Chapter 3

Search Strategy

In a sense, adaptive control represents a combination of Adaptive Control Constraint (ACC) and Adaptive Control Optimization (ACO). As with ACO, an index of performance (IP) must be defined for the system which indicates overall system performance. As with ACC, measurements of the process are taken to determine adjustments in the controller input variables. An adaptive control system is one which operates in an environment that changes over time in an unpredictable fashion.

When model identification is infeasible, it may still be possible to make a determination of the IP for the process. This can either be done through direct observation of the IP or by calculating the IP from measurements of the process variables that determine its value. Even though the IP can be evaluated, the problem still remains that the relationship between the IP and the process inputs is unknown. Therefore, the values of these inputs optimizing the system performance cannot be determined directly. In this type of situation, especially in the cutting process, some form of search procedure must be resorted.

The general strategy in any of the search techniques is to make adjustments to controller input variables and observe the effects of the output variables. Based on the effects, decisions are made to systematically change the inputs so as to improve the process performance.

3.1 Properties of a search strategy (6)

A search strategy is a procedure of logical computations used to adjust the process inputs in order to try to improve the IP. By repeating the logical procedure, the search strategy tends to move toward the optimum value of IP. As mentioned above, there are many different search strategies. One strategy may be appropriate for some search problems and inappropriate for others. Because the nature of the search problem is that we are dealing with an unknown process, it is often difficult to decide in advance which search strategy would work best. The success of the search is sometimes reduced to a matter of luck, on the part of the programmer selecting the strategy. The general criteria used to judge the effectiveness of a search strategy are

3.1.1 Speed of arriving at the optimum

It is desirable for the strategy to arrive at the optimum IP value in the minimum number of steps. In process control, this is especially important when the industrial process is subjected to frequent shifts and a new optimum set of operating conditions must continually be sought.

3.1.2 Simplicity of the strategy

This is desirable from the viewpoints of the operating personnel who must supervise the process. If the strategy is complicated, it will be difficult to understand by those using it. This may result in the operator overriding the strategy. Simplicity of the strategy is also an advantage in programming the strategy on the control computer.

3.1.3 Capability to deal with difficult search problems

Some search strategies work very well on certain problems, while other strategies are more suited to other

situations. A desirable search strategy is one that is versatile enough to cope with a variety of different search problems.

3.1.4 Stopping criteria

When the optimum has been reached, the search should be terminated. Because of the stochastic nature of many manufacturing processes, the optimum is often disguised by the presence of random noise. This makes it difficult to identify the optimum operating conditions. One criteria used to judge a search strategy is the ability to discern this stopping point.

3.2 Basic definitions

In order to explain the operation of any of the search techniques, it is necessary to establish certainly basic definitions.

3.2.1 Response surface

Response surface is perhaps the most fundamental concept required to understand how search strategy works in the concept of the response surface. A response surface is a mathematical relationship of the IP (or other dependent variable) as a function of the input variables. For two inputs, x_1 and x_2, the response surface can be plotted very conveniently as shown in Fig. 3.1. The plot reads something like a geological survey map. The contour lines are lines of constant IP value. In Fig. 3.1, the value of each contour line is identified.

In conception in Fig. 3.1, a response surface can be defined mathematically as:

$$Z = f(X_1, X_2) \tag{3.1}$$

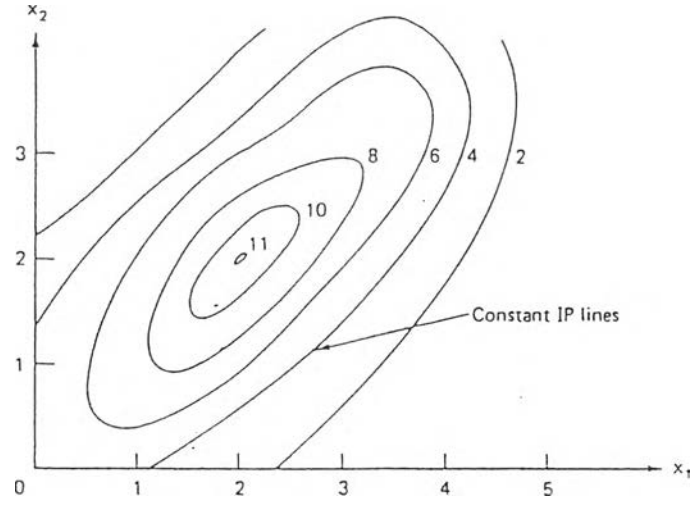

Fig. 3.1 Response surface for IP as a function of X_1 ,and X_2 (6)

where Z = index of performance or dependent variable; and,

x_1 , x_2 = inputs or independent variables on which Z is functionally dependent.

