ADAPTIVE SAMPLED-DATA SYSTEMS – A STATISTICAL THEORY OF ADAPTATION

Bernard Widrow

Department of Electrical Engineering
and
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, Massachusetts

Summary. An adaptive system can be devised by combining two systems, one an adjustable "worker," which receives an input and operates upon it to yield an output, the other a "boss" whose input is a measure of the performance of the worker and whose output is the adjustment of the worker. Performance feedback can be used to achieve automatic system synthesis, that is, the selection of an optimum worker from a predetermined class of possibilities.

In this paper, adaptive sampled-data "Wiener-Lee systems" are described and analyzed. An example is a predictor whose purpose is to adapt in order to be able to predict with minimum mean-square error the next sample of a correlated random input sequence.

Iterative gradient methods used in the adjustment of the impulses of the worker's impulse response can be represented by linear sampled feedback models. Except for measurement noise, adjustment transients are multidimensional geometric progressions.

Small sample size noise causes imperfect system adjustment. The "misadjustment" is shown to be approximately equal to the reciprocal of the number of input samples required to adapt to a step change in process statistics multiplied by the number of interacting adjustments. This result has been verified on an IBM 704 digital computer.

For nonstationary inputs, the choice of adaptation time constant can be shown to be optimum when the misadjustment due to measurement noise is equal to twice the misadjustment due to imperfect process "tracking." This choice can be made adaptive, and would require a higher level of supervision (the boss's boss). The duration of performance averaging that takes place at the higher level is much longer term.

Adaptive systems can adjust to changes in environment and to partial system failure and can "learn" to imitate complex dynamic systems by monitoring their inputs and outputs. The combination of prediction and imitation may permit an accuracy in statistical prediction that has not been heretofore possible.

* * * * * * * * * * *

A fixed "optimum" system design implies that the designer has foreseen all possible input conditions (at least statistically) and, knowing what he would like the system to do under each condition, has chosen the best (according to his criterion) within a class of designs to which he has restricted himself, a priori. An example of this is the use of the Wiener-Lee optimization for the design of linear control systems.[1, 2, 3]

In many instances however, the characteristics of input signals are not known, even statistically. In other cases, the nature of the input might be known to be changeable; for example, input statistics can be nonstationary. An adaptive system that continually searches for the optimum within the allowed class by an orderly trial-and-error process would give vastly superior performance in many of these instances. Sampled-data systems that involve digital computers could be readily made adaptive. A digital filter whose structure is determined by the program of a computer could be modified by a program to change the program. The objective is always to improve performance.

A. Feedback and Trial-and-Error Processes

Iterative or trial-and-error processes are integral parts of adaptive systems. They provide the mechanism of adaptation. It is often convenient to represent such processes as feedback systems; the error of trial and error is analogous to

the "error" of feedback control.

Many of the relaxation and iterative methods commonly employed by numerical analysts appear to be linear feedback systems when represented in this manner. These might be called linear trial-and-error processes. An example of great importance in this discussion is that of surface exploration for stationary points.

Stationary points, or maxima and minima, are characterized by zero partial derivatives with respect to the independent variables. These partial derivatives generally increase with distance from the stationary point and, moreover, increase linearly for a quadratic surface. In many of the commonly used gradient methods,[4] the surface is explored by making changes in the independent variables (starting with an initial guess) in proportion to measured partial derivatives to obtain the next guess, and so forth. These methods give rise to geometric (exponential) decays in the independent variables as they approach a stationary point for second-degree or quadratic surfaces. This is illustrated by the one-dimensional model of Fig. 1.

The "surface" being explored in Fig. 1 is given by Eq. 1. The first and second derivatives are given by Eqs. 2 and 3.

$$y = a(x-b)^2 + c \tag{1}$$

$$\frac{dy}{dx} = 2a(x-b) \tag{2}$$

$$\frac{d^2y}{dx^2} = 2a \tag{3}$$

Let the proportionality constant between change in guess and derivative be -k. This constant could be so chosen that the error in x decreases by one half with each iteration cycle, as illustrated in Fig. 1a. A sampled-data feedback model of the iterative process is shown in Fig. 1b.[5,6,7] The initial numerical guess is injected once at the beginning of the process, whereas the numerical reference or stationary value b is injected synchronously during each cycle. The numerical sequence at the point x(n) begins with the initial guess and proceeds as a sampled transient that relaxes geometrically toward the stationary point, exactly like the sequence of guesses in the surface exploration. For the present, we shall disregard the source of "derivative measurement noise."

The following is an explanation of the feedback model. The next guess is equal to the present guess (this accounts for the unity feedback branch) plus a constant (-k) times the derivative. From Eq. 2, the derivative equals 2a times x(n) minus the constant 2ab. Since the next guess will become the present guess for the next iteration cycle, it is stored by the unit delay (the feedforward branch of transfer function z) to supply the signal at node x(n) at the proper time. It is clear from the flow-graph model that if the iterative process is stable, equilibrium will be reached when x(n) reaches the value b.

The flow-graph can be reduced, and the transfer function from any point to any other point can thus be found. The resulting characteristic equation is

$$(2ak-1)\,z + 1 = 0 \tag{4}$$

The iterative process is stable when $0 < k < \frac{1}{a}$. In order to choose the "loop gain" k to get a specific transient decay rate, one would have to measure the second derivative (2a) at some point on the curve.

