H. J. KUSHNER
RIAS, Inc.,
Baltimore, Md.

A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise[1]

A versatile and practical method of searching a parameter space is presented. Theoretical and experimental results illustrate the usefulness of the method for such problems as the experimental optimization of the performance of a system with a very general multipeak performance function when the only available information is noise-distributed samples of the function. At present, its usefulness is restricted to optimization with respect to one system parameter. The observations are taken sequentially; but, as opposed to the gradient method, the observation may be located anywhere on the parameter interval. A sequence of estimates of the location of the curve maximum is generated. The location of the next observation may be interpreted as the location of the most likely competitor (with the current best estimate) for the location of the curve maximum. A Brownian motion stochastic process is selected as a model for the unknown function, and the observations are interpreted with respect to the model. The model gives the results a simple intuitive interpretation and allows the use of simple but efficient sampling procedures. The resulting process possesses some powerful convergence properties in the presence of noise; it is nonparametric and, despite its generality, is efficient in the use of observations. The approach seems quite promising as a solution to many of the problems of experimental system optimization.

Introduction

The problem of locating the maximum point of an unknown function $X(t)$ (i.e., the parameter point t at which a system performs best) is of great current interest in technology. Often, the larger part of one's knowledge of the form of $X(t)$ or of the location of its maximum is obtained from the noise-perturbed samples $X(t) + \xi$, where ξ is a zero-mean random variable, taken at selected values of the function parameter t. We will be concerned with the problem where the observations are taken sequentially. In addition, we will confine ourselves to the problem of optimization with respect to a single system parameter. At a given point in the search for the maximum, we have the n numbers, $X(t_i) + \xi_i$, $i = 1, \ldots, n$, and the object of the design of the search policy is to use the information contained in these n numbers to determine a suitable location for the next observation.

A substantial literature exists on many forms of this problem.[2-5] The most common method of attack involves the iterative technique known as the gradient method. The method generates a sequence of estimates of the location of the maximum. If t_n is the location of the nth estimate, then the gradient method locates t_{n+1} at a distance from t_n that is determined entirely by an estimate of the gradient of $X(t)$ in the vicinity of t_n. Several shortcomings derive from the fact that the movement of the t_n sequence depends only on the local properties of $X(t)$; the procedure cannot be used unless there is only one local maximum of $X(t)$; large areas which are relatively flat may prevent convergence or make it exceedingly slow (the plateau problem); $X(t)$ must be constant with respect to time. In addition, a

[1] This work was done at Lincoln Laboratory, M.I.T., Cambridge, Mass., operated with support from the U. S. Army, Navy, and Air Force.

Contributed by the Automatic Control Division of The American Society of Mechanical Engineers and presented at the Joint Automatic Control Conference, Minneapolis, Minn., June 17–21, 1963. Manuscript received at ASME Headquarters, July 19, 1963.

[2] G. E. P. Box, "The Exploration and Exploitation of Response Surfaces: Some General Considerations and Examples," *Biometrics*, vol. 10, 1954, pp. 16–60.

[3] G. E. P. Box and J. S. Hunter, "Multifactor Experimental Designs for Exploring Response Surfaces," *Annals of Mathematical Statistics*, vol. 28, 1957, pp. 195–241.

[4] H. J. Kushner, "Hill Climbing Methods for the Optimization of Multiparameter Noise Disturbed Systems," Journal of Basic Engineering, Trans. ASME, vol. 85, Series D, June, 1963, pp. 157–164.

[5] H. J. Kushner, "A Versatile Stochastic Model of a Function of Unknown and Time-Varying Form," *Journal of Mathematical Analysis and Applications*, vol. 5, 1962, pp. 150–167.

Nomenclature

t = adjustable parameter

t_i, τ_i = locations of observations

$X(t)$ = unknown function whose absolute maximum we are searching for

ξ = observation noise

$Y = X + \xi$ = noise perturbed observation

$\bar{X}_n$ = conditional expectation of $X(t)$ given n-observations

Var $X(t)$ = conditional variance

E = the expectation operator

$N(m,v)$ = normal distribution with mean m and variance v

T_i = width of the ith unsampled interval

$\bar{X}_n{}^*$ = maximum, over t, of $\bar{X}_n$

I_i = ith unsampled interval

c = a smoothness constant for $X(t)$

σ^2 = variance of ξ

Φ = cumulative normal density function

$$\left.\begin{array}{r} A^*, A_i^*, A_T \\ t^*, t_i^* \\ d_\alpha \\ B_i, B \\ P_\alpha \\ \epsilon_n, \epsilon \end{array}\right\} = \text{quantities relating to design of sampling procedure}$$

A_{ki} = smoothing coefficient

σ_{ni} = "effective" noise variance general difficulty with the gradient procedure is that it does not provide for itself a natural starting point (and region of operation) in the parameter space. In practical situations, the rate of convergence may depend strongly on the chosen starting point.

These problems suggest that a search method that is not confined by local properties, but rather obtains and utilizes sufficient information on the entire curve, would be useful. In this paper we investigate some theoretical and practical aspects of this alternate (global) approach to parameter space search. Although this is definitely an area which should be studied, there is, at present, practically no literature on the global search approach. The approach to be developed here requires fewer restrictions than the gradient methods, and hence will yield useful results under more general conditions. In addition, even when the gradient method is to be used, our method will be useful for the location of a good starting point for the gradient process or for the location of the region of parameter space in which the gradient process is to operate.

The design of a global search procedure may be divided into two parts; i.e., the choice of a suitable *model* for the unknown $X(t)$, and the choice of a suitable *search policy*. The model is merely a framework within which the results of the observations may be interpreted. Using this assumed model, the amount of information received from, or the actual usefulness of, an observation are computable quantities. In this regard, the procedure differs considerably from the gradient method.

One choice of model is the parameterized form: Assume $X(t)$ is of the form $\sum_0^r a_i f_i(t)$, where the f_i are given and the a_i are (a priori) jointly normally distributed with some given mean and covariance matrix. Let the observations

$$Y(t_k) = X(t_k) + \xi_k,\ k = 1, \ldots, n,$$

where ξ_k is a Gaussian random variable, be given. It is a straightforward, although perhaps lengthy, process to compute the conditional expectation

$$\bar{X}(t) = E(X(t)|Y(t_1), \ldots, Y(t_n)) \qquad (1)$$

and the conditional variance

$$\text{Var}\ X(t) = \text{Var}\ (X(t)|Y(t_1),\ \ldots,\ Y(t_n)). \qquad (2)$$

This particular model is not desirable, in general, since the required computation is vast, and the true curve may not be of the form

$$\sum_0^r a_i f_i(t)$$

The model which will be discussed in the sequel is much more general, will require less computation, and is more intuitively motivated than the model just described.

When the observations are noise-corrupted, the problem is largely statistical, and whatever the model selected for $X(t)$, the appropriate statistical properties would be used to determine the location of the next observation. An approach that we have found quite promising, and which is to be developed in the paper, is to avoid completely analytic or parametric definitions of $X(t)$ and to define it completely from a statistical point of view; that is, in terms of the joint distribution of its values at several arbitrary points. We will define $X(t)$ as a realization of a particular Gaussian stochastic process. The particular process that we have found most convenient is the process of Gaussian, independent, and infinitely divisible increments (the classical Brownian motion process). Among the many advantages of this model, one of the most useful is that the conditional expectation of $X(t)$ is a piecewise linear approximation of $X(t)$. It is completely nonrestrictive, and *any* system performance function can be considered to be of this form. For the purposes of the rest of the development, the noise components of the observations, ξ_k, are assumed to be independent of each other and normally distributed. The assumption of normality is not important and is useful only for giving an exact probabilistic interpretation to the results.

