JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS 5, 150-167 (1962)

A Versatile Stochastic Model of a Function of Unknown and Time Varying Form

HAROLD J. KUSHNER

Massachusetts Institute of Technology,
Lincoln Laboratories, Lexington 73, Massachusetts

Submitted by Lotfi Zadeh

Properties of a random walk model of an unknown function are studied. The model is suitable for use in the following (among others) problem. Given a system with a performance function of unknown, time varying, and possibly multipeak form (with respect to a single system parameter), and given that the only information available are noise perturbed samples of the function at selected parameter settings, then determine the successive parameter settings such that the sum of the values of the observations is maximum. An attempt to avoid the optimal search problem through the use of several intuitively reasonable heuristics is presented.

INTRODUCTION

Problems of maximizing some function of the observations on a curve X_t of unknown form, where the only available information is noise disturbed samples at various t, are of great current interest in technology. Two (among many) examples are the determination of the location of the absolute maximum of X_t and maximizing the expected value of the sum of N (noisy) observations on X_t. These problems may arise (for example) in the design of adaptive control systems or in the optimization of the performance of existing systems. In these cases X_t will be the systems average performance function and t the variable parameter.

The unknown function is generally assumed to be imbedded in a large class of functions. The properties of this (model) class are determined by the (generally analytic) restrictions placed upon the unknown function. The observations are generally taken sequentially and information (in reference to the model) from past observations is used to determine a suitable location for the next observation.

There are four important ways in which assumptions that are generally made on the model may not correspond to the physical situation.

1. The unknown curve may have more than one local maximum.
2. The form of the unknown curve may vary with time.
3. There may be regions which are relatively flat (the plateau problem).
4. The unknown curve may not be continuous or differentiable.

Methods of maximum locating by gradient estimation cannot be used in any of these cases. Conditions 1 and 3 require a search procedure that obtains information on all parts of the curve (global procedure). It seems that the curve model must either be very complicated logically or include provisions for estimation of curve values at points other than those directly observed.

An approach that we have found promising as well as interesting is to avoid analytic restrictions as much as possible and imbed the unknown regression function in a family of curves that have been generated by a suitable stochastic process. The admissable functions are then points (ω) in a particular sample space (Ω). The ω value is determined by and determines the particular function. The procedure is as follows. Observations are taken (sequentially) at selected points. The estimated curve and its variance (or the expected à posteriori value of an observation at any point and the uncertainty of the estimate) is then computed in accordance with the model. (In engineering terms we perform a filtering and prediction operation.) From this (statistical) information we determine the location of the next sample point (using a Bayesian or other suitable criterion).

It has been found most convenient, from the points of view of intuitive relatedness to the physical functions, computational simplicity and ease of use of the information to assume that the (model) X_t is a process of independent, infinitely divisible Gaussian increments.

In Section I we discuss the important properties of the model and derive the forms of the expected X_t and its variance conditioned upon the values of the (noisy) observations. Since the recurrent computation of the curve mean and variance requires the use of all past observations, it may rapidly get out of hand. We have therefore devoted section II to some techniques that substantially simplify the computation.

If we have control over the location of the sample points (as in the two problems of the first paragraph) an optimum sampling procedure is desired. These are, for most useful criteria of optimality, still unobtainable. In Section III we advance two search policies that are especially suitable for use with the form in which the curve information is presented and whose "usefulness" is justified on heuristic grounds. It is felt, because some of the mathematical problems remain intractable, that a simulation study is required for more thorough justification. Because of this, we rest (in Section III) with a presentation of ideas. A more complete analysis will be undertaken in the report on the results of the simulations.

Sections IV and V contain some useful generalizations; e.g. the form of the

model for a time varying X_t. Section VI contains several miscellaneous remarks and extensions.

The use of the model has been emphasized for the case of noise biased observations. It is suitable (and much simpler to use) when we have an unknown, multipeak, time varying curve and observations that are unbiased by noise.

I. The Curve Model

The important properties of the model will be reviewed in this section. If X_t is the value of the unknown function at parameter value t ($t, t_i \in T$, a closed bounded but not necessarily connected set on the real line), then

$$X_t = X_{t_1} + \xi(t_1, t)$$

$$\xi(t_1, t) \sim N(0, c \mid t - t_1 \mid) \tag{1.1}$$

If the intervals $[t_1, t_2]$ and $[t_3, t_4]$ are not disjoint, then

$$\begin{aligned} \mathscr{E}(X_{t_1} - X_{t_2})(X_{t_3} - X_{t_4}) &= \mathscr{E}\,\xi(t_1, t_2)\,\xi(t_3, t_4) \\ &= c \mid [t_1, t_2] \cap [t_3, t_4] \mid \end{aligned} \tag{1.2}$$

where $[t_1, t_2] \cap [t_3, t_4]$ is the length of the overlap. Letting $Y_t = X_t + \eta_t$ be an observation taken at t with noise $\eta_t \sim N(0, \sigma_t^2)$ we have

$$\mathscr{E}(X_t - Y_{t_1})^2 = c \mid t - t_1 \mid + \sigma_{t_1}^2. \tag{1.3}$$

Henceforth, it will be assumed that the observation noises are independent of each other and of the X_t. We will determine the expectation and variance of X_t conditioned on the observations Y_{t_i}, $i = 1, \cdots, n$, where $t_{i-1} \leq t_i$. The subscript i will replace t_i when a specific t point is referred to. The symbol t will refer to a generic parameter point.