Each time a guess in x is to be made, the derivative is physically measured (Fig. 1a), whereas in the model (Fig. 1b) it is obtained as a quantity proportional to x. If the surface were of higher degree than second, the derivative would not be simply proportional to x but would be some polynomial in x. The model could still be made, but it would not be a linear system, and transients would not be geometric. Nevertheless, the iteration process will locate the stationary point. In its vicinity, transients will be geometric because the second and lower degree terms of the Taylor expansion of any continuous surface become the dominating ones. For this and other reasons, exploration of the parabolic surface is given special attention.

The derivatives of a parabola and the partial derivatives of a parabolic surface could be measured in the manner illustrated by Fig. 2. The dimensionless ratio of γ to B is defined as the perturbation P of the measurement.

The first and second derivatives are given by Eqs. 5 and 6. These relations are precise for parabolas, and are approximate for higher degree curves.

$$\left.\frac{dy}{dx}\right|_{x_B} = 1/\delta[1/2C - 1/2A] \tag{5}$$

$$\left.\frac{d^2y}{dx^2}\right|_{x_B} = 1/\delta^2[1/2C-B + 1/2A] \tag{6}$$

Figure 3 shows a two-dimensional paraboloid and a plan view of a sequence of vector changes in the independent variables x_1 and x_2 while a minimum is being sought. Each component of a vector change is a linear combination of the local partial derivatives. The resulting transients are multidimensional geometric progressions.

The surface being searched is given by Eq. 7, the partial derivatives by Eqs. 8, and the second partial derivatives by Eqs. 9.

$$y = ax_1^2 + bx_2^2 + cx_1 + dx_2 + ex_1x_2 + f \tag{7}$$

$$\frac{\partial y}{\partial x_1} = 2ax_1 + c + ex_2 \qquad \frac{\partial y}{\partial x_2} = 2bx_2 + d + ex_1 \tag{8}$$

$$\frac{\partial^2 y}{\partial x_1^2} = 2a \qquad \frac{\partial^2 y}{\partial x_1 \partial x_2} = e \qquad \frac{\partial^2 y}{\partial x_2^2} = 2b \tag{9}$$

A vector flow-graph model of the iterative process is given in Fig. 4. The branches in this graph are capable of carrying two-dimensional samples, indicated by column matrices, and the matrix gains of the branches signify that outputs equal inputs premultiplied by gains. The two-dimensional flow graph is completely analogous to the one-dimensional graph. The feedforward branch is merely a delay with no crosscoupling of the coordinates, and the unit feedback branch is simply that with no crosscoupling. The first partial derivatives are formed, as indicated by Eqs. 8, from the linear combination of the constants c and d and of the x's premultiplied by the matrix of second partials.

This flow graph can be reduced straightforwardly by making use of the rules of matrix algebra. There are as many natural frequencies (decay rates) as there are independent coordinates. The multidimensional loop gain in this case is determined by choice of the matrix of k's.

There are many surface-searching methods in common use. Among these are the method of steepest descent, Newton's method, and the Southwell Relaxation method. These can all be represented by feedback models, like Fig. 4. They differ mainly in the choice of the k's in the feedback matrix.

B. Analysis of an Adaptive Statistical Predictor

The point of view taken in Sec. A on certain trial-and-error processes has been helpful in analyzing the behavior of adaptive systems. Consider the general linear sampled-data system formed of a tapped delay line, shown in Fig. 5. This system is intended to be a statistical predictor. The present output sample g(n) is a linear combination of present and past input samples. The constants in this combination are h_o, h_1, h_2, etc., the predictor impulse-response samples, or the gains associated with the delay-line taps. Their choice constitutes the predictor design. They may be adjusted in the following manner. Apply a mean-square reading meter to $\epsilon(n)$, the difference between the present input and the delayed prediction. This meter will measure mean-square error in prediction. Adjust h_o, h_1, h_2, . . . , until the meter reading is minimized.

The problem of adjusting the h's is not trivial because their effects upon performance interact. Suppose that the predictor has only two impulses in its impulse response, h_o and h_1. The mean-square error for any setting of h_o and h_1 can be readily derived:

$$\epsilon(n) = f(n) - h_o f(n-1) - h_1 f(n-2)$$

$$\epsilon^2(n) = f^2(n) + h_o^2 f^2(n-1) + h_1^2 f^2(n-2) - 2h_o f(n) f(n-1) - 2h_1 f(n) f(n-2) + 2h_o h_1 f(n-1) f(n-2) \quad (10)$$

$$\overline{\epsilon^2}(n) = \phi_{ff}(0) h_o^2 + \phi_{ff}(0) h_1^2 - 2\phi_{ff}(1) h_o - 2\phi_{ff}(2) h_1 + 2\phi_{ff}(1) h_o h_1 + \phi_{ff}(0)$$

The discrete autocorrelation function of the input is $\phi_{ff}(k)$.

The mean-square error of Eqs. 10 is what the mean-square meter would read if it were to average over very large sample size. The mean-square error is a parabolic function of the predictor adjustments h_o and h_1, and in general can easily be shown to be a quadratic function of such adjustments, regardless of how many there are. Any of the Wiener-Lee problems such as linear prediction, interpolation, or noise filtering give quadratic functions. If a mean-fourth error criterion were used, this would be a fourth-degree function.