Once the model for $X(t)$ is chosen, it remains to select the quantity that we wish to maximize (or the error that we wish to minimize) and to determine the appropriate sampling policy, that is to say, a procedure for the use of the first n-observations to compute the location of the $n + 1^{st}$. The sampling policies for the most desirable error criteria are generally the most difficult to implement. For example, let N be the total number of additional observations that are to be taken, let the random variable $\bar{X}_N$ be the conditional expectation of $X(t)$ at the termination of all sampling, and let the quantity to be maximized be the maximum over t of $E\bar{X}_N$, the expected value of $\bar{X}_N$. (At time n, the quantity $\bar{X}_N$ is a random variable since it depends on the observations $Y(t_{n+1}),\ \ldots,\ Y(t_{N+n})$ which have not yet been taken.) The maximization of

$$\max_t E\bar{X}_N$$

is certainly a reasonable goal since we are, in fact, looking for the maximum of $X(t)$. $E\bar{X}_N$ depends on the locations of the future observations, and it is generally so complicated that it usually is not a practical calculation with present-day computing equipment. The primary source of the computational difficulty derives from the fact that the procedure is required to look ahead N-samples and to compute the effects of the next sample through the subsequent N-1 stages.

Since the location of the optimum sample points for the foregoing criterion (as well as for other reasonable criteria)[6] is virtually impossible to compute, we have developed, on the basis of heuristic or intuitive considerations, a suboptimum alternate. The alternate, however, does yield quite satisfactory results. The location of the next observation is a simple function of the value of the two functions[7]: $\bar{X}(t)$ and Var $X(t)$.

The expressions for $\bar{X}(t)$ and Var $X(t)$ in terms of the $Y(t_i)$ are much simpler than one would expect.[5] A sample taken at a point where $\bar{X}(t)$ is maximum will yield the largest expected value, while an observation taken where Var $X(t)$ is maximum yields much more information on the unknown curve. The point at which the observation is actually taken is determined by a suitable weighting of the two quantities. Although the design of the search policy requires little effort, behavior is rather sophisticated. The design can be refined to include many forms of a priori information.

The work presented here is a further development of the work reported,[5] where there appears a detailed discussion of the model, its versatility, and methods of substantially simplifying the computations required in the search procedure. This paper is concerned with the study of suitable search policies and with a description of the experimental results that have been attained in practical situations.

The next section will review briefly several properties of the model. The search procedure and the results of the computer experiments are discussed in subsequent sections for situations without and with observation noise, respectively. The *no* observation noise case provides the basis for the more general case.

There are several possible extensions of the results to be described to multiparameter systems. At the present state of our study, these extensions are computationally quite lengthy and hence, in this paper, we will restrict our concern to the single-parameter system.

[6] Since we are only interested in the maximum of $X(t)$, the most useful quantities to maximize will be functions of mean and variance functions at the termination of all sampling.

[7] If ξ_k is not normal, then $\bar{X}(t)$ is the least-squares estimator of $X(t)$, where $X(t)$ is defined according to the foregoing model and Var $X(t)$ is the mean-square error in the estimate of $X(t)$.

If $X(t_i) - X(t_j)$ can be considered as a signal and $\xi_i - \xi_j$ as noise, then the method is applicable to *any* signal-to-noise ratio (SNR). Since the design of the sampling procedure depends on what is known or assumed about the SNR, the more accurately the SNR is known, the better the sampling procedure will be. The method is also quite useful if there is no observation noise.

The Curve Model

The model that we have selected for the unknown function $X(t)$ has many interesting properties, many of which have been described elsewhere.[5] Here we will give without proof the properties that will be used or referred to in later sections. The proofs of all statements may be found in the previous paper.[5]

The model is a Gaussian random function with independent, infinitely diversible increments. Hence, for arbitrary t_1 and t_2, we have

$$X(t_2) - X(t_1) \sim N(0, c|t_2 - t_1|) \tag{3}$$

where c is the mean-square rate of variation of the curve. In order to use the model when the observations are corrupted by additive noise, it is necessary to select a value for c, and this will be pursued further in the last section. The definition of equation (3) in terms of increments is useful for the problem of locating the function maximum because there we are interested only in the functional differences and not in the absolute values.

The autocorrelation

$$E(X(t_1) - X(t_2))(X(t_3) - X(t_4))$$

equals c times the length of overlap of the two intervals (t_1, t_2) and (t_3, t_4). From this and the joint normality of the collection $Y(t_i) - X(t)$, $i = 1, \ldots n$, the values of $\bar{X}(t)$ and Var $X(t)$ may be computed.

A particularly interesting property of the model is that $\bar{X}(t)$ is piecewise linear and the breakpoints are the locations of the observations; i.e., $\bar{X}(t)$ is a piecewise linear approximation to $X(t)$. If $t_n \geq t_j$ for all j, then $\bar{X}(t) = \bar{X}(t_n)$ for $t \geq t_n$. In the particular case where Var $\xi = 0$, $\bar{X}(t_k) = X(t_k)$ and Var $X(t_k) = 0$. In general, the variance is quadratic in the interval between observed points and increases linearly as $c|t - t_n|$ for $t \geq t_n$.

The formulas for the case $n = 2$, $t_1 \leq t \leq t_2$, and Var $\xi_i = \sigma^2$, are

$$\bar{X}(t) = \left[\frac{(\sigma^2 + c(t_2 - t))Y(t_1) + (\sigma^2 + c(t - t_1))Y(t_2)}{2\sigma^2 + c(t_2 - t_1)}\right] \tag{4}$$

$$\text{Var } X(t) = \frac{c^2(t - t_1)(t_2 - t) + \sigma^2(\sigma^2 + c(t_2 - t_1))}{2\sigma^2 + c(t_2 - t_1)} \tag{5}$$

In general, when t_k is a sampled point,

$$\bar{X}(t_k) = \sum_1^n A_{ki}Y(t_i) \tag{6}$$

Owing to equation (6) and the piecewise linearity of $\bar{X}(t)$, $\bar{X}(t)$ is linear in the observation $Y(t_i)$. The general equation for Var X(t) is

$$\text{Var } X(t) = d(1 - d)cT_k + \sigma^2{}_{ni} \tag{7}$$

where

$$\sigma_{ni}{}^2 = \sigma^2[A_{kk}(1 - d)^2 + A_{k+1,\,k+1}d^2 + 2\mathrm{A}_{k+1,\,k}d(\mathrm{l{-}d})] \tag{8}$$

$$T_k = t_{k+1} - t_k,\ t_k \leq t \leq t_{k+1}, \quad \text{and} \quad d = (t - t_k)/T_k$$

The piecewise linearity of $\bar{X}(t)$ and the piecewise linear or quadratic nature of Var $X(t)$ facilitate the visual interpretation of the data and the intuitive relation between the data and the true curve. Some of the foregoing properties for the case of two observations $Y(t_1)$ and $Y(t_2)$ are illustrated in Fig. 1.