Since X_t is a process of independent increments, it is well defined only in terms of differences. It will be convenient to deal with it in terms of the increments $\Delta X_t = X_t - X_0$, $\Delta Y_i = Y_i - X_0$. (X_0 is at present an arbitrary observation at $t_0 \leq t_1$.) Henceforth, since we will be attempting to predict the curve form from the observations Y_i, $i = 1, n$, we will be interested in the quantities

$$\overline{\Delta X}_t = \mathscr{E}(\Delta X_t \mid \Delta Y_1, \cdots, \Delta Y_n)$$

$$\text{Var}\ \Delta X_t = \text{Var}\,(\Delta X_t \mid \Delta Y_1, \cdots, \Delta Y_n)$$

The entire collection ΔX_t, ΔY_i has a joint normal distribution. The cova-

riance matrix is Eq. (1.4). ΔX_t corresponds to the zeroth row and column and ΔY_i to row and column i. Let $t_k \leq t \leq t_{k+1}$, $r_{ij} = | t_i - t_j |$, $t_i \leq t_{i+1}$, and $r = t - t_k$.

$$\sum = \begin{Vmatrix} cr_{0t} & cr_{01} & cr_{02} \cdots cr_{0k} & cr_{0k} + cr & \cdots & cr_{0k} + cr \\ cr_{01} & cr_{01} + \sigma_1^2 & cr_{01} \cdots \vdots & & & cr_{01} \\ cr_{0k} & cr_{01} & cr_{02} \cdots cr_{0k} + \sigma_k^2 & cr_{0k} & \cdots & cr_{0k} \\ cr_{0k} + cr & & \vdots & & & \\ cr_{0k} + cr & cr_{01} & cr_{02} \cdots cr_{0k} & \cdots & cr_{0,n-1} & cr_{0n} + \sigma_n^2 \end{Vmatrix} \tag{1.4}$$

Letting Λ be the determinant of Σ and Λ_{ij} the cofactor of the ith row and jth column of Σ we have [1]

$$\overline{\Delta X_t} = -\sum_1^n \frac{\Lambda_{0i}}{\Lambda_{00}} \Delta Y_i = \sum_1^n A_{ti} \Delta Y_i \tag{1.5}$$

$$\text{Var } \Delta X_t = \Lambda / \Lambda_{00}. \tag{1.6}$$

Equations (1.4)-(1.6) yield all the necessary information; however, since it will be more convenient to work directly in terms of X_t and Y_t rather than in terms of ΔX_t and ΔY_t we will proceed with the computations in an indirect way. If the assumption $X_0 = 0$ is made, then $\Delta X_t = X_t$ and $\Delta Y_i = Y_i$. The restrictiveness of this assumption disappears, however, as $t_0 \to -\infty$. This suggests our procedure; perform the computations, then take limits. (This procedure helps simplify our computations, as will be seen later.)

Several simple properties of $\bar{X}_t$ and Var X_t will now be demonstrated; first, that $\bar{X}_t$ is linear in $t - t_i$ in the interval $[t_i, t_{i+1}]$ and equals $\bar{X}_n$ for $t \geq t_n$. $\bar{X}_t$ is thus a piecewise linear estimate of X_t, and second, that Var X_t is quadratic in $t - t_i$ in the interval $[t_i, t_{i+1}]$ and equals Var $X_n + c(t - t_n)$ for $t \geq t_n$. These results are true for finite or infinite t_0. When $t_0 = -\infty$, the results for $t \geq t_n$ have a counterpart for $t \leq t_1$.

The linearity property of $\bar{X}_t$ can be seen by recalling 1.5 and noticing (in 1.4) that Λ_{0i} is linear in r. For $t \geq t_n$, Λ_{0i} is constant, thus $\bar{X}_t = \bar{X}_n$. For $t_1 \leq t \leq t_n$, Λ is quadratic in r, thus Var X_t is quadratic in this region. With $t \geq t_n$, r appears only in the zero-zeroth entry of Σ. Upon expanding about this entry we obtain

$$\text{Var } X_t = \frac{\Lambda}{\Lambda_{00}} (r = 0) + cr \frac{\Lambda_{00}}{\Lambda_{00}} = \text{Var } \bar{X}_n + c \mid t - t_n \mid. \tag{1.7}$$

These simple properties will prove to be of great help in the computations of $\bar{X}_t$ and Var X_t and in the subsequent use of these quantities for the determination of the next sample point and, in addition, for the ease of application of intuition to the results. We have this simplicity because the curve can be studied interval $[t_i, t_{i+1}]$ by interval and within each interval $\bar{X}_t$ is linear and Var $\bar{X}_t$ is quadratic.

When $t_0 = -\infty$, the computations of the ratios Λ/Λ_{00}, $\Lambda_{0i}/\Lambda_{00}$ are handled as follows. Subtract some row (say the first) of the matrices Σ and Σ_{0i} (the matrix obtained from Σ by deleting the 0th row and ith column) from all other rows, divide all entries in this row by r_{01}, take the ratio of the appropriate determinants and let $r_{01} \to \infty$. For example when $t_k \leq t \leq t_{k+1}$ we have

$$\frac{\Lambda}{\Lambda_{00}} = \frac{r_{01} \cdot \det \begin{Vmatrix} cr_{1t} & -\sigma_1^2 & cr_{12} & \cdots & cr_{1k} & cr_{1t} & \cdots & cr_{1t} \\ 1 & \dfrac{cr_{01}+\sigma_1^2}{cr_{01}} & 1 & & & & & 1 \\ cr_{1k} & & & & & & & \\ cr_{1t} & & & & & & & \\ cr_{1t} & -\sigma_1^2 & cr_{12} & & & cr_{1,n-1} & \sigma_n^2 + cr_{1n} & \end{Vmatrix}}{r_{01} \cdot \det \begin{Vmatrix} \dfrac{cr_{01}+\sigma_1^2}{cr_{01}} & 1 & \cdots & & 1 \\ -\sigma_1^2 & & & & \\ -\sigma_1^2 & cr_{12} & \cdots & & \sigma_n^2 + cr_{1n} \end{Vmatrix}} \tag{1.8}$$

The r_{01} coefficients cancel and in the limit $(cr_{01} + \sigma_1^2)/cr_{01}$ equals one.

