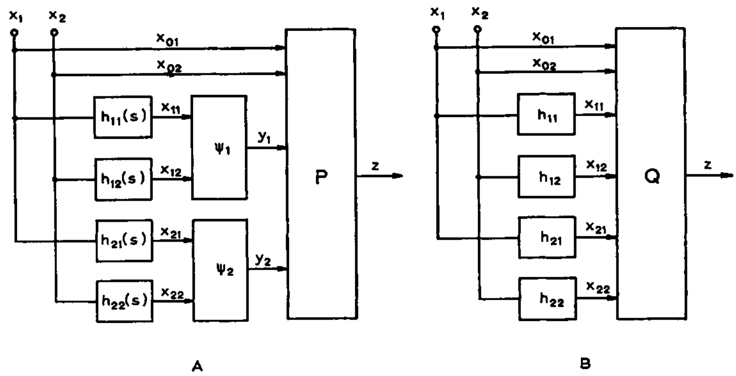

Fig. 4.—Dynamic model of simple process.

are dummy variables with the property that in the steady-state

and
$$\left.\begin{aligned} x_{01} = x_{11} = x_{21} = x_1 \\ x_{02} = x_{12} = x_{22} = x_2 \end{aligned}\right\} \tag{5}$$

As a result of the approximation chosen we see that the non-linear function relating the performance index z to these dummy variables is completely non-dynamic. Therefore, if we wish to maximize (or minimize) the instantaneous value of z by correcting x_1 we only have to satisfy the following relations:

$$\frac{\partial z}{\partial x_{01}} = 0, \quad \frac{\partial z}{\partial x_{11}} = 0 \quad \text{and} \quad \frac{\partial z}{\partial x_{21}} = 0. \tag{6}$$

The dummy variables appearing in (6), however, are all dynamically related and it is impossible to satisfy all the equations simultaneously by correcting x_1 only. Because we have too many equations to determine one variable we are led to a problem in calculus of variations in order to determine the proper way of correcting x_1 to reach a realizable dynamic optimum. Such an approach, however, will become extremely complicated if not impossible when the number of variables involved is large and the dynamic relations are of higher than first order. Therefore, a simplified approach is suggested which yields a result very close to the best possible solution. Rather than satisfying the requirements of (6), the following expression is used:

$$\sum_{i=0}^{2} \frac{\partial z}{\partial x_{i1}} = 0. \tag{7}$$

In particular, if the non-linear function relating z to the dummy variables x_{ij} is a quadratic form (Fig. 4*B*) we find

$$\sum_{i=0}^{2} \frac{\partial z}{\partial x_{i1}} = \sum_{i=0}^{2} r_{i1} + \sum_{i=0}^{2} \sum_{l=1}^{2} \sum_{k=0}^{2} \left(t_{\substack{kl\\i1}} + t_{\substack{i1\\kl}}\right) x_{kl}, \tag{8}$$

which is a linear equation. In this expression the coefficients r and t are constants determining the shape of the performance surface. The steady-state optimum is determined by

$$\frac{\partial z}{\partial x_1} = r_1 + \sum_{l=1}^{2} \left(t_{\substack{l\\1}} + t_{\substack{1\\l}}\right) x_l = 0, \tag{9}$$

where
$$r_1 = \sum_{i=0}^{2} r_{i1},$$

$$t_{\substack{l\\1}} = \sum_{i=0}^{2} \sum_{k=0}^{2} t_{\substack{kl\\i1}},$$

$$t_{\substack{1\\l}} = \sum_{i=0}^{2} \sum_{k=0}^{2} t_{\substack{i1\\kl}}.$$

Thus by comparing coefficients of (8) and (9) we find that the linear combination (7) yields the exact optimum solution at least in the steady-state. Furthermore, it can be shown that by using the linear combination (7) we get the reachable solution which is least apart from the exact but non-reachable solution (6). Because we know that