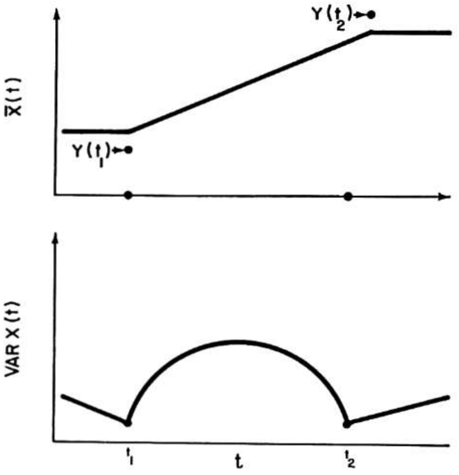

Fig. 1 Mean and variance of $X(t)$ conditioned upon values of observations at t_1 and t_2

Sampling Procedures—No Observation Noise

Since a mathematical model of $X(t)$ is available, it is theoretically possible, once a criterion of optimality is given, to determine the mathematically optimum sampling policy. However, as mentioned in the introduction, the determination of the optimum sampling policies is extremely difficult. Because of this, the development of our sampling laws has been guided primarily by heuristic considerations. There are some advantages to the approximate approach. From the following reasoning, we see that its use may yield better results than would a procedure that is optimum for the model. Although the model selected for $X(t)$ is the best that we have found for our purposes, it is sometimes too general; for example, the Fourier series expansion of a true random walk function has components of all frequencies, while the true curve will generally be bounded in frequency. Information concerning the frequency bound is often available in the form of the relative smoothness of $X(t)$ and the distribution of the maxima, and so on. This information may be taken into account in the intuitively developed sampling procedure, thereby increasing the efficiency of use of the samples. This is, in fact, one way of interpreting the procedure to be developed in this section.

We will now describe in detail a sampling procedure that we have studied both theoretically and experimentally. Let[8]

$$\bar{X}_n{}^* = \max_t E(X(t)|Y(t_1), \ldots, Y(t_n))$$
$$= \max\,(\bar{X}(t_1), \ldots, \bar{X}(t_n))$$

Although we will occasionally use the more general terminology, in this section it should be kept in mind that $Y(t_i) = X(t_i)$ and

$$\bar{X}_n{}^* = \max\,(X(t_1), \ldots, X(t_n))$$

in the noiseless case. Let t^*, the location of the next observation, be the point that maximizes

$$\text{Prob } (X(t) \geq \bar{X}_n{}^* + \epsilon_n) \tag{9}$$

where ϵ_n is a series of positive constants.

The procedure of equation (9) may be thought of as follows: It samples at the point that is the most likely competitor (to the location of $\bar{X}_n{}^*$) for the location of the maximum of $X(t)$. The criterion of competitiveness is the probability that $X(t)$ exceeds $\bar{X}_n{}^* + \epsilon_n$, and we sample at the point where the probability of

[8] **Assume, when necessary, that $\bar{X}_n{}^*$ is unique.**

this event is greatest. The choice of the ϵ_n governs our control over the sampling process. It determines, in the foregoing, the strictness of the criterion of competitiveness. This will be made clear in the discussion to follow. In addition to this intuitive significance of the rule of equation (9), it is computationally simple. The choice of a sampling procedure is now reduced to the choice of a suitable ϵ_n-sequence.

Since $X(t)$ is normally distributed, equation (9) equals

$$1 - \Phi\left\{\frac{\bar{X}_n{}^* - \bar{X}(t) + \epsilon_n}{[\text{Var } X(t)]^{1/2}}\right\},$$

where Φ is the cumulative normal density function. Since Φ is a monotonically increasing function, t^* also minimizes the simpler quantity

$$A = (\bar{X}_n{}^* - \bar{X}(t) + \epsilon_n)^2/\text{Var } X(t) \qquad (10)$$

Now, before considering the method of implementation and the properties of the sampling rule in the general case, let us investigate the situation after two observations have been taken. Let the sampling interval be $(0, T)$, take one observation at each endpoint, and assume

$$X(0) = 0, \qquad X(T) = -B$$

Thus, from the properties described in the preceding section, we have, for $0 \le t \le T$,

$$\text{Var } X(t) = ct(T - t)/T$$

$$\bar{X}(t) = -tB/T$$

$$\bar{X}_2{}^* = \max (X(t_1), X(t_2)) = 0$$

In addition

$$A = (tB/T + \epsilon_2)^2/[ct(T - t)/T] \qquad (11)$$

$$t^* = \epsilon_2 T/(2\epsilon_2 + B) \qquad (12)$$

The minimum value of A is

$$A^* = 4\epsilon_2(\epsilon_2 + B)/cT \qquad (13)$$

Since t^* is independent of c, the noiseless sampling policy does not depend on c; hence, we will set $c = 1$ where convenient.

Equation (12) illustrates one of the salient characteristics of the sampling policy. As $\epsilon_2 \to \infty$, $t^* \to T/2$, the point of maximum variance [the point where the value of $X(t)$ is most uncertain]. As $\epsilon_2 \to 0$, $t^* \to 0$, the location of the maximum of $\bar{X}(t)$ (the point where the expected value of the observation is greatest). This remark suggests that ϵ_n should be large at the start of the sampling process (when we are interested in obtaining information such as the locations of the regions of the various peaks of $X(t)$), and decrease as we approach the end of the sampling process (when we wish to search closely the previously located regions of the peaks and spend only a relatively few observations searching for a narrow peak that may possibly have been missed).

We now discuss the implementation of the computation when the n-observations $X(t_1), \ldots, X(t_n)$ have been taken. Let $0 = t_1 \le t_2 \le \ldots t_n = T$, let I_i denote the ith interval $(t_i \le t \le t_{i+1})$, and $T_i = t_{i+1} - t_i$ be the width of I_i. Since $X(t)$ is piecewise linear, the computation in the general case may be done interval by interval in the following way: Define the quantity A_i as the value of A in I_i. Thus

$$A_i = \frac{(\epsilon_n + \bar{X}_n{}^* - X(t_i) + B_i(t - t_i)/T_i)^2}{(t - t_i)(t_{i+1} - t)/T_i} \qquad (14)$$

where $B_i = X(t_i) - X(t_{i+1})$. The location of the minimum of A_i is

$$t_i{}^* = \frac{(\epsilon_n + \bar{X}_n{}^* - X(t_i))T_i}{2(\epsilon_n + \bar{X}_n{}^* - X(t_i)) + B_i} + t_i \qquad (15)$$