Some formulas for the case $n = 2$, will now be given (see fig. 1). For $t_1 \leq t \leq t_2$,

$$\bar{X}_t = \frac{1}{(\sigma_1^2 + \sigma_2^2 + cr_{12})} [(\sigma_2^2 + c(r_{12} - r))\, Y_1 + (\sigma_1^2 + cr)\, Y_2] \tag{1.9}$$

$$\text{Var } X_t = \frac{c^2 r[r_{12} - r] + cr[\sigma_2^2 - \sigma_1^2] + \sigma_1^2(\sigma_2^2 + cr_{12})}{\sigma_1^2 + \sigma_2^2 + cr_{12}} \tag{1.10}$$

In the no noise case ($\sigma_i^2 = 0$) the curve goes through the observed values. A quantity of importance in the smoothing terms is c/σ_i^2. The larger this ratio, the less the smoothing. This results since smoothing is a balance between two factors, the noise variations (reflected in the value of σ_i^2) and the mean square rate at which the curve is assumed to fluctuate.

Another feature of interest is the separation of noise and curve effects in the variance [Eq. (1.10)]. The noise part exists as a result of measurement uncertainties. The curve part exists since the curve, X_t, being the sort of random variable that it is, can never be determined exactly at points at which no observations have been taken.

If n observations are taken at t_1, the expected results $A_{1i} = 1/n$ and Var $X = \sigma_1^2/n$ are obtained. Since X_t is a Brownian motion curve, almost all sample functions are uniformly continuous and bounded on any finite interval. Some à priori information can be added. For example, if it is known or estimated that the function has a particular value X_k at $t = t_k$, this information may be included as the result of an observation with $\sigma_k^2 = 0$.

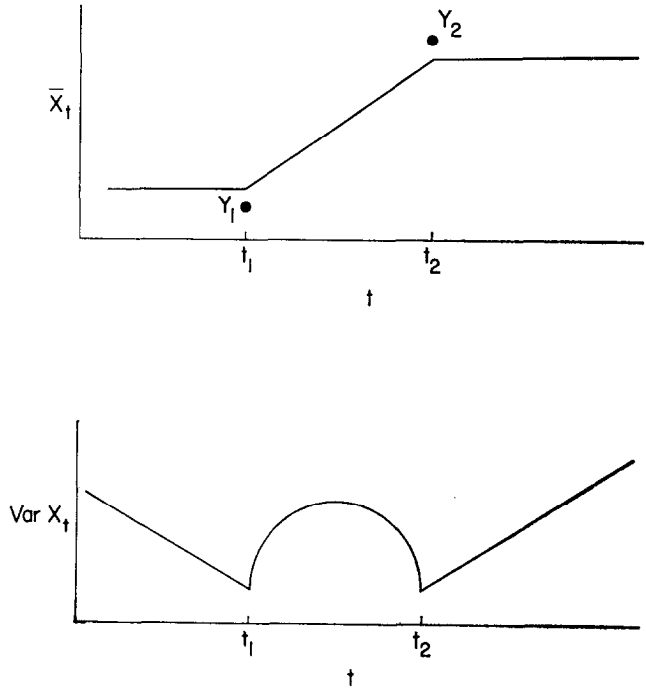

FIG. 1. The smoothed curve and its variance with observations at t_1 and t_2.

II. COMPUTATION SIMPLIFICATIONS

A serious drawback, from the viewpoint usage, is the vast amount of computation necessary to determine $\bar{X}_t$ and Var X_t. These computations involve the evaluation of at least n determinants of order n, where n is the number of observations, in addition to much additional calculation. Clearly, simplification of the computational procedure is desirable. Several methods

of simplification are possible. We will follow the path that seems to yield the simplest results. First we will derive equations expressing the variances in terms of the A_{ki} and then give a simpler method of evaluating the A_{ki} and $\bar{X}_i$.

THEOREM I. *If t_k is an observed point, then Var $X_k = \sigma_k^2 A_{kk}$.*

Letting $t = t_1 (k = 1)$ in Eq. (1.8) puts all the series in the zeroth row of the numerator, except $-\sigma_1^2$, equal to zero. Expanding the matrix about this term yields

$$\frac{\Lambda}{\Lambda_{00}} = -\sigma_1^2 \frac{\Lambda_{01}}{\Lambda_{00}} = \sigma_1^2 A_{11} \tag{2.1}$$

In general, subtracting the kth row of Σ and Σ_{00} from every other row, letting $t = t_k$ and repeating the limiting procedure discussed in connection with Eq. (1.8) yields

$$\frac{\Lambda}{\Lambda_{00}} = -\sigma_k^2 \frac{\Lambda_{0k}}{\Lambda_{00}} = \sigma_k^2 A_{kk}. \tag{2.2}$$

This result has the correct intuitive properties. If the A_{ki} are equal, then Var $X_k = \sigma_k^2/n$.