The optimum m-impulse predictor can be derived analytically by setting the partial derivatives of $\overline{\epsilon^2}$ of Eq. 10 equal to zero. Finding the optimum system experimentally is the same as finding a minimum of a paraboloid in m dimensions. This could be done manually by having a human operator read the meter and set the adjustments, or it could be done automatically by making use of the iterative gradient methods for surface-searching, as described in the previous section. When either of these schemes is employed, an adaptive system results that consists essentially of a "worker" and a "boss." The worker in this case predicts, whereas the boss has the job of adjusting the worker.

Figure 6 is a block-diagram representation of such a basic adaptive unit. The boss continually seeks a better worker and sees the combination of random process, worker, and desired objective of prediction with minimum mean-square error as a multidimensional transducer that connects adjustment to performance. Adaptation is a multidimensional feedback process. The "error" signal is the gradient of mean-square error with respect to adjustment.

Noise enters the adaptation feedback system because the input process cannot be continued indefinitely for each measurement of mean-square error (A, B, C of Fig. 2, needed for gradient measurement), and thereby places a basic limitation upon adaptability. It will be shown that the slower the adaptation, the more precise it is. The faster the adaptation, the more noisy (and poor) are the system adjustments.

The model considered first in deriving a measure of the quality of adaptation versus the speed of adaptation will have only one adjustment. The theory will be generalized to multiple dimensions later. A plot of mean-square error versus h_o for the simplest system having only one impulse in its impulse response would be a parabola, analogous to the parabola of Fig. 1. During each cycle of adjustment, the derivative of $y = \overline{\epsilon^2}$ with respect to $x = h_o$ would have to be measured according to the scheme of Fig. 2.

Equation 5 gives the derivative as the difference of "forward" and "backward" measured values of y multiplied by $(1/2\delta)$. The variance in derivative errors is equal to the sum of the variances in the forward and backward measurements of $y = \overline{\epsilon^2}$ multiplied by $(1/4\delta^2)$, under the assumption that they are essentially independent. "Noisy" measurements of y due to finite sample size cause noisy derivative measurements. These in turn cause noisy settings of $x = h_o$ and losses in system performance. The over-all objective is to find a relation between the loss in average performance and the sample size used in measurement of A and C of Fig. 2.

Set h_o to some fixed value. The corresponding mean-square error $\overline{\epsilon^2}$, a long-term average, can be estimated by taking the square of a single error sample. The variance in such a measurement is

$$\overline{(\epsilon^2 - \overline{\epsilon^2})^2} = (\overline{\epsilon^2})^2 + \overline{\epsilon^4} - 2(\overline{\epsilon^2})^2 = \overline{\epsilon^4} - (\overline{\epsilon^2})^2 \quad (11)$$

If ϵ were Gaussian-distributed with zero mean, the mean fourth minus the square of the mean

square would equal 2 times the square of the mean square, whereas if ϵ were flat-top distributed with zero mean, this would equal 0.8 times the square of the mean square. An average component added to ϵ causes the coefficient to diminish only slightly. No value of this coefficient greater than 2 has been found, and it could range from 2 to 0. A reasonable all-round value is taken as 4/3. Since the variance in an average of N independent measurements is 1/N times that of a single one, the variance in a point on the $\overline{\epsilon^2}$ curve taken from N measurements is approximately $\frac{4}{3}\frac{1}{N}(\overline{\epsilon^2})^2$. The effective sample size is a fraction of N when measurements are not independent. It can be shown however that this fraction is close to unity even when the $\epsilon(n)$'s are highly correlated, as high as 90 per cent, for example. It follows that the variance in the measurement of a first derivative (as in Fig. 2) from N forward and N backward samples is given by

$$\left(\frac{1}{4\delta^2}\right)\frac{4}{3N}(A^2+C^2) \tag{12}$$

It can be shown that when the perturbation P is less than 20 per cent, as it will usually be, the following approximation will be essentially an equality:

$$A^2 + C^2 \approx 2B^2 \tag{13}$$

The perturbation has been defined to be $P = (\gamma/B)$. It can be shown that γ is precisely equal to $(a\delta^2)$. Combination of all these allows expression (12) to be replaced by

$$\begin{pmatrix}\text{variance in}\\ \text{derivative}\\ \text{measurement}\end{pmatrix} = \frac{2B^2}{3N\delta^2} \tag{14}$$

During each iteration cycle, a "noise" in derivative measurement occurs, the variance of which is given by (14). These noises are almost statistically independent from cycle to cycle. Assume that B^2 is relatively constant and approximately equal to c^2. This will be essentially the case in the vicinity of the optimum. To find the effect of these noises upon the adjustment x, consider the flow graph of Fig. 1b. The variance in x is equal to the variance in derivative noise multiplied by the sum of the squares of the impulse values of the unit impulse response from the derivative point to x(n). The sum of squares is conveniently expressed in terms of the "time constant" τ of the flow graph rather than in terms of the feedback constant k. A unit value of τ means that the transients decay by a factor of (1/e) with each iteration cycle. The sum of squares turn out to be very close to $(1/8a^2\tau)$, with the result that the variance in x is

$$\frac{1}{12(N\tau)}\left(\frac{c^2}{a^2\delta^2}\right) \tag{15}$$

In steady state after transients have disappeared, x will vary randomly about the optimum point from iteration cycle to iteration cycle. Because x is not always at optimum, y on the average will be greater than c. The "misadjustment" M is defined as

$$M = \frac{\overline{y} - c}{c} \tag{16}$$

The misadjustment is a very useful parameter, being the dimensionless ratio of the mean increase in mean-square error to the minimum mean-square error. It is a measure of how the adaptive system performs on the average, after adapting transients have died out, compared with the fixed optimum Wiener-Lee system.