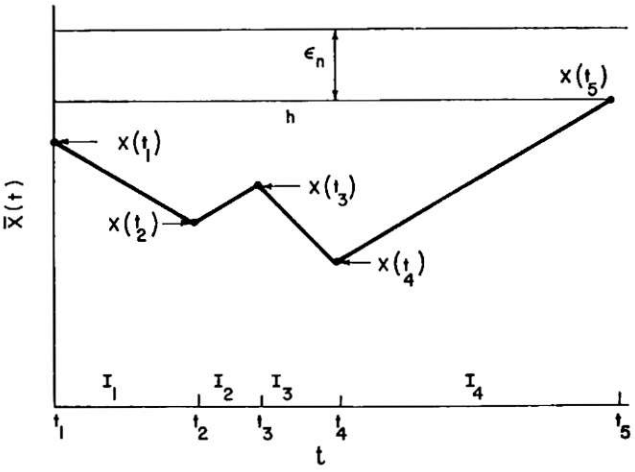

Fig. 2

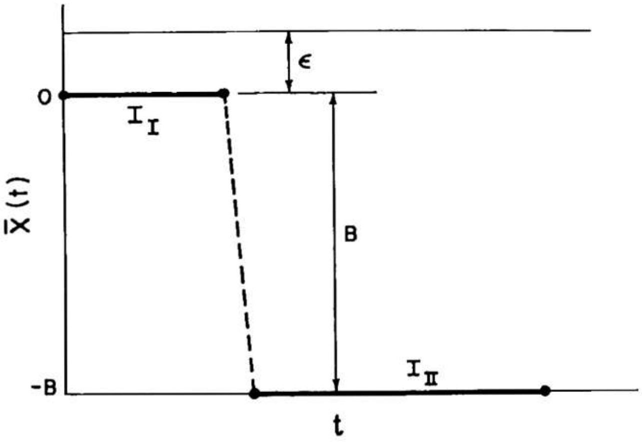

Fig. 3

The minimum value of A_i is

$$A_i{}^* = 4(\epsilon_n + \bar{X}_n{}^* - X(t_i))(\epsilon_n + \bar{X}_n{}^* - X(t_i) + B_i)/T \quad (16)$$

These concepts are illustrated in Fig. 2. Thus, in the genera case, the sampling procedure involves the computation of the minimum of the numbers $A_1{}^*, \ldots, A^*{}_{n-1}$. The location (t^*) of the $n + 1^{st}$ observation is the $t_i{}^*$ which corresponds to the minimum of $A_i{}^*$. The entire process may be realized by a very simple computer program.

It should be observed that, as in the case of two observations, t^* tends to the center of the widest unsampled interval as $\epsilon_n \to \infty$, and to the location of $\bar{X}_n{}^*$ as $\epsilon_n \to 0$.

We will next describe a method for selecting the ϵ_n that we have found to yield satisfactory results. The following idea underlies the method of selection. The search is divided into stages. We do a rough search first. This rough search gives us a feeling for the general character of the curve and locates the regions in which the large maxima lie. A finer search is then done. Most of the effort of the finer search is concentrated about the regions that the rough search has shown to be most promising, although several observations may be used to continue the rough search. This must be done when there is a possibility that there exists an important peak which had been missed by the initial rough search. The process continues through a sequence of finer and finer search stages. The value of ϵ_n is held constant during each stage, but decreases from stage to stage. The length of each stage is a random variable whose value is determined by the outcome. The roughness of search in each stage and the relative degree of emphasis on the more promising and the less promising areas must be determined by the experimenter.

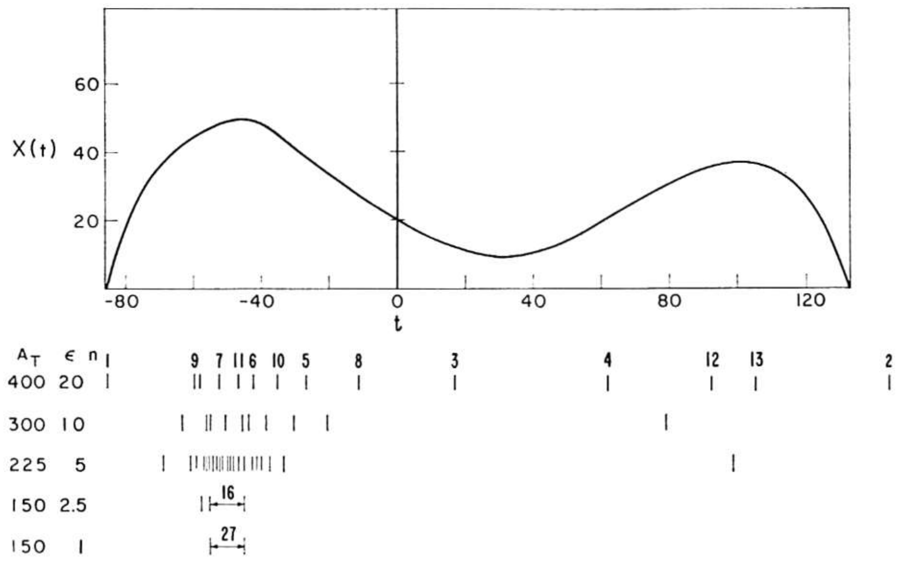

Fig. 4 Experimental results with no observation noise; locations of observations

Table 1

Stage	1	2	3	4	5
ϵ	20	10	5	2.5	1
A_T	400	300	220	150	150
d_0	4*	1.33*	.45*	.14*	.027*
d_2	4.85	1.93	.89	.54	0.24*
d_5	6.25	3	1.8	1.5*	.96
d_{10}	9	5.33	4.1*	4.15	3.23
d_{20}	16	12*	11.3	13.5	11.8
d_{40}	36*	33	37	48	45

Table 2

Stage	1	2	3
ϵ	20	10	5
A_T	424	424	424
d_0	3.77*	.94*	.24*
d_2	4.55	1.36	.46
d_5	5.9	2.12	.94
d_{10}	8.75	3.75	2.12*
d_{20}	15	8.75*	4.95
d_{40}	34*	23.6	19.1

It is here that all a priori knowledge on $X(t)$ may be made use of. Henceforth, the subscript on ϵ_n will be dropped.

We will now illustrate the design of the first stage of search. The desired roughness of search, i.e., distance between the locations of the observations, in a region is a function of $\bar{X}_n^* - \bar{X}(t)$. As this quantity increases, the likelihood of a peak in that region decreases, and the observations may be located further apart. We assume that the following two hypothetical regions will occur in the problem: Region I, where $\bar{X}(t) \approx \bar{X}_n^*$, and Region II, where $\bar{X}(t) \approx \bar{X}_n^* - B$ for some given B; see Fig. 3. For each region, we then determine the desired minimum distance between observations. Denote these minimum distances by d_0 and d_B, respectively. The choice of these distances makes use of all the available a priori information. They should be determined from our knowledge of, or feeling for, the relative likelihood of a peak in intervals of widths d_0 and d_B in the two regions described. Information that is sufficient to yield good sampling procedures is almost always available. At the time when the desired type of search is complete, we have from equation (14)

$$A_{\text{I}}^* \geq 4\epsilon^2/d_0 \qquad (17a)$$

$$A_{\text{II}}^* \geq 4(\epsilon + B)^2/d_B \qquad (17b)$$

Equating the right-hand sides of equationᴾ (17a) and (17b) yields the ϵ which gives the desired ratio d_0/d_B. The resulting common value of the right-hand sides of equations (17a) and (17b) is the minimum value of A^* (denoted by A_T) that the A_i^* will attain upon completion of the stage. A_T determines the minimum values of d_0 and d_B. For the given problem, ϵ is determined on the basis of such a hypothetical situation and sampling continues, using the rule of equation (10) [or equations (14–16)] until all the A_i^* are greater than or equal to A_T. The sampling rule of equation (10) may be viewed as an interpolation formula, determining the roughness of search for all situations other than the selected two.