THEOREM II. *If*

$$t_1 \leq t_k \leq t \leq t_{k+1} \leq t_n \qquad \text{and} \qquad d \equiv \frac{t - t_k}{t_{k+1} - t_k} \equiv \frac{r}{r_{k,k+1}},$$

then

$$\text{Var } X_t = (1-d)^2 \sigma_k^2 A_{kk} + d^2 \sigma_{k+1}^2 A_{k+1,k+1}$$

$$+ d(1-d)(\sigma_k^2 A_{k+1,k} + \sigma_{k+1}^2 A_{k,k+1}) + cd(1-d) r_{k,k+1}. \tag{2.3}$$

Evaluate Λ/Λ_{00} as follows. Subtract $(1-d)$ times row k plus d times row $(k+1)$ from all rows except the kth and $(k+1)$th. Subtract row k from row $k+1$ (do all subtractions in Σ and Σ_{00}). Take the ratio of the determinants of the resultant matrices and repeat the limiting procedure (with respect to r_{01}) on the elements of row k of both numerator and denominator. The zeroth row elements of the resultant numerator matrix are

$$d(1-d)\, c r_{k,k+1}, 0, \cdots, 0, -\sigma_k^2(1-d), -\sigma_{k+1}^2 d, 0, \cdots, 0.$$

The only nonzero entries are the zeroth, kth, and $(k+1)$th. Λ/Λ_{00} is determined by expanding the numerator about row zero as follows.

$$\frac{\Lambda}{\Lambda_{00}} = \frac{1}{\Lambda_{00}} [d(1-d)\, c r_{k,k+1} \Lambda_{00} - \sigma_k^2 (1-d) \Lambda_{0k} - d\sigma_{k+1}^2 \Lambda_{0k+1}]$$

But $-\Lambda_{0i}/\Lambda_{00} = A_{ti} = (1-d)\,A_{ik} + dA_{k+1,i}$ since $\bar{X}_t$ linear in t in the interval $[t_k, t_{k+1}]$. Thus

$$\begin{aligned}\operatorname{Var} X_t = d(1-d)\,cr_{k,k+1} + \sigma_k^2(1-d)\,[(1-d)\,A_{kk} + dA_{k+1,k}] \\ + \sigma_{k+1}^2 d[(1-d)\,A_{k,k+1} + dA_{k+1,k+1}]\end{aligned}$$

which can be reordered to equal the equation in the statement. Theorem I is a special case of Theorem II ($d = 0$).

Statement 1.7 is a simplified variance equation that may be used in the case $t \geq t_n$; similarly for $t \leq t_1$ we have

$$\operatorname{Var} X_t = \operatorname{Var} X_1 + c(t_1 - t). \tag{2.4}$$

We now consider the computation of A_{ki} and will work from the set of equations determined by the linear least squares criterion

$$\frac{\partial}{\partial A_{ki}}\,\mathscr{E}\left(X_k - \sum_1^n A_{ki}\Delta Y_i\right)^2 = 0.$$

Since the system is Gaussian, this set of equations (2.5) yield the desired A_{ki}.

$$\begin{aligned}
cr_{01} &= A_{k1}(\sigma_1^2 + cr_{01}) + A_{k2}cr_{01} + \cdots + A_{kn}cr_{01} \\
cr_{02} &= A_{k1}cr_{01} + A_{k2}(\sigma_2^2 + cr_{02}) + \cdots + A_{kn}cr_{02} \\
cr_{0k} &= A_{k1}cr_{01} + \cdots + A_{kk}(\sigma_k^2 + cr_{0k}) + \cdots + A_{kn}cr_{0k} \\
cr_{0k} &= A_{k1}cr_{01} + \cdots + A_{kn}cr_{0k+1} \\
cr_{0k} &= A_{k1}cr_{01} + \cdots + A_{kn}(\sigma_n^2 + cr_{0n})
\end{aligned} \tag{2.5}$$

First, we prove the useful relation

$$\lim_{r_{01}\to\infty}\left(\sum_1^n A_{ki}\right) = 1. \tag{2.6}$$

Take any of the equations of (2.5), say the first, and divide both sides by r_{01}. This yields

$$1 = A_{k1}\frac{(\sigma_1^2 + cr_{01})}{cr_{01}} + A_{k2} + \cdots + A_{kn},$$

from which (2.6) follows.

Now, commencing with the computation of A_{nj} $(t_n \geq t_i)$, we will describe a relatively simple method of computing all the A_{ij}.

THEOREM III. *A_{nj} can be written as*

$$A_{nj} = b_j/B_n \tag{2.7}$$

where

$$B_j = \sum_1^j b_i, \qquad b_1 = 1 \qquad \text{and} \qquad b_j = (\sigma_{j-1}^2 + cr_{j,j-1}B_{j-1})/\sigma_j^2. \tag{2.8}$$

Letting $k = n$ in (2.5) and transforming (2.5) into a new system by adding $r_{i,i+1}$ to each side of the ith equation and subtracting the ith equation from the $(i+1)$th yields the following set of $n-1$ equations that are independent of r_{01}. The limit condition ($r_{01} \to \infty$) is imposed by using (2.6) as the nth equation.