3.2.2 Optimum point

The optimum point on a response surface is the combination of x_1 ,and x_2 values at which IP is optimized. In Fig. 3.1 the optimum (maximum) IP value is slightly greater than 11.0 and its location is approximately x_1 = 2.0, x_2 = 2.0. For a maximization problem, the optimum point is at the peak or summit of the response surface. For minimization problem, the optimum point is located at the deepest point in the valley of the response surface.

3.2.3 Unimodality

Search strategies commonly rely on the assumption that the response surface is unimodal. What this means is that there is only one peak in the response surface. The objective of

the search strategy is to find that single peak. If more than one peak exists, the strategy might seek out a peak that is not the highest one. Hence, the true optimum point would be neglected.

3.2.4 Gradient

Many search strategies are based on the use of gradients. The gradient is a vector quantity whose components are along the axes of the independent variables (x_1 , x_2). The magnitude of each component is equal to the partial derivative of the IP with respect to the corresponding independent variable. This interest will be limited to two independent variables, although the concept of the gradient applies to n-dimensional response surfaces. For two inputs, x_1 and x_2 , the components of the gradient are defined as:

$$G_{1p} = \partial Z / \partial X_1 \tag{3.2}$$

$$G_{2p} = \partial Z / \partial X_2 \tag{3.3}$$

where G_{1p}, G_{2p} = components of the gradient in the x_1 and x_2 directions, respectively, these components must be evaluated at a particular location on the response surface and this location is identified as a point p; and,

Z = index of performance, a function of x_1 , x_2 .

The two components add together to form the gradient

$$G_p = iG_{1p} + jG_{2p} \tag{3.4}$$

where i and j represent unit vectors parallel to the x_1 and x_2 axes. The gradient points in the direction of the steepest slope.

Moving in the direction of steepest slope is a reasonable strategy to reach the top of the response surface. This is why many search strategies are based on the use of gradients. The magnitude of the gradient is a scalar quantity given by:

$$M_p = [(\partial Z / \partial X_1)^2 + (\partial Z / \partial X_1)^2]^{1/2} \quad (3.5)$$

Again, the magnitude of the gradient is defined at a particular point, p, on the $x_1 - x_2$ surface.

The direction of the gradient is a unit vector defined by means of the following equations:

$$D_p = G_p / M_p \quad (3.6)$$

It is sometimes more convenient to work with the direction of the gradient rather than the gradient itself because the length of the direction vector does not vary. Its length is always 1 unit.

The definitions of the gradient, the magnitude, and the direction of the gradient given above can all be extended to response surfaces with more than two independent variables. The visualization is more difficult, but the concepts are identical in multidimensional space.

3.2.5 Trajectory

Whether or not the search strategy makes use of gradients to find its way, the trajectory is the sequence of moves followed by the strategy to seek out the optimum.

3.3 Gradient search strategies

The most familiar gradient search strategy is called the method of steepest ascent. This method begins by estimating the gradient at the current operating point. It then moves the operating point to a new position in the direction of the gradient. The gradient is determined again at the new position in anticipation of the next move toward the optimum. The cycle of gradient determination and step move is repeated until the optimum point is achieved.

3.3.1 Determining the gradient

In a practical problem, the mathematical equation for the response surface is not usually known. Accordingly, the gradient cannot simply be found by employing Eq. 3.2 and 3.3. Instead, the slope of the response surface is determined by making several exploratory moves centered around the current operating point. The exploratory moves are arranged in the form of a factorial experiment. That is, a square of experimental points is established around the current operating point, as illustrated in Fig. 3.2. At each point, the IP is determined. Then, for the minimization, the gradient components are estimated by means of the following equation:

$$G_{1p}=[(Z_1 + Z_4) - (Z_2 + Z_3)]/2\Delta x_1 \qquad (3.7)$$

$$G_{2p}=[(Z_1 + Z_3) - (Z_2 + Z_4)]/2\Delta x_2 \qquad (3.8)$$

where Z_i = values of the performance index at the four experimental points ($i=1,2,3,4$);

Δx_1 = difference in the independent variable x_1 separating the experimental points; and,

Δx_2 = difference in the independent variable x_2 separating the experimental points.