$$x_{kl}(s) = h_{kl}(s) \cdot x_l(s), \tag{10}$$

we can solve (7) directly and get

$$x_1 = -[H_{12}(s)/H_{11}(s)]\, x_2, \tag{11}$$

or in general
$$x_j = -\frac{1}{H_{jj}} [H_{j1} \cdot x_1 + H_{j2} \cdot x_2 + \cdots], \tag{12}$$

where
$$H_{jl} = \sum_{i=0}^{q} \sum_{k=0}^{q} \left(t_{\substack{kl\\ij}} + t_{\substack{ij\\kl}}\right) h_{kl}(s). \tag{13}$$

The result of (12) says that in case of a quadratic surface we get linear transfer functions $H_{jl}(s)$ relating disturbance x_l to the correction x_j. Fig. 5 shows a general block diagram for generation of $x_{1,\text{opt}}$ in case the controllable variables are $x_1 ... x_m$ and the non-controllable but measureable disturbances are $x_{m+1} ... x_n$. In the simple case of only one disturbance and one controllable quantity the block diagram of Fig. 5 will be reduced to one single block. This block then constitutes the optimum feed forward compensator to generate the proper corrections based on a knowledge of the actual disturbance.

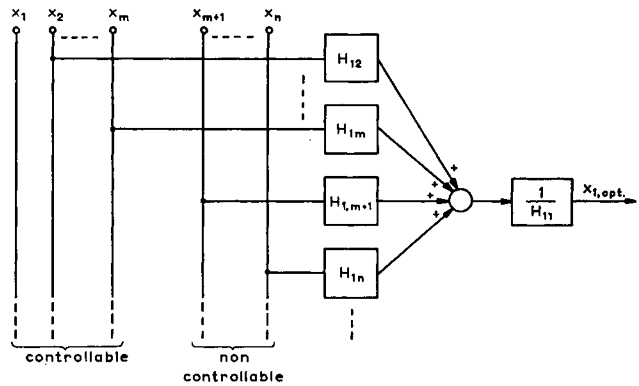

Fig. 5.—General feed forward compensator.

The preceding development was based on the assumption that the performance surface including the non-linearities of the process could be adequately approximated by a quadratic surface. If this is not so, a piecewise quadratic surface will be necessary. This implies that the coefficients

$$r_{ij},\ \underset{ij}{t_{kl}} \quad \text{and} \quad \underset{kl}{t_{ij}}$$

are piecewise constant. As these coefficients enter into the transfer functions of Fig. 5 it is a relatively simple matter to have them corrected provided that the number of variables determining the value of each coefficients is reasonable. It can be expected, however, that only a few of the coefficients will have to be considered as piecewise constants rather than constants. This statement is based on the fact that the feed forward compensation will never be perfect anyway because of complete lack of knowledge about some of the disturbances. The accuracy required from the realizable part of the compensator will therefore have to be governed by engineering judgement.

AUTOMATIC EXPERIMENTATION

The direct feed forward compensator generating the corrective actions in the controllable variables based on the measured disturbances takes care of the coarse corrections. If the accuracy achieved in this manner is not adequate, due to non-predictable disturbances of appreciable effect, the optimum working conditions will have to be found by experimentation. Consider again the simple example of Fig. 1. This time, however, the disturbance x_2 is non-measureable. The steady-state performance surface as sketched in Fig. 2 still exist, but the coordinates of the working point on the surface are not known. Redrawing the information from Fig. 2

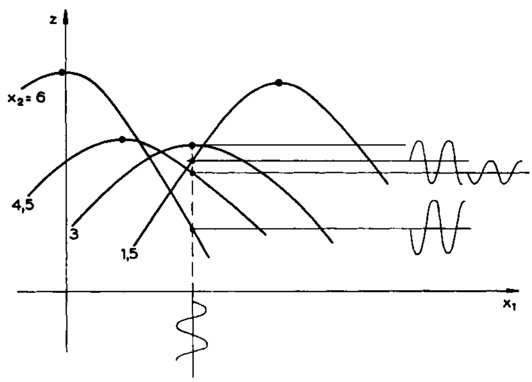

Fig. 6.—Steady state performance surface of Fig. 2 redrawn to show the results of perturbing x_1.