The next stage is similarly designed, but with smaller B, d_0, and d_B. The procedure generally works well and provides a reasonable method for locating the next observation.

The results of the computer studies illustrate the entire approach and the performance that one may expect. For the first stage of the computer run illustrated in Fig. 4, we selected $B = 40$, $d_0 = 4$, and $d_{40} = 36$. This choice implies $\epsilon = 20$ and $A_T = 400$ and, in addition, $d_{10} = 9$ and $d_{20} = 16$. The design of the subsequent stages is given in Table 1. The starred quantities are those used to determine ϵ and A_T for the particular stage of search. Table 2 gives the design of the sampling law for the results in Fig. 5.

Two distinct views toward sampling policy design are illustrated in Tables 1 and 2. The results of the sampling policy of Table 1 and the curve to which it was applied are given in Fig. 4. In this case, the available a priori information indicated that there could be no $X(t) \geq \bar{X}_n^*$ in regions which were more than a given distance (B_1) below $\bar{X}_n^*$, and whose unsampled width was less than a given width (d_{B1}); i.e., a bound on the rate of rise of the curve was known. We shall designate any such region as (d_{B1}, B_1). Thus once such regions are discovered, there is no need to sample them further. This effect is easily provided for before the run is started (without need for any of the empirical samples). We decide how closely (d_0) we should search the more promising regions that will turn up during the first stage.[9] From d_0, d_{B1}, and B_1, ϵ and A_T are computed, and the first sampling stage continues until all regions (d_{B1}, B_1) are found, and the desired type of search in the more promising regions has been realized. For the second stage, B_2, d_{B2}, and d_0 are similarly selected. In order to eliminate the regions in which there can be no absolute maxima, such as (d_{B1}, B_1), from further consider-

[9] In connection with the discussion of the results, there will be some remarks on the choice of d_0.

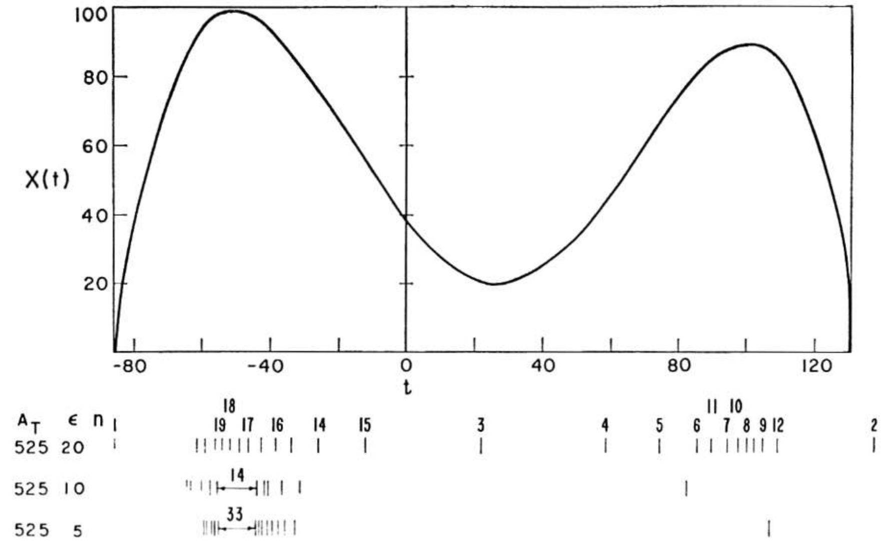

Fig. 5 Experimental results with no observation noise; locations of observations

ation, it is only necessary that d_{B1} (second stage) $\geq d_{B1}$ (first stage). This is illustrated by the final increase in the d_{20} sequence of Table 1.

In the run of Table 2 (corresponding to Fig. 5), the search in the more promising areas was emphasized, but an interest in the search of all regions was maintained. (This is indicated by the fact that A_T is constant.) The a priori information available in this case indicated a strong possibility that the highest hills were rather close in amplitude; thus we required a finer search of the neighborhoods of the peaks of the higher hills.

The behavior that is typical of the process is illustrated by the observation numbers listed above the $\epsilon = 20$ lines in Figs. 4 and 5. Let X_n and τ_n refer to the nth observation and its location, respectively. In the run of Fig. 4, X_1 and X_2 were taken at the endpoints of the sampling interval $(-85, 130)$. Since X_2 is slightly less than X_1, τ_3 is slightly to the left of the center of the interval. Since ϵ is large relative to the variations in the observations already taken, the variation (with respect to t) of the denominator of equation (10) is greater than the variation in the numerator; hence, the process concentrates on a rather rough search, and X_4 and X_5 are taken. Since X_5 is sufficiently larger than the previous observations, the region in the vicinity of τ_5 is searched for a peak by X_6 to X_{11}. (The location of a new sample maximum immediately increases the A_i* of all other intervals.) Once the search points on the hill are sufficiently close (the A_i* in this region exceed the A_i* in other regions), the process moves to regions with the wider unsampled intervals, and takes X_{12} and X_{13}. At this point, the threshold criterion for the end of the rough search is satisfied everywhere, and a finer search $(\epsilon = 10)$ commences. The relative distribution of the observations is determined by the design given in Table 1.

There are two large hills to be investigated in Fig. 5, and the process first climbs the hill on the left. It is characteristic of the process that whenever the last observation is sufficiently larger than the previous observations, the area of the last observation will immediately be searched for the peak of the hill upon which that observation lies. The behavior of the process in climbing both hills in Fig. 5 (the direction from which it proceeds and the decreasing step size) is due to the relative smallness of X_1 and X_2.

Some remarks on the selection of d_0 will now be made: d_0 should be sufficiently small for the process to distinguish the significant variations in the neighborhood of a maximum. For example, if there are several competing peaks, these regions should be searched well enough to determine, with a high probability, which one is the highest. By judiciously selecting the d_0, d_B, and B of the subsequent stages, the search of the apparently smaller but still competitive peaks may still continue to some extent (as illustrated in Figs. 4 and 5). Hence, it is not necessary to establish with certainty the region of the absolute maximum in the first stage. The observations are used more efficiently if the first stage just accomplishes the job of isolating the more promising regions. The sizes of the steps which the process takes when climbing a hill (e.g., the distance between the points τ_4 to τ_9 in Fig. 5), decrease as d_0 decreases; hence, for small d_0, the rate of climb of the major peaks is slow and, in addition, too much effort is spent in climbing the minor peaks.

Although we have gone into considerable detail in the discussion of the design and the behavior of the process, the design procedure is actually rather simple and yields a process with a very sophisticated sampling behavior.

An interesting alternate sampling policy, whose properties are close to the properties of equation (9), is given in Appendix 1. It should be emphasized that the procedure of equation (9) is only one of many reasonable and simple possibilities. A study of some of these possibilities may yield procedures that are particularly suited to special classes of problems.