Consideration of Eq. 1 shows that the increase in $\overline{y}$ is equal to the variance in x multiplied by a. If use is made of Eq. 15, the misadjustment can be expressed as

$$M = \frac{a}{c}\frac{1}{12(N\tau)}\left(\frac{c^2}{a^2\delta^2}\right) \tag{17}$$

Recall that $P = a\delta^2/c$ in the vicinity of the optimum. Eq. 17 becomes

$$M = \frac{1}{12(N\tau)P} \tag{18}$$

The $(N\tau)$ product is related to the total number of samples "seen" by the system in adapting to a transient in input process statistics. Notice that a given effect could be achieved by using many samples per cycle (large N) and few cycles to adapt (small τ), or by using few samples per cycle

and proceeding toward the optimum with small steps (large τ). The important quantity is the $N\tau$ product. The misadjustment is always inversely proportional to it. $2N\tau$ can be represented by the symbol Γ. This is the "adaptation time constant," that is, the number of process samples that elapse in one time constant of adaptation. Equation 18 can be rewritten as follows; notice that all symbols are dimensionless:

$$M = \frac{1}{6\Gamma P} \tag{19}$$

Another point of view on finding the derivative at a point on the mean-square-error curve postulates a "small-sample size" mean-square-error curve. This entire curve could be investigated point by point by "playing" the same small piece of input record over and over again to the system, each time with different adjustment of h_o. The derivative of this curve can be obtained precisely by using the same data for the forward and backward measurements of mean-square error. Perturbation amplitude would have no effect. The small sample size curve is actually a property of the piece of input record.

It can be shown that the variance in the measurement of the derivative using N samples for both forward and backward measurements in the vicinity of the optimum is

$$8\,ac/3N \tag{20}$$

Expression 20 is by no means obvious and has been derived by first finding the variance in the position of the minimum point of the small sample size curve. This turns out to be the quantity $(2c/3a)$. Since the derivative is equal to $2ax$, the variance in derivative near $x = 0$ is $4a^2$ times the variance in x, which is the same as the variance in minimum-point position. From this we obtain expression 20. The variance in x is

$$\left(\frac{8ac}{3N}\right)\left(\frac{1}{8a^2\tau}\right) = \frac{c}{3aN\tau} \tag{21}$$

The average increase in y is

$$c/3N\tau \tag{22}$$

The misadjustment is therefore

$$M = \frac{1}{3N\tau} = \frac{1}{3\Gamma} \tag{23}$$

In this case, the adaptation time constant Γ or the number of samples that elapse per time constant is $(N\tau)$.

There is often a premium on operating with small perturbation. If the adaptive system is connected "in line" with an actual process, a component of misadjustment exactly equal to the perturbation must be added to the appropriate misadjustment expression. Setting an adjustment forward for some time and then backward for some time always makes performance worse, on the average, by the amount γ than if the adjustment had remained at center all the while. If the adaptation does not take place in real time or if another identical system not "in line" is available for experimentation, advantage could be taken of data-repeating or the use of large perturbations. There is another way that the derivatives could be measured,[8] which is equivalent to data-repeating. This method requires the measurement of the input autocorrelation function. Differentiation of expression 10 shows how the derivatives are related to the input correlations. The auxiliary system in this case is a correlator, and the technique can be used only where there are simple relations between derivatives and correlations.

These ideas can be extended to multidimensional adaptation. The most efficient adaptation scheme has the partial derivatives measured along the major and minor elliptical axes. These directions would be at an angle of 45° with the original coordinate axes because coefficients in the mean-square expression (Eq. 10 for example) associated with the squares of the h's are always of equal magnitude for any of the Wiener-Lee problems. If partial derivatives are measured along these directions, and the vector changes were to have components along these directions proportional to the respective partial derivatives, transients and measurement noise propagations along these directions would be isolated. The branch in the

flow graph of Fig. 4 connecting the x(n) matrix with the partial-derivative matrix would have a gain matrix with only diagonal elements. The feedback-gain matrix would likewise have only diagonal elements.

The time constants for the two variables in the flow graph could be chosen separately, but a reasonable procedure is to make them the same. It is also reasonable to make N and P the same for both directions. The increase in mean-square error due to variance in adjustment along one of the major axes plus that due to variance in adjustment along the other major axis equals the total increase in mean-square error. Misadjustments add. The previous expressions for misadjustment are generalized by multiplying them by the number of adjustments, m.

If the orthogonalization scheme is not used, it is quite difficult to predict closely the misadjustment for a given situation. "Ball-park" answers can surely be obtained from the same misadjustment expressions by taking the τ of the flow graph to be an average over its various modes. A few generalizations can be made however. Doubling N halves M, and halving the over-all adaptation rate halves M. To achieve a given M, the number of process samples that must elapse before the system will adapt increases with the square of m when different data for each performance measurement are used, and increases in direct proportion to m if data-repeating is used.