$$\begin{aligned}
0 &= A_{n1}(\sigma_1^2 + cr_{12}) + A_{n2}\sigma_2^2 \\
0 &= -A_{n1}cr_{23} - A_{n2}(\sigma_2^2 + cr_{23}) + A_{n3}\sigma_3^2 \\
&\vdots \\
0 &= -A_{n1}cr_{n,n-1} \cdots - A_{n,n-1}(cr_{n,n-1} + \sigma_{n-1}^2) + A_{nn}\sigma_n^2 \qquad (2.9) \\
&\sum A_{ni} = 1
\end{aligned}$$

Because of (2.6), A_{nj} may be replaced by the ratio (2.7) and we can substitute b_i for A_{ni} in (2.9). From this new set we see that the b_i are arbitrary up to a multiplicative factor. Thus we may divide b_i by b_1 (in other words set $b_1 = 1$). The set of equations can now be solved for the b_i.

$$\begin{aligned}
b_1 &= 1 \\
b_2 &= (\sigma_1^2 + cr_{12})/\sigma_2^2 \\
b_j &= (\sigma_{j-1}^2 b_{j-1} + cr_{j,j-1}B_{j-1})/\sigma_j^2. \qquad (2.10)
\end{aligned}$$

The A_{1j} can be obtained in a similar manner.

If we define

$$c_n = 1, \qquad C_{i+1} = \sum_{i+1}^n c_j \qquad \text{and} \qquad c_k = (\sigma_{k+1}^2 c_{k+1} + cr_{k,k+1}C_{k+1})/\sigma_k^2 \tag{2.11}$$

then A_{1j} can be written as

$$A_{1j} = c_j/C_1. \tag{2.12}$$

The smoothing coefficients A_{1j}, A_{nj} have now been computed. Next we shall show that all the A_{ij} may be computed by operations on the two sequences b_i and c_i.

THEOREM IV. *If b_i and c_i are defined as in (2.8) and (2.11) and A_{ij} is written as*

$$A_{ij} = A_j \Big/ \sum_1^n A_k, \tag{2.13}$$

then

$$\begin{aligned} A_j &= b_j && (j \leq i) \\ A_j &= (b_i/c_i)\, c_j && (j > i). \end{aligned} \tag{2.14}$$

In the set (2.5), let $k = i$ and replace A_{ij} by A_j. Proceeding as in Theorem III, add $r_{j,j+1}$ to each side of equation j, for $j < i$, and subtract that equation from the $(j+1)$st. For $j \geq i$, subtract the $(j+1)$st from the jth equation. The set (2.15) is obtained by this procedure.

$$\left.\begin{aligned} 0 &= -A_1(\sigma_1^2 + cr_{12}) + A_2\sigma_2^2 \\ 0 &= -A_1 cr_{23} - A_2(\sigma_2^2 + cr_{23}) + A_3\sigma_3^2 \\ &\vdots \\ 0 &= -A_1 cr_{i-1,i} \cdots - A_{i-1}(\sigma_{i-1}^2 + cr_{i,i-1}) + A_i\sigma_i^2 \end{aligned}\right\} \tag{2.15a}$$

$$\left.\begin{aligned} 0 &= A_i\sigma_i^2 - A_{i+1}(\sigma_{i+1}^2 + cr_{i,i+1}) - A_{i+2}cr_{i,i+1} \cdots - A_n cr_{i,i+1} \\ &\vdots \\ 0 &= A_{n-1}\sigma_{n-1}^2 - A_n(\sigma_n^2 + cr_{n,n-1}) \end{aligned}\right\} \tag{2.15b}$$

Proceeding again as in the proof of Theorem III we set $A_1 = 1$ and solve the set (2.15a) for A_j $(j \leq i)$. This yields $A_j = b_j$.

The set (2.15b) is solved similarly to yield (in terms of A_n) $A_j = c_j A_n$. In order for the values of A_i obtained from (2.15a) and (2.15b) to be equal, we must have $A_n = b_i/c_i$ from which follows the statement of the Theorem.

In summary, the computational procedure for the A_{ij} is as follows. Let $h_i = b_i/c_i$,

$$\begin{aligned} m_i &= \sum_1^i b_j + h_i \sum_{i+1}^n c_j \\ &= B_i + h_i C_{i+1} \end{aligned} \tag{2.16}$$

then

$$\begin{aligned} A_{ij} &= b_j/m_i && (j < i) \\ &= h_i c_j/m_i && (j \geq i) \end{aligned} \tag{2.17}$$

It is necessary to recompute only half of the b_j, c_j (hence half the h_j, m_j, B_j

and C_j) after each iterate. Let the $(n+1)$st observation be taken between t_j and t_{j+1}. The situation existing before and after this observation is taken can be seen by referring to Fig. 2(a) and (b) respectively.

c_1		c_j	c_{j+1}	c_n
t_1 t_2		t_j $\hat{t}$	t_{j+1}	t_n
b_1		b_j	b_{j+1}	b_n

FIG. 2(a). Notation prior to new observation at t.

c_1	c_2				c_{n+1}
t_1	t_2	t_j	t_{j+1}	t_{j+2}	t_{n+1}
b_1	b_2				b_{n+1}

FIG. 2(b). Notation post new observation at new $t_{j+1} = t$.

For $i \leq j$, the new b_i equals the old b_i and for $i \geq j+2$ the new c_i equals the old c_{i-1}. Thus the b_i $(i = j+1, \cdots, n)$ and the c_i $(i = 1, \cdots, j+1)$ must be recomputed and these are obtained from Eq. (2.8) and (2.11).