Sampling Procedure: Observation Noise Case

The basic sampling rule of equation (9), or its equivalent equation (10), and the stagewise type of search discussed in the preceding section also yield quite satisfactory results when the observations are corrupted by additive noise $(Y(t_k) = X(t_k) + \xi_k)$. In this case, however, it is necessary to select a value for c. The computation required to evaluate $\bar{X}(t)$ and Var $X(t)$ is greater here[10] than where $\xi_k = 0$ but, as mentioned in an earlier section, $\bar{X}(t)$ is still piecewise linear and Var $X(t)$ is still piecewise quadratic. The method of computation of these quantities is given elsewhere.[5]

With the use of equation (7), equation (14) may be written as

$$A_i = \frac{(\epsilon + (\bar{X}_n^* - \bar{X}(t_i)) + B_i(t - t_i)/T_i)^2}{\sigma_{ni}^2 + c(t - t_i)(t_{i+1} - t)/T_i} \qquad (18)[11]$$

The exact expression for A_i* and t_i* are easily derivable from equation (18), but will not be given here for the following reason: Approximations are often useful for the simplification of the computation of A_i*. For example, when the $c(t_{i+1} - t_i)/\sigma^2 \ll 1$,

[10] The time required for the computation of the location of the next sample is proportional to n.[5]

[11] Recall that

$$A_i^* = \max_{t \epsilon I_i} A_i$$

The direct maximization of equation (18) may yield a t_i* not in I_i, when there is observation noise. In this case t_i* is the nearest endpoint of I_i.

the very simple approximation to Var $X(t)$ that is given in Appendix 2 may be used. An alternate approximation to Var $X(t)$ is obtained by assuming the A_{ki} in equation (8) to be equal; hence, $\sigma_{ni}^2 \approx A_{ii}\sigma^2$.

There are two points of view that may be taken toward the sampling policy discussed in the preceding section. The values of ϵ and A_T could be selected, as discussed in the preceding section, on the basis of the desired fineness of search at various selected levels below the sample maximum. An alternate is to select an ϵ and use the form of equation (9) to select a probability threshold, P_T. The two approaches are strongly related, since for any given A_T, P_T may be uniquely determined, and vice versa.

For the purpose of design of the sampling procedure in the noise case, it is useful to keep the equality in mind. Our procedure is to neglect the noise and select ϵ and A_T as in the preceding section. One selects B, d_0, and d_B, and then computes ϵ and A_T. The only difference is that in order to normalize the threshold with respect to c, as was done in the noiseless case, we use A_T/c as the threshold. In terms of the probabilistic interpretation, we have selected an ϵ and a threshold P_T, and we sample until the inequality

$$P(X(t) \geq \bar{X}_n^* + \epsilon) \leq P_T \tag{19}$$

is satisfied uniformly in t. This will require more observations than the no-noise case owing to the presence of σ_{ni}^2 in the denominator of equation (18). The results, however, in the sense of the relation between equation (9) and equation (19) are equivalent. We thus design the procedure as we would if there were no noise. The noise effects are automatically taken into account.

One may design a procedure which guarantees

$$\bar{X}_n^* \to X > \max_t X(t) - \delta$$

in probability[12] for an arbitrarily small δ. This will be true regardless of the (nonzero) value of c selected. In rough outline[13] this is constructed as follows. Select the (nonzero) $\epsilon_n \to 0$ as previously. Let A_T, or some infinite subsequence of the A_T, increase without bound. Hence, a subsequence of the equivalent P_T converges to zero. Since $\epsilon_n < \delta$ for large n, we have, uniformly in t,

$$P(X(t) \geq \bar{X}_n^* + \delta) \leq P(X(t) \geq \bar{X}_n^* + \epsilon_n) \to 0 \tag{20}$$

The result also holds in the no-noise case.

A reasonable approximation to the asymptotic variance is given in Appendix 2, and one may use this to compute the approximate distribution of the observations.

Table 3

Stage	1	2	3
A_T	400	200	100
ϵ	20	10	5
d_0	4*	2*	1*
d_1	4.4	2.42	1.45
d_2	4.85	2.8	1.95*
d_5	6.25	4.5*	4
d_{10}	9	8	9
d_{20}	16*	18	25

The designs used in runs in Figs. 6 to 10 are given in Table 3.

Recall that A_T/c is the true threshold for the A_i of equation (18). Several salient features of the method are illustrated by the figures. The smoothing of the observations is proportional to σ^2/c; the $\bar{X}(t)$ in Fig. 6, with the lower value of c, are smoother than the corresponding curves in Fig. 7. Owing to this greater smoothness in Fig. 6 and the greater importance of noise there, the satisfaction of the threshold (everywhere) for $\epsilon = 20$, required more observations there than in Fig. 7. The same phenomenon repeats in Figs. 8 and 9. The desired smoothing properties are the primary considerations in the selection of c.

As σ^2 increases, the tendency of the observations to cluster about the sample maximum decreases. There is always some differentiation, however, and more effort is generally spent in the neighborhood of the sample peak than is spent elsewhere in the sampling interval. As in the no-noise case, the search of an area persists until the process is satisfied that the area is no more promising with respect to its immediate goal [minimizing equation (9)] than the results previously attained elsewhere on the parameter interval.

The Choice of *c*

The use of the sampling procedure of equation (9) in the presence of observation noise requires a choice of c. The smoothing increases with decreasing c. As the smoothing increases, the noise component of Var $X(t)$ decreases; and, in addition, the curve variations become less distinguishable [the distances $\bar{X}_n^* - \bar{X}(t)$ decrease]. For large n, or small distances between observations, the noise component constitutes the greater part of Var $X(t)$. [See Appendix 2 or equation (5).] Hence, an increase in c often decreases both the numerator and denominator of the A_i^*. This effect allows the satisfactory use of a range of c. A suitable range must nevertheless be selected.

We have defined c as

$$c = E(X(t + T) - X(t))^2/T \tag{21}$$

A suitable value of c may be selected by considering the anticipated relative effects of the noise and of the curve variations on the differences between the values of observations taken at different points. Select a distance, T, and estimate the variance of a Gaussian random variable which would have about the same distribution as $X(t + T) - X(t)$. For the runs in Figs. 6 to 10, we assumed that the variations in the curve values across a t-interval of unit width would be distributed about the same as samples drawn from a distribution whose variance was in the range of 1 to 3.

As an alternate point of view, consider the case of observations $Y(0)$ and $Y(T)$. $\bar{X}(0)$ may be obtained from equation (4). The relative effect of the observation at T to that at 0 on $\bar{X}(0)$ is $h/(h + 1)$ where $h = \sigma^2/cT$. For a given T, one may select c so that a satisfactory smoothing is obtained.

Generally, c will depend on the selected distance T. In particular, if the true $X(t)$ is differentiable or of bounded variation, the anticipated c will approach 0 as T approaches 0.[5] Thus one may construct [from a priori knowledge on the behavior of $X(t)$] a curve relating c to the comparison interval T, and let the c that is used in a region depend upon the T_i in that region. Although the use of $c(T)$ is sufficiently simple, our program used a single value, the maximum of the anticipated $c(T)$. The maximum was chosen in order to minimize the effects on $\bar{X}(t)$ of the observations taken in the low areas between peaks

Conclusions

A new, versatile, and practical method of parameter space search has been presented. The experimental results illustrate its usefulness for the problem of experimentally and sequentially locating the maximum of a multipeak performance function of a single variable when the only available information is noise-disturbed samples of the function. The method is extremely

[12] Convergence of any type implies that the procedure is unterminating. Although this will never be the case in a practical situation, such asymptotic properties are of interest because they indicate the power or limitations of the method.