C. Simulation of Adaptive Sampled-Data Systems

Relations 19 and 23 are very general and apply regardless of the nature of the input processes and the goals of adaptation. Several assumptions were made in their derivation, however, and in order to verify the reasonableness of the assumptions, an extensive series of simulation studies was undertaken with the aid of an IBM 704 digital computer. The results of these experiments have shown that the measured misadjustments usually fall within 10 per cent of their predicted values, and rarely differ by as much as 20 per cent.

The results of one of these simulations is shown in Fig. 7. The adaptive system in this case was a predictor. Its sampled input was a stationary Markov process generated by injecting a random sequence into a sampled-data system of transfer function $1/(1-1/2z)$. The adaptation process was orthogonalized, and the initial guess of system structure was $h_o = h_1 = 0$. The optimum impulse response point and contours of constant mean-square error are shown in the figure.

Ten samples per iteration cycle were used in the data-repeating scheme. The adaptation rate was set so that transients along each coordinate diminished by a factor of 1/2 during each cycle, so that $\tau = 1.45$, and therefore $\Gamma = 14.5$. Since $m = 2$, the theoretical misadjustment is

$$M = \frac{2}{3\Gamma} = \frac{2}{43.5} = 4.6 \text{ per cent}$$

It can be seen from Fig. 7 that within about four cycles, the adjustment transient had essentially died out (it theoretically was down to $1/2^4$ of its original value). The system adjustments take on a "random walk" about the optimum point. The experimental misadjustment has been calculated by subtracting the minimum mean-square error from the average mean-square error of iterations 4 through 10 (this allows the initial transient to disappear) and then dividing by the minimum mean-square error. This turns out to be 20 per cent. This value, however, is only 5.4 per cent if points 9 and 10 in Fig. 7 are ignored. The statistical phenomenon giving rise to the large excursion in point 9 is a rare one, although quite normal. By the tenth cycle, the adaptation system has almost recovered. If averages were taken over more cycles, there would be very close agreement between the theoretical and measured values of M.

D. Application of Adaptive Filters to Nonstationary Signals

The adaptive predictor considered in the previous section is able to adapt to a new process environment within approximately four adaptation

cycles, which is less than three adaptation time constants. During this time, 40 input samples are processed by the system, and M = 4.6 per cent. When a misadjustment of approximately 10 per cent is tolerable, a two-impulse filter could be made to adapt to a major change in input process statistics after seeing about 20 samples. A ten-impulse filter would require 200 input samples to adapt and would have a steady-state misadjustment of 10 per cent. This degree of complexity is all that would be required for most practical applications.

Speed of adaptation can always be acquired at the expense of increase in misadjustment. Adaptation speed is controlled by N and τ. Fast adaptation is highly desirable when the input process statistics change rapidly. When process changes are slow, slow adaptation has the advantage of small misadjustment.

Nonstationariness of input processes adds another component to expression 23 because of imperfect process "tracking"; the larger the adaptation time constant, the greater the average lag between the system adjustments and the instantaneously optimum adjustments. It can be shown[9] that the misadjustment for nonstationary inputs is

$$M = \frac{\alpha}{\Gamma} + \beta\Gamma^2 \qquad (24)$$

The constants α and β are properties of the input process. If the derivative of M is set to zero,

$$\frac{dM}{d\Gamma} = \frac{-\alpha}{\Gamma^2} + 2\beta\Gamma = 0 \qquad (25)$$

The condition for optimum choice of the adaptation time constant Γ is therefore

$$\frac{\alpha}{\Gamma} = 2\beta\Gamma^2 \qquad (26)$$

This important result means that the adaptation time constant is optimized when the misadjustment component resulting from small sample size "noise" equals twice that due to poor process tracking. The same result can be generalized for any number of adjustment coordinates.

When the nature of the nonstationariness is known, it is possible (but not simple) to calculate the optimum Γ. If an adaptation process is to be designed to realize a certain Γ, second partial derivatives must be measured at the same time as the first partial derivatives. The former change quite slowly. It is feasible to maintain Γ quite close to any desired value by continually updating the adaptation loop gain, and this has been demonstrated experimentally.

An alternative that would require much less knowledge of the input process and no elaborate calculation of the value of Γ is to achieve self-optimization of Γ. This can be done by experimentally varying Γ, much as the impulses in the impulse response are varied, with the objective of selecting the adaptation rate that optimizes long-term performance. A block diagram of such a system, an adaptive predictor in this case, is shown in Fig. 8. The box W is the worker, B_1 is the boss, and B_2 is the boss's boss. The duration of performance-averaging that takes place at the level of B_2 (over many variations of the nonstationary input process) is much longer term than at B_1 (over a fraction of a variation).

E. Conclusions

We have described and evaluated an adaptive sampled-data system model that is quasistatically linear and makes use of performance feedback for self-optimization. The goal is the minimization of a mean-square error.

There are two main reasons why this model of adaptation is both important and interesting. First, it provides a solution to many practical problems in the control systems and communication theory areas; and second, it obeys simple mathematical laws and seems to have behavioral characteristics in common with other kinds of evolving systems. An example of the latter is the living animal whose structure adapts to "optimize" existence in its environment. Another example is the policy of an industrial firm which, in order to maximize profits, evolves with changes in economic conditions, with technological advances, and with competition. Performance feedback is akin to what evolutionists call "natural selection,"[10]

except that in nature, structural perturbations are random in frequency and amplitude.