The new b_i and c_i could be computed from their old values and the numbers $\sigma_t^2 | \hat{t} - t_j |$, $| t_{j+1} - \hat{t} |$ [from Fig. 2(a)] but there does not appear to be a saving in computation.

Some further savings are possible if the $\bar{X}_k$ are computed directly. $\bar{X}_k$ is given by

$$\bar{X}_k = \frac{\sum_1^k b_i Y_i + h_k \sum_{k+1}^n c_i Y_i}{m_k}. \tag{2.18}$$

With the running sums

$$P_k = \sum_1^k b_i Y_i, \qquad Q_{k+1} = \sum_{k+1}^n c_i Y_i$$

(2.18) can be written as

$$\bar{X}_k = \frac{P_k + h_k Q_{k+1}}{m_k} \tag{2.19}$$

Thus, to evaluate $\bar{X}_k$ we determine the recursive sequences $\{b, c; B, C; P, Q\}$ and $\{m, h\}$. Half of these $8n$ numbers must be recalculated after each sample.

The A_{ii} and $A_{i,i\pm1}$, that are necessary for the variance calculations are

$$\begin{aligned} A_{ii} &= \frac{b_i}{m_i} \\ A_{i,i-1} &= b_{i-1}/m_i \\ A_{i,i+1} &= h_i c_{i+1}/m_i. \end{aligned} \tag{2.20}$$

III. Sampling Procedures

The sampling procedures will be considered first with no observation noise. The usual desire is to determine a procedure that maximizes (or equivalently minimizes) some function of the observations; for example, their sum or maximum value. This is in general notoriously difficult if the procedure involves looking ahead more than two or three sampling time units.

A natural alternative is to use procedures that look ahead only the distance that can be conveniently handled. If, for example, this alternate procedure sampled at each instant as if to maximize the sum of the next two observations, it will degenerate, in many cases, to sampling the same point all the time. This occurs since the necessary balance between obtaining information and maximizing the quantity of interest is altered in favor of maximizing immediately the quantity of interest.

A possible solution to this problem is the following. Determine the qualitative, and as much as possible the quantitative, variation in the optimum location of the next observation when k rather than two additional observations are to be taken. Sum up this information in a suitable explicit equation whose minimum (or maximum) will yield the sampling point. Preferably, the form of the equation will not depend on k or the exact form of the expected curve or the variance but rather will contain these as parameters.

Several properties held by a wide variety of optimal search processes which a desirable approximate process should have are the following.

1. As N (the total number of observations) tends to infinity, every region of greater than zero size is sampled at least once.

2. For large N, the initial observations will tend to be information gathering (or play the long shot) and be taken near the point of maximum curve variance.

3. The final observations are taken at points where the expected "pay off" (in whatever sense the observations pay off) will be maximum.

Two approximations to optimum policies will now be presented. The definitions

$$\bar{X}^* = \max_t \bar{X}_t, \qquad X^* = \max_t X_t \qquad \text{and} \qquad X^* = X^*_{t^*}$$

will be used.

A. The location of every observation is selected on the basis of a balance between properties 2 and 3. The simplest such balance is a linear weighing. We select the point at which

$$\sqrt{\text{Var } X_t} + f(N, n)\,(\bar{X}_t - \bar{X}^*) \tag{3.1}$$

is maximum. $\bar{X}_t$, $\bar{X}^*$, and Var X_t are the sample curve mean, maximum, and variance just prior to the nth observation and $f(N, n)$ is a positive sequence. If $f(N, n)$ is constant, property 1 may not hold. A sufficient condition for property 1 to hold is $\sqrt{n}\, f(\infty, n) \to 0$.

We will not go into the details of any of the methods discussed in this section but plan to present them somewhat more fully in a future report which will include, in addition, results of computer simulations.

B. Sample at the t point ($\hat{t}$) at which ($\epsilon = \epsilon(N, n)$ is a positive sequence)

$$P(X_t \geq \bar{X}^* + \epsilon) = 1 - \Phi\left(\frac{\bar{X}_t + \epsilon}{\sqrt{\text{Var } X_t}}\right) \tag{3.2}$$

is maximum. Let the absolute value of the slope of the ith interval be K_i, the width T_i, and the optimum point (measured from the endpoint of maximum expected value) be t_i. Within the ith interval, the t_i maximizing (3.2) also minimizes the simpler quantity (3.3).

$$M_i \equiv \frac{(K_i t + \epsilon_i)}{\text{Var } X_t} \tag{3.3}$$

where

$$\text{Var } X_t = \frac{ct}{T_i}(T_i - t) \qquad \epsilon_i = \epsilon + [\bar{X}^* - \max_{t \in I_i} \bar{X}_t]$$

$$\hat{t}_i = \frac{\epsilon_i T_i}{(K_i T_i + 2\epsilon_i)} \tag{3.4}$$

The sample is taken at the $\hat{t}_i$ that minimizes $M_i(\hat{t}_i)$, $i = 1, \cdots, n$.

To concentrate the samples about the location of the sample curve maximum $\epsilon = \epsilon(N, n)$ must tend to zero with n. If property 1 ($N = \infty$) is to hold the additional restraint

$$\sqrt{n}\, \epsilon\,(\infty, n) \to \infty.$$

must be imposed. Under this condition, the points of observation may not be sufficiently concentrated about X^*. It is possible, however, to select an oscillating $\epsilon(\infty, n)$ (see Fig. 3) with the property, for some subsequence n_k,

$$\sqrt{k}\, \epsilon\,(\infty, n_k) \to \infty$$

which will guarantee

$$\lim_{n\to\infty} \frac{1}{n} \sum_{1}^{n} X_i \xrightarrow{\text{a.s.}} X^*,$$

(i.e., that the infinite procedure is consistent).