as shown in Fig. 6 the performance index z is a function of x_1 for four different values of x_2. The problem now is to find the value of x_1 which yields

$$\partial z/\partial x_1 = 0$$

without knowing the magnitude of x_2. A number of different experimental procedures have been suggested for this problem, but only one shall be considered here, namely the one originally suggested by Draper and Li [2] and further studied by Eykhoff and Smith [3]. The method involves applying a small sinusoidal perturbation in a controllable variable and detecting the magnitude and phase angle of the response in the performance index z. Fig. 6 indicates that if the working point is to the left of the optimum point then the nearly sinusoidal response will have zero phase angle with respect to the input perturbation whereas 180° phase shift would result if the working point is to the right of the optimum point. At the optimum point the response has zero magnitude. If the searching frequency (perturbation frequency) is chosen so low that the phase shift introduced by the process dynamics is less than, say, 45° a scheme as sketched in Fig. 7 can be used to automatically adjust the controllable variable x_1 to reach optimum steady-state conditions. The quantity z (assumed to be available as continuous measurement) is passed through a highpass filter to remove D.C. offsets. The output of the highpass filter is multiplied by the input sinewave producing a product with a mean value proportional to

$$\partial z/\partial x_1.$$

The multiplier output is positive if the working point is to the left of the optimum and vice versa. By integrating this signal and adding the result to x_1, the working point will continuously move towards the optimum point at which the multiplier output has zero mean value. The speed of recovery of the controllable variable x_1 after an initial offset from optimum conditions

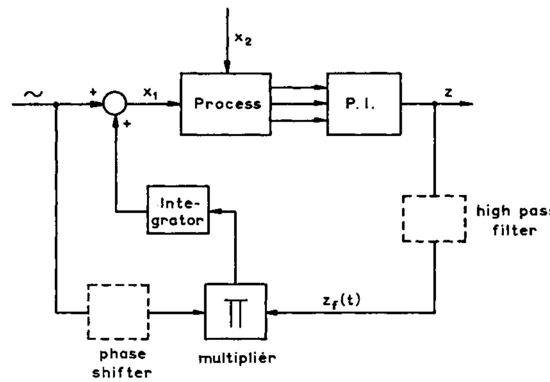

Fig. 7.—Simple scheme for direct, experimental optimalizing control.

is mainly dependent upon the searching frequency used and the amount of "noise" present in z. By shifting the phase of the input sinusoid fed to the multiplier some compensation can be achieved for the influence of the phase shift produced by the process. There is, however, a limit to what extent this can be done because the dependent variables determining the performance index might have different phase shifts and it is highly possible that the phase shifts are changing with the disturbances. For stability reasons the maximum permissible phase shift between the two sinusoids entering the multiplier is 90°.

Recovery times in the order magnitude of 2–5 times the period of the searching frequency are reasonable for such a system. If the process has one significant time constant and the searching frequency is chosen such that the phase shift in the process is 45° then the recovery time of this optimalizing loop will be in the order magnitude of 10–30 times the process time constant. In case of more than one controllable variable different searching frequencies are chosen in each channel. Naturally adding more loops will increase the recovery time further. Quite a high number of inputs can be handled by such a multiloop experimental optimalizing scheme if only the speed of recovery is made low enough.