The reason for sequencing the exploratory points 1, 2, 3, and 4 as shown in Fig.3.2 is to reduce the effect of any drift in the process. The values of Δx_1 and Δx_2 must be decided according to two opposing factors. First, Eq. 3.7 and 3.8 approximate the true partial derivatives of Eq. 3.2 and 3.3 more as Δx_1 and Δx_2 become smaller. But if experimental error is present in the measure of Z, the separation between experimental points must be large enough to overcome the effect of the errors. And if Δx_1 and Δx_2 are very different in value, the dimensionless would be applied. Judgement must be used by the analyst in order to decide on the values of Δx_1 and Δx_2 as well as the number of experimental replications.

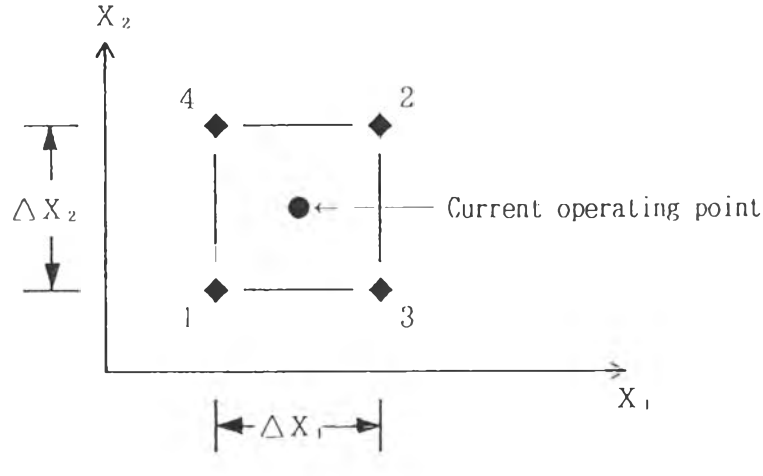

Fig. 3.2 Fortorial experiments to estimate the gradient

3.3.2 Step moves

Exploratory moves are used for the purpose of determining the gradient. Once the gradient has been determined, a step move is made at the new operating point. The step move is taken in the direction of the gradient. The input variable x_1 and x_2 are incremented in proportion to the components of the direction vector as in the following equation:

$$\text{new } x_1 = \text{old } x_1 + SM * G_{1p} / M_p \qquad (3.9)$$

$$\text{new } x_2 = \text{old } x_2 + SM * G_{2p} / M_p \qquad (3.10)$$

where SM is a scalar quantity that determines the size of the step move. The use of the constant SM in Eq. 3.9 and 3.10 means that the length of the step move is the same for every cycle. This occurs in spite of the fact that the gradient components will change in value both relatively and absolutely with every cycle.

3.3.3 Stopping criteria

The search continues until the optimum is reached. At the optimum value of the index of performance, the gradient has a value of zero. It would be sheer coincidence if a step move were to land exactly on the optimum point. A more likely occurrence is for the strategy to overshoot the optimum. When this happens, it can be indentified by the fact that the next gradient changes direction very abruptly, perhaps heading in roughly the opposite direction from previous step moves.

When the vicinity of the optimum is found, it is beneficial to reduce the size of the step move. This is accomplished by reducing the value of the constant SM in Eg. 3.9 and 3.10. In the beginning of the search, a large step size would

be used to speed convergence to the optimum. The final resolution of the optimum must be achieved with smaller step moves. A reasonable criteria for stopping the search is proper when repeated step moves produce no significant improvement in the IP.

3.4 Optimum Gradient Method

There are other gradient search strategies in addition to the method of steepest ascent. A close cousin is the optimum gradient method. The procedure of the optimum gradient method is as follows:

3.4.1 The strategy begins with a determination of the index of performance and its gradient at the starting point.

3.4.2 A step move is made in the direction of the gradient.

3.4.3 The index of performance is determined at the new operating point. If the current index of performance is less than the previous index of performance (in minimization), take another step move in the direction of the previous gradient.

3.4.4 Repeat step 3.4.3 until there is not further improvement in the results of the index of performance. When the current index of performance is greater than the previous index of performance, determine the gradient at the previous point. Go to step 3.4.2.

The advantage of the optimum gradient method is that in most search problems it will not be necessary to make exploratory moves to find the gradient after every move. Since there are time and cost associated with each exploratory move, the optimum gradient method will be less time consuming and less expensive to operate in most search problems.