Adaptation makes possible control and communications systems that perform almost as well with nonstationary inputs as fixed systems do with stationary inputs. Adaptive systems can be designed to control highly nonlinear processes by adjusting to changes in their operating regions. A new, unique area for the application of adaptive systems is that of imitation of unknown complex dynamic systems by observation of their inputs and outputs.

The adaptive system in Fig. 9 is connected to combine imitation and prediction. It can be shown that $\overline{\epsilon^2}$ is a parabolic function of the h's, if the unknown system is linear, and that the model applies exactly. Because of normal lags in a dynamic process, "early warning" information is present in the input signal, and could be used in prediction if the characteristics of the complex system were understood. The adaptive system can imitate these characteristics. This principle might be applied to weather prediction, for example. Assume that early warning information for a one-day forecast for a certain city is contained in a multitude of meteorologic measurements taken within a 500-mile radius. By observation of current and previous weather records, the predicting system could "learn" to imitate (approximately) the weather dynamics and to use this in prediction.

Closed-loop adaptation that makes use of performance feedback permits direct, automatic system synthesis. It has the advantage of being usable where no analytic synthesis procedure is known; for example, where error criteria other than mean-square are used and where systems are quasistatically nonlinear. In the event of a partial system failure, an adaptation system that continually monitors performance will optimize this performance by adjusting the intact parts. System reliability is thereby improved.

Acknowledgment

I would like to thank Professor P. M. Lewis II, and Professor R. A. Howard for many stimulating and enjoyable discussions of this and related subjects. I would also like to thank my students, R. D. Buzzard, L. Maisel, D. F. DeLong, R. L. Mattson, and R. R. Brown, who worked with me in this field during the last two years.

This work was supported in part by the U.S. Navy (Bureau of Ships) under Contract NObsr-72716; and in part by the U.S. Army (Signal Corps), the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U.S. Navy (Office of Naval Research).

* * * * * * * * * * *

Professor Widrow is now with the Department of Electrical Engineering, Stanford University, Stanford, California.

References

1. N. Wiener, "Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications," John Wiley and Sons, Inc., New York; 1949.
2. Y. W. Lee, "Application of statistical methods to communication problems," Technical Report 181, Research Laboratory of Electronics, M.I.T.; September 1, 1950.
3. G. C. Newton, L. A. Gould, and J. F. Kaiser, "Analytical Design of Linear Feedback Controls," John Wiley and Sons, Inc., New York; 1949.
4. R. R. Brown, "Gradient methods for the computer solution of system optimization problems," WADC Technical Note 57-159, Servomechanisms Research Laboratory, M.I.T.; September, 1957.
5. W. K. Linvill, R. W. Sittler, and B. Widrow, "Pulse-Data Systems," Course 6.54 Notes, Department of Electrical Engineering, M.I.T.; 1959.
6. J. R. Ragazzini and G. F. Franklin, "Sampled-Data Control System," McGraw-Hill Publishing Company, New York; 1958.
7. S. J. Mason, "Feedback theory. 1. Some properties of signal flow graphs," Technical Report 153, Research Laboratory of Electronics, M.I.T.; February 2, 1953. "Feedback theory – Further properties of signal flow graphs," Technical Report 303, Research Laboratory of Electronics, M.I.T.; July 20, 1955.
8. D. F. DeLong, "Analysis of an adaptive sampled-data system", S.M. Thesis, Department of Electrical Engineering, M.I.T., January 1959.
9. B. Widrow, "Adaptive sampled-data systems," Quarterly Progress Report No. 6, Computer Components and Systems Group, M.I.T.; 30 April 1959, pp. 17-21.
10. G. G. Simpson, "The Meaning of Evolution," Mentor Books, New York; 1951.

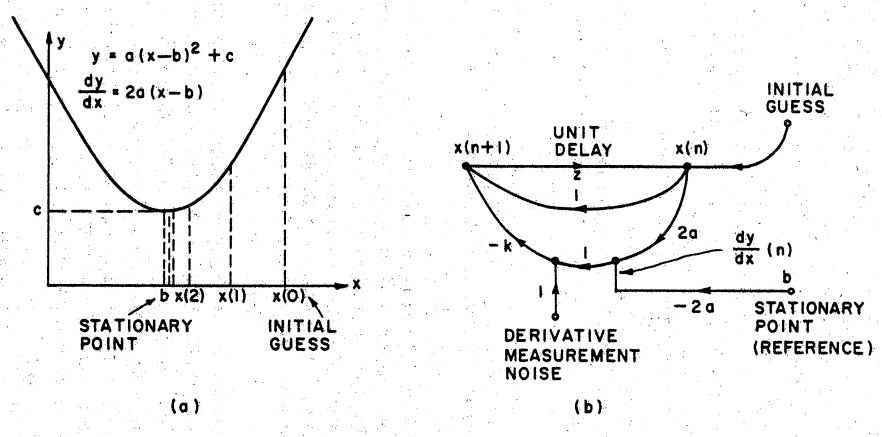


Fig. 1. One-dimensional surface exploration.