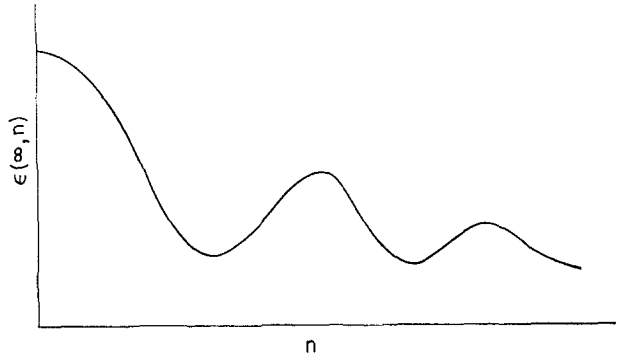

FIG. 3

This modification is useful when no N is given. The process is then unterminating and oscillates between gaining information (when $\epsilon(\infty, n)$ is relatively maximum) and concentrating the observations in the region of $\bar{X}^*(\bar{X}^* \xrightarrow{\text{a.s.}} X^*)$.

IV. THE CHOICE OF c

The efficiency of use of the model is strongly dependent on the value selected for c. This value is the expected mean square rate of amplitude variation (with t) of the family of curves. It is a part of the model that must be determined by the experimenter on the basis of the expected system behavior. In the absence of information and in the event that one of the sampling procedures of Section III is used, a larger rather than a smaller value is desirable and the process should be run until a better idea of the system (or desirable sampling and smoothing procedure) behavior is obtained.

There are some considerations that are helpful in the choice of c. Let us assume that we have two noise free observations (at $t = 0$ and $t = T$) of value zero (see Fig. 4) and that, from past experience, we expect that $X_{T/2}$ will deviate from zero by less than B about half the time.

The fact that about half the area of a normal density function lies within $\pm$ 2/3 of a standard deviation from the mean implies that

$$\tfrac{2}{3}\sqrt{\frac{ct}{T}(T-t)} = \tfrac{2}{3}\sqrt{\frac{cT}{4}} \approx B$$

$$c \approx 9B^2/T$$

A useful value of c cannot be estimated directly from experimental results since (following the model and locating the observations at t_i where $0 = t_0 < t_1 < \cdots < t_n = T$)

$$cT = \sum_{0}^{n} c(t_k - t_{k-1}) = \lim_{\substack{\max(t_k - t_{k-1}) \to 0 \\ n \to \infty}} \sum_{1}^{n} c(t_k - t_{k-1})$$

$$= \lim_{\substack{\max(t_k - t_{k-1}) \to 0 \\ n \to \infty}} \mathcal{E} \sum_{1}^{n} (X_k - X_{k-1})^2$$

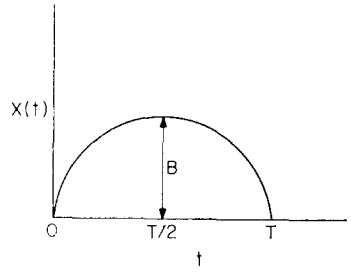

FIG. 4

which is zero if X_t is of bounded variation. Thus the limit of any direct estimate of c will be zero. The value of c and, as a matter of fact, the entire Browian Motion nature of X_t are artifices whose sole function is to aid the design of simple sampling and smoothing procedures.

The assumption of a constant c over the entire parameter space can be generalized. If the assumption of a constant mean square rate of variation of the curve is not reasonable a process of independent increments where

$$\mathcal{E}(X_t - X_s)^2 = |f(t) - f(s)| \equiv \int_s^t c(\tau)\, d(\tau)$$

for $f(t)$ monotone nondecreasing and $c \geq 0$ can be used in lieu of the process of Eq. (1.3). $f(t)$ or $c(t)$ can be chosen to obtain the desired smoothing properties.

The computational difficulties attendent upon this generalization (the expected value is not piecewise linear) can be simplified by replacing the variance increment $cr_{j,j+1} = \mathcal{E}(X_{j+1} - X_j)^2$ by $c(t)\, r_{j,j+1}(t_j \leq t \leq t_{j+1})$. The exact value of t will not be important if dc/dt is small.

V. Time Variable Systems

The variation of the form of X_t with time can be included in the model. A number of forms of this modification are possible; the central feature of all of them is the increase of the curve variance with time. The most convenient

manner of effecting this variance increase is to add a variance increment ΔV to the noise variance before each recomputation of the curve properties. This increase weighs an observation in accordance with the time that has elapsed since it was taken. (It is clear that the effect on the smoothed curve of an observation whose variance tends to infinity will decrease to zero.) This will be demonstrated below. Letting the variance at time n of a past observation located at t_i be $\sigma_i^2(n)$, we have

$$\sigma_i^2(n) = \sigma_i^2(n-1) + \Delta V(t_i).$$

If the observation was taken at $n = k$ then $\sigma_i^2(k) = \sigma_i^2$, the actual noise variance at t_i. The weighing of an observation decreases with the time that has elapsed since it was taken. If the properties of the time variation of X_t are independent of t, then ΔV will be a constant. A useful variation occurs in the case of a moving maximum where we may wish to step up the rate of search in a near neighborhood of the current sample maximum. A $\Delta V(t)$ appropriate to this case is drawn in Fig. 5 (t_m denotes the location of the sample maximum and d and ΔV_m are determined by the expected properties of the time variation).