[13] The proof is only sketched here. One must show that at each search level only finitely many observations will be taken and also that the convergence of equation (19), which is for a fixed t, holds for t simultaneously.

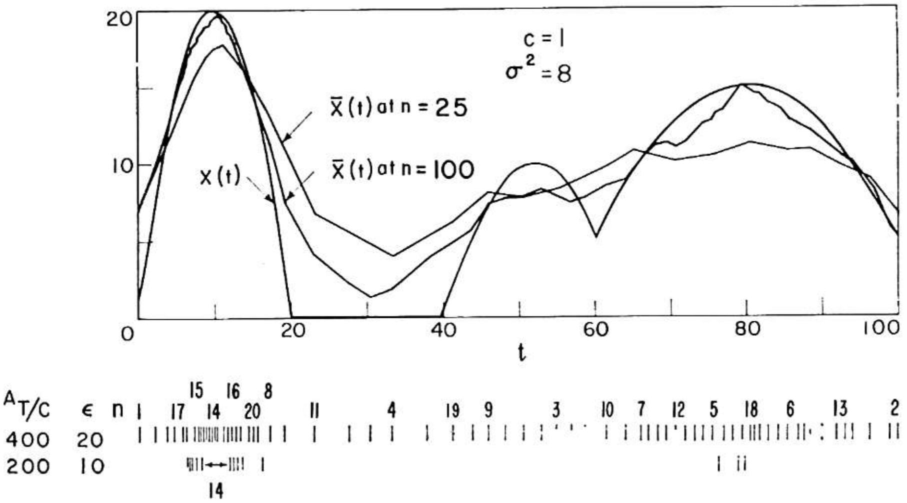

Fig. 6 Experimental results with observation noise

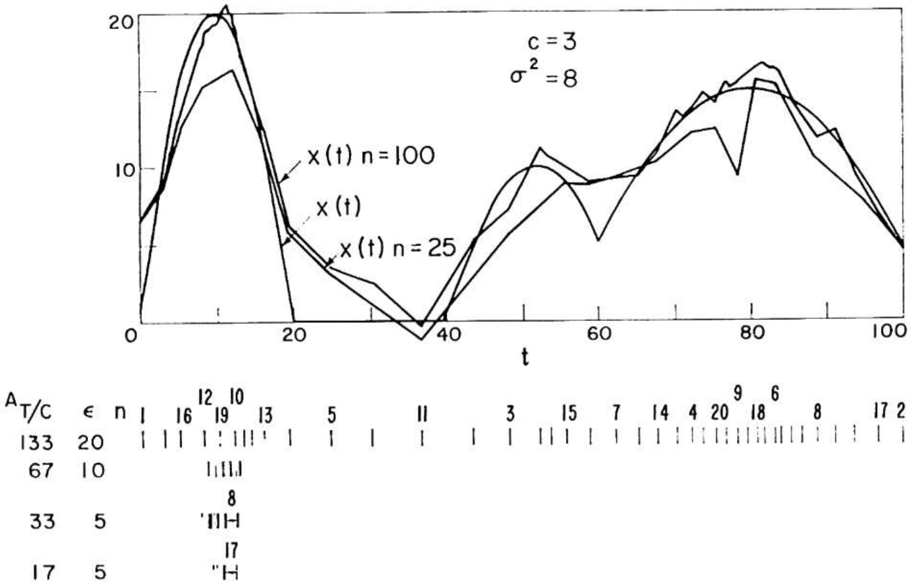

Fig. 7 Experimental results with observation noise

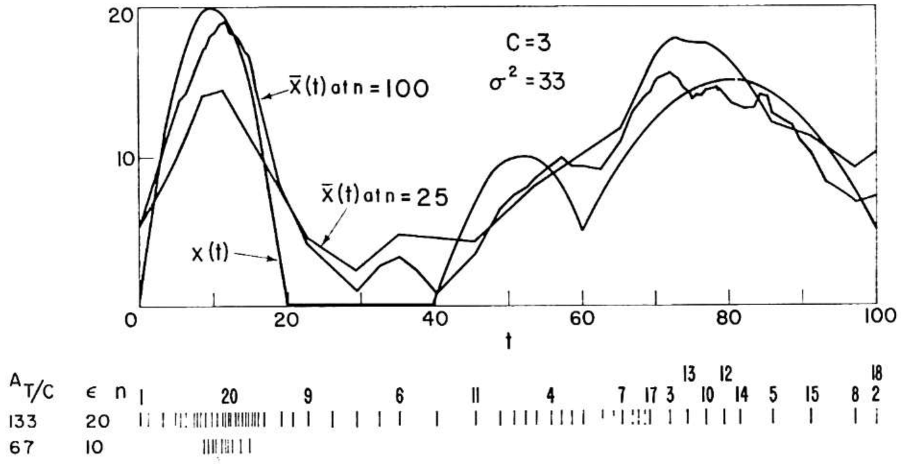

Fig. 8 Experimental results with observation noise

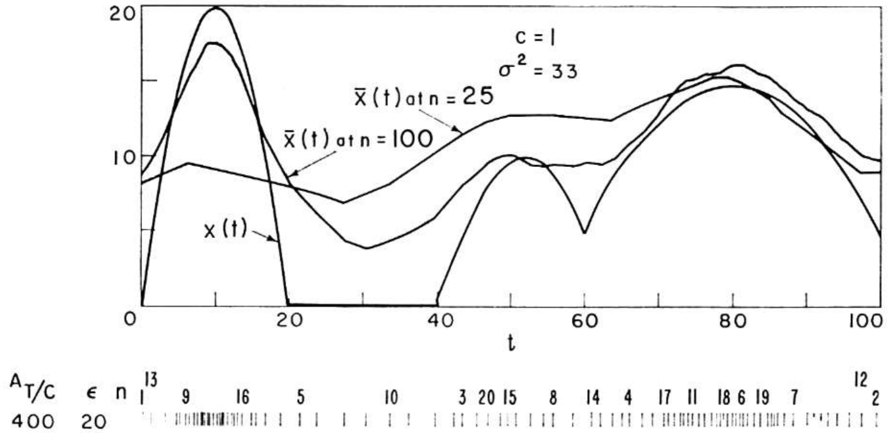

Fig. 9 Experimental results with observation noise

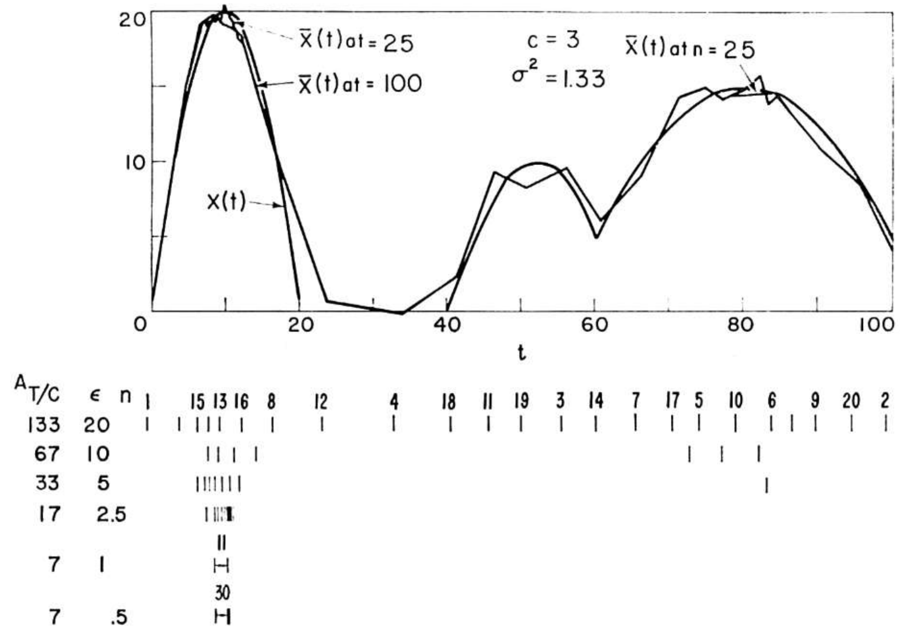

Fig. 10 Experimental results with observation noise

general and the search policy yields a procedure which allows us continually to isolate and explore the more promising regions of parameter space. The approach seems quite promising as a solution to many of the problems of experimental system optimization.

APPENDIX 1

An alternate point of view which supports the logic which led to the form equation (9) will be discussed here. The probability that $X(t) \geq \bar{X}_n^* + \epsilon$, for *some* $t\epsilon I_i$, and no observation noise is given by[14]

$$P_i = \exp - \left\{ \frac{2(\epsilon + \bar{X}_n^* - \bar{X}(t_i))(\epsilon + \bar{X}_n^* - \bar{X}(t_{i+1}))}{cT_i} \right\} \quad (22)$$

The difference between equation (22) and equation (9) is that equation (22) is a property of an interval, while equation (9) is a property of a point.

Since the $n - 1$ events, $X(t) \geq \bar{X}_n^* + \epsilon$ for some $t\epsilon I_i$, are independent, the probability that $X(t) \geq \bar{X}_n^* + \epsilon$ for some t in the sampling interval is given by

$$P_n \equiv 1 - \prod_{1}^{n-1} (1 - P_i) \quad (23)$$

The use of the maximum, over i, of (23) as a criterion for selecting the interval within which the sample is to be taken, leads to the same choice as does equation (10). If the interval selection was made with the use of (23), then the particular point to be sampled within the interval may be selected as the point minimizing the a posteriori probability that there exists an $X(t) \geq \bar{X}_{n+1}^* + \epsilon$. The results attainable with this method are close to the results attainable with the method previously described. The method described here gives a useful probabilistic interpretation to the results. A threshold P_T would be selected instead of A_T, and sampling with a fixed ϵ would continue until $P_n < P_T$. ϵ would then be reduced and a new P_T selected, and so on, as with the first method. The latter method, although closer in keeping with the model, necessitates a choice for c. This can often be done; however, when it is not desirable to do this, one must resort to a logic, such as equation (9), which does not depend on c.

APPENDIX 2

Let $T_i = \tau$, $\alpha = c\tau/\sigma^2$ and define $c_n = b_1 = 1$ and

$$c_j = c_{j+1} + \alpha \sum_{j+1}^{n} c_i \quad (24)$$

[14] Equation (22) is derived in Appendix 3.