From an instrumental point of view, particular attention should be paid to the unit performing the multiplication. Because the useful signal in the quantity z can be expected to be rather small compared with the noise, it is required that the multiplier has almost perfect resolution and linearity for a large range of signal magnitudes. Analog multipliers of the types ordinarily used in analog computers do not meet these requirements. A special kind of potensiometric multipliers has been developed for this purpose. The potentiometer is made up of a printed circuit commutator with 30–50 equally spaced segments. Each segment is connected to a voltage divider so arranged that the voltage distribution along the commutator is sinusoidal. The brushes, located diametrically are made to rotate around the commu-

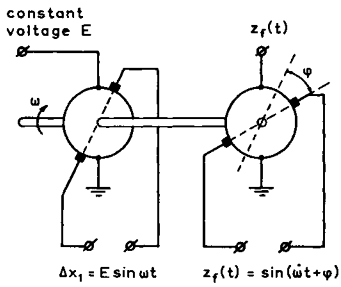

Fig. 8.

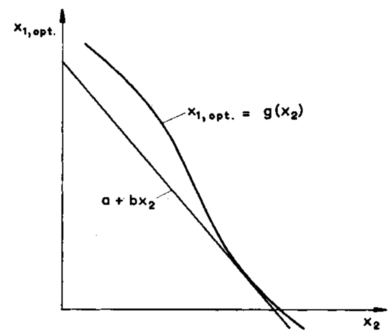

Fig. 9.

Fig. 8.—The sine wave generator and multiplier used in Fig. 7.

Fig. 9.—Interpretation of the steady state parameters (a and b) in a simple feedforward compensator.

tator surface. The brushes of two such commutators are driven by the same shaft at the same speed, only the brushes are shifted an angle φ as indicated in Fig. 8. One commutator is fed with a constant voltage. The voltage between the brushes will then be a sinusoid which is used as the searching signal. The other commutator is fed by the output from the highpass filter producing between the brushes the desired product function $z_f(t) \cdot \sin(\omega t + \varphi)$. In cases of relatively high searching frequencies ordinary electronic integrators can be used to integrate the multiplier output whereas electric motor integrators will prove advantageous for very slow searching frequencies.

Because the speed of response of the purely experimental optimalizing system is very low it is obviously desirable to combine the experimental method with the direct feed forward compensation. In that case, the output of the integrator succeeding the multiplier should adjust one of the parameters of the feed forward compensator rather than directly adjust the magnitude of the controllable variable. This principle is in accordance with the methods used by a human operator when the process of learning is involved. Also because of the low speed the experiment can offer in gaining information about the state of the process this information can only be used for adjusting steady-state parameters of the feed forward compensator. Again referring to the simple example of Fig. 1 and the steady-state performance surface of Fig. 2, we have the relationship

$$x_{1,\,\mathrm{opt}} = g(x_2) \approx a + bx_2 \tag{14}$$

as sketched in Fig. 9. There are two parameters, a and b, to be adjusted in this steady-state feed forward compensator. Because a sinusoidal perturbation of either one of these parameters passes through identical process dynamics it is conceivable that the same searching frequency can be used for both a and b provided that the two sinusoids are shifted 90°. The perturbations are thus

$$\Delta a \cdot \sin \omega t \quad \text{and} \quad \Delta b \cdot \cos \omega t$$

and the outputs of the multipliers preceding the integrators in the loops correcting a and b respectively are

$$z_f(t) \cdot \sin(\omega t + \varphi) \quad \text{and} \quad z_f(t) \cdot \sin(\omega t + \varphi).$$

The advantages of using the same frequency for both perturbations is obviously that the generation of the cosine wave and the multiplication by the cosine wave can be performed by the same unit as shown in Fig. 8 only by adding two more brushes to each commutator.

ACKNOWLEDGEMENT

Part of the work briefly reported here has been done by the author at University of California, Berkeley, on a grant from the Royal Norwegian Council of Scientific and Industrial Research.

REFERENCES

1. BALCHEN, J. G., A Study of Dynamic Optimization of Continuous Processes, Norwegian Institute of Technology. (Report to be published.)
2. DRAPER, C. S. and LI, Y. T., Am. Soc. Mech. Eng. Publication, Sept. 1951.
3. EYKHOFF, P. and SMITH, O. J. M., Optimalizing Control with Process-dynamics Identification. (To be published.)