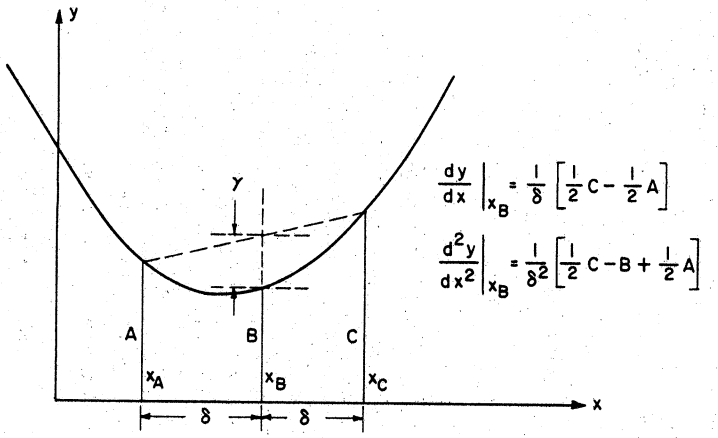
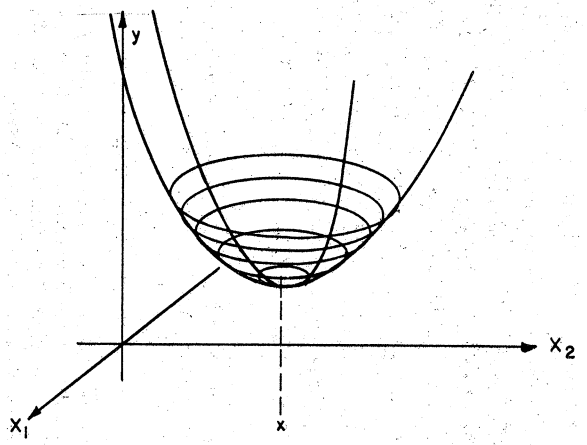


Fig. 2. Measurement of first and second derivatives of a parabola.



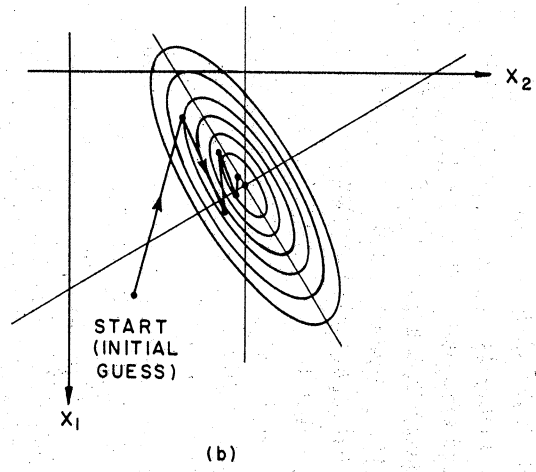


Fig. 3. Two-dimensional surface searching.

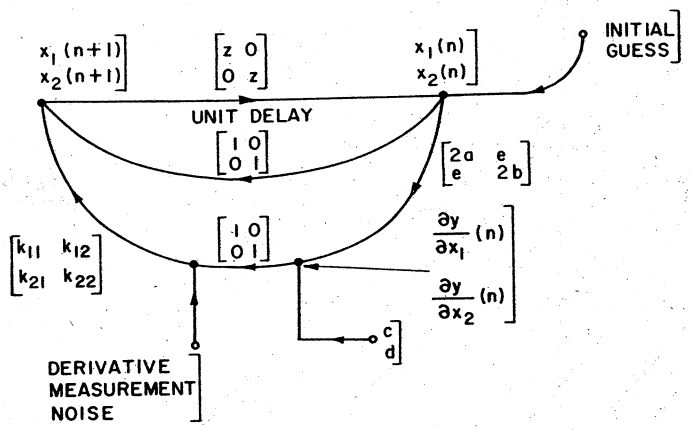


Fig. 4. Feedback model of two-dimensional surface searching.

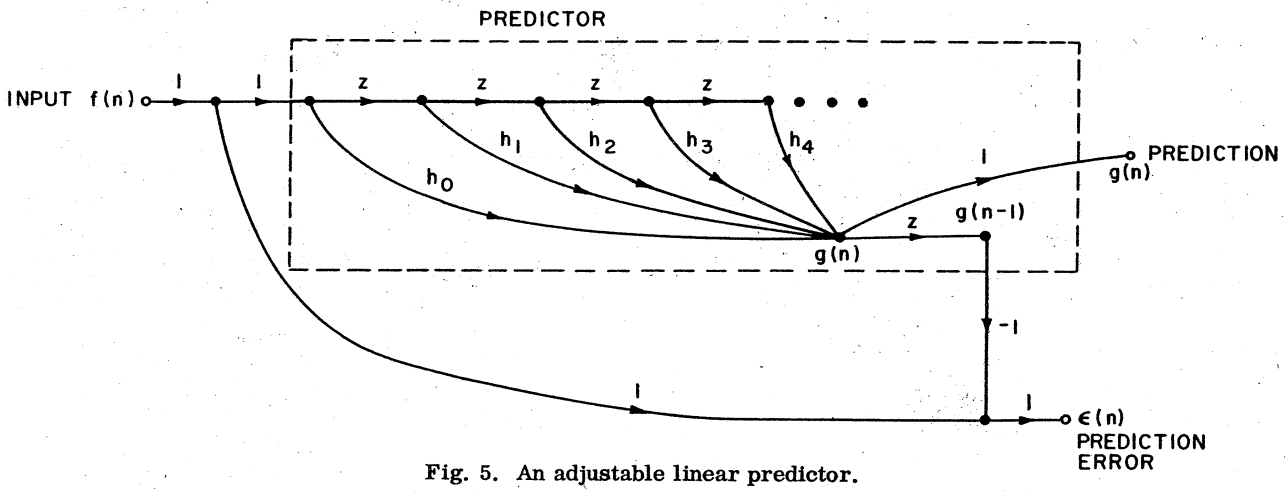


Fig. 5. An adjustable linear predictor.