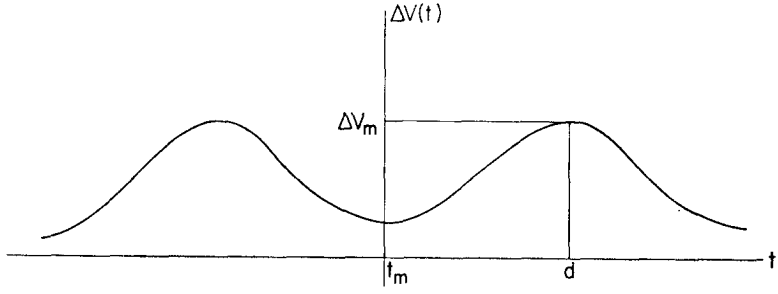

FIG. 5

Due to the increasing variances, the smoothing of the observations in a region is a function of the time distribution of samples in that region. The expected form of the curve in a region that has not been sampled for a long time will become independent of the observations in that region. This may be seen from a computation of a particular smoothing coefficient. Referring to Eqs. (2.10), (2.13), and (2.14) and letting only $\sigma_i^2(n)$ increase as above, we can write a typical A_{ki} (at time n) as

$$A_{ki} = \frac{b_i}{\sum_1^n A_i} = \frac{1}{\sigma_i^2(n)} \frac{\left(\sigma_{i-1}^2 b_{i-1} + c r_{i,i-1} \sum_1^{i-1} b_j\right)}{\sum_{\substack{j=i \\ j=1}}^{n} A_j + A_i}$$

which tends to zero with n since $\sigma_i^2(n) \to \infty$.

It may be desirable to use different sampling sequences $f(N, n)$ and

$\epsilon(N, n)$ here; however, this will not be discussed further. The sampling methods and rate are not necessarily determined by $\Delta V(t)$ and, within the framework of the model, all observations have useful meaning regardless of the value of $\Delta V(t)$. We have tacitly assumed that the system remains nearly stationary during the taking of a single observation. This may not hold in certain cases. However, it will often be possible to add that non-stationary into the noise part of the observation.

VI. Remarks

A heuristic idea inherent in the (sum maximizing) search procedures is the following. Let t_m be the location of the sample maximum X_m. Sample at the point that is (in some intuitive sense) the "most likely competitor" of t_m. Considering method B in particular, the most likely competitor of t_m is the t at which there is the greatest probability that $X_t \geq X_m + \epsilon$. We select $\epsilon > 0$ to prevent the procedure from degenerating and selecting t_m. As more samples are taken and our confidence in the estimate of X_m increases we feel that the most likely competitor criterion should become stricter; hence $\epsilon \to 0$.

Another problem is the modification of the sampling rules when an upper bound ($< \infty$) on the number of local maxima and a lower bound (> 0) on their separation can safely be assumed. The results that have been obtained to date for this case make use of restrictions (functions of the random samples) on the allowable sample points.

Our assumptions of normal and uncorrelated noise may not always be valid. Unfortunately, correlation can so far be handled only at the expense of considerable complication in the computations. Normality is usually approximated in the large sample case but, in any case, the $\bar{X}_t$ curve derived here is the best least squares linear smoothing. The model has been selected to give generally useful and simple computational and search procedures and the results will be useful in any case.

Noise (or uncertainties) in the parameter settings can also be handled. In this case the true parameter setting will be the random variable $t + \eta$, where η has zero mean. Neither the curve maximum nor its location are generally the same (as compared with the case $\eta \equiv 0$) under these conditions. Boundedness of more moments of η than the second are necessary in general but these extra conditions will give us no trouble if the parameter range is finite.

As a simple example of such a problem find the location of the minimum of the expected value of $X = (t + \eta)^2$ where η is a random variable of mean zero and variance σ^2

$$X = t^2 + 2t\eta + \eta^2 \qquad \mathscr{E}X = t^2 + \sigma^2$$
$$X = \mathscr{E}X + (X - \mathscr{E}X) = (t^2 + \sigma^2) + (2t\eta + (\eta^2 - \sigma^2)). \tag{6.1}$$

The first term on the right of (6.1) is the regression function to be minimized, the second term is the (unbiased) observation noise.

The model is applicable to curve fitting in general and is particularly suitable if the curve has a complex form or is of high order or if we wish to trade some accuracy for computation simplification.

If the sample locations increase with time and if we are interested in $\bar{X}_t$ and Var X_t only in the neighborhood of the last sample point, the calculations reduce to a very simple recursive computation. Referring to the results of Theorems I and III, we have, for $t = t_n$

$$\bar{X}_n = \frac{1}{B_n} \sum_1^n b_i Y_i$$

for $t \geq t_n$

$$\text{Var } X_t = \frac{b_n}{B_n} \sigma_n^2 + c \mid t - t_n \mid$$

$$\bar{X}_n = \frac{B_{n-1}}{B_n} \cdot \frac{1}{B_{n-1}} \sum_1^{n-1} b_i Y_i + \frac{b_n}{B_n} Y_n = \frac{B_{n-1}}{B_n} \bar{X}_{n-1} + \frac{b_n}{B_n} Y_n.$$

ACKNOWLEDGMENT

The author wishes to express his indebtness to the earlier unpublished work on this model by Robert A. Sittler.

REFERENCE

1. CRAMER. "Mathematical Methods of Statistics," pp. 314-315. Princeton Univ. Press, 1945.