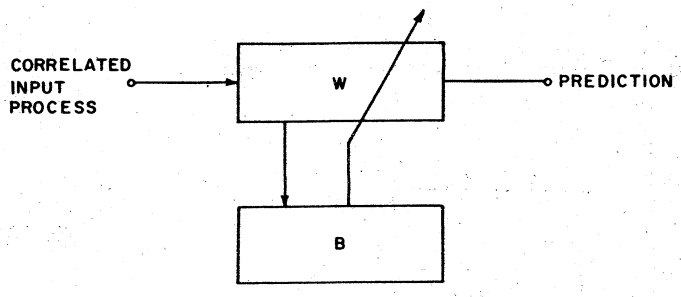
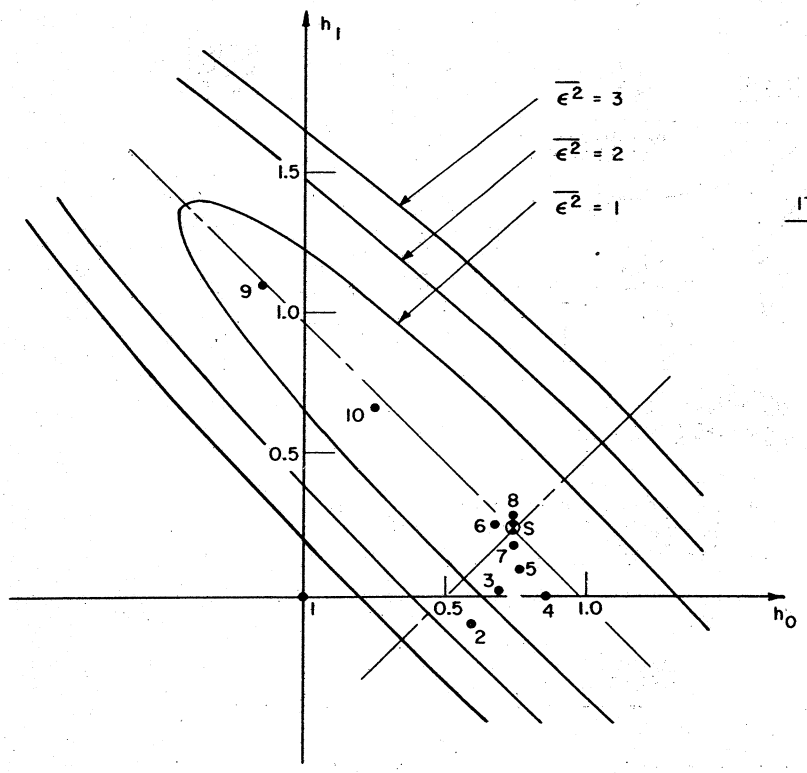


Fig. 6. An elementary adaptive predictor.



ITERATION	$\overline{\epsilon^2}$
1	4.43
2	1.43
3	0.686
4	0.473
5	0.440
6	0.417
7	0.413
8	0.417
9	0.764
10	0.523

$\epsilon^2_{min} = 0.410$

Fig. 7. Adaptation steps of a simulated predictor.

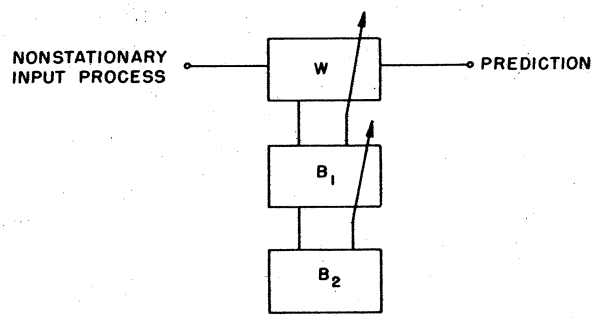


Fig. 8. An adaptive predictor with two levels of adaptation.

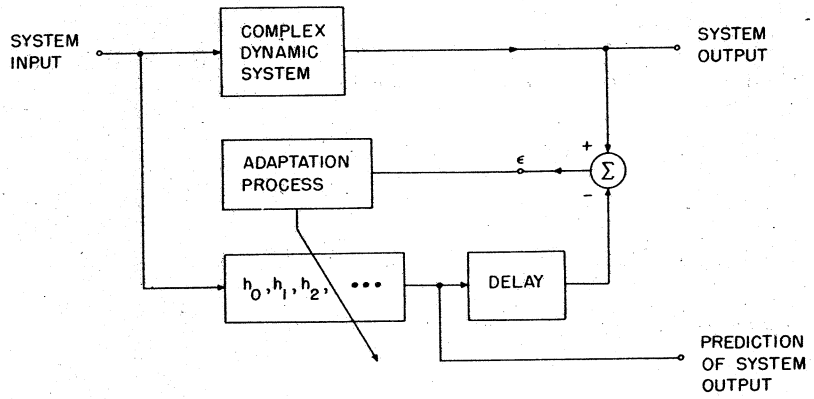


Fig. 9. Adaptive imitation and prediction.

Reprinted from:
1959 IRE WESCON CONVENTION RECORD
Part 4