$$b_j = b_{j-1} + \alpha \sum_{1}^{j-1} b_i \qquad (25)$$

$$\bar{X}(t_\kappa) = \sum_{1}^{n} A_{ki} Y_i$$

The coefficients A_{ki} are derived elsewhere[5] and, in particular, it is there shown that

$$A_{jj} = b_j \bigg/ \left(\sum_{1}^{j} b_i + \frac{b_j}{c_j} \sum_{j+1}^{n} c_i \right) \qquad (26)$$

$$A_{j+1,\, j} = b_j \bigg/ \left(\sum_{1}^{j+1} b_i + \frac{b_{j+1}}{c_{j+1}} \sum_{j+2}^{n} c_i \right) \qquad (27)$$

Obtaining b_{j+1} and b_j with the use of equation (25) and subtracting, we obtain the difference equation

$$b_{j+1} - b_j(2 + \alpha) - b_{j-1} = 0 \qquad (28)$$

If $\alpha \ll 1$,

$$b_j \rightarrow a(1 + \sqrt{\alpha)^j} = a\lambda^j$$

for some constant a.

Let $n = 2m + 1$ and n be large. Hence, $c_m = b_m$ and $b_{m+1}/c_{m+1} \approx 1$. By symmetry $c_{n-i} = b_{1+i}$. The substitution of these results into equations (26) and (27) yields

$$A_{m+1,\, m} \approx A_{mm} = \lambda^j \bigg/ \left(2 \sum_{1}^{j} \lambda^j \right) \rightarrow \tfrac{1}{2}(c\tau/\sigma^2)^{1/2} \qquad (29)$$

Similarly,

$$A_{n,\, n-1} \approx A_{nn} \rightarrow (c\tau/\sigma^2)^{1/2} \qquad (30)$$

Now an approximation to Var $X(t)$, for small τ or large σ^2, may be obtained for both the center and end intervals (the intervals $t_{m+1} - t_m$ and $t_n - t_{n-1}$, respectively).

Upon substituting equations (29) and (30) into equation (7), we obtain the variance expression

$$\text{Var } X(t) = cr\left(1 - \frac{r}{\tau}\right) + b\sigma^2 \left(\frac{c\tau}{\sigma^2}\right)^{1/2} \qquad (31)$$

where $b = 1/2$ for $t \epsilon I_m$ and $b = 1$ for $t \epsilon I_n$. The curve contribution decreases as τ, and the noise contribution decreases as $\tau^{1/2}$. In general $1/2 \leq b \leq 1$, and approaches 1 or 1/2 as t approaches the center or the endpoint of $(0, T)$, respectively.

In our experiments, when the number of observations taken was large, the true variance as computed by our program was generally close to the value estimated by the second half of equation (31) (with τ equal to the true T_i).

One may use the approximation

$$\frac{(\epsilon + \bar{X}_n{}^* - \bar{X}(t_i))^2}{b\sigma(cT_i)^{1/2}} \qquad (32)$$

to estimate the average T_i for small T_i and large n.

APPENDIX 3

Let the observations $X(0) = 0$ and $X(T) = -B$ be given, and neglect noise effects. Let X^* be the absolute maximum of $X(t)$ on the interval $(0, T)$. We will now compute the probability that $X^* \geq \epsilon \geq 0$.

It is convenient to assume temporarily a discrete random walk model for $X(t)$. Let $0 = t_0 \leq t_1 \leq \ldots \leq t_n = T$ and $X(t_i) = X_i$; then

$$X_k = X_{k-1} + \psi_k$$

$$\psi_k \sim N[0, c(t_k - t_{k-1})]$$

Let v be the first k, if any, for which $X_k \geq \epsilon$.

$$\begin{aligned} P(\max_k X_k \geq \epsilon, X_n < -B) \\ &= \sum_{k=1}^{n-1} P(v = k, X_n < -B) \\ &\leq \sum_{k=1}^{n-1} P(v = k, X_n - X_k < -B - \epsilon) \\ &= \sum_{k=1}^{n-1} P(v = k, X_n - X_k \geq B + \epsilon) \\ &\leq \sum_{k=1}^{n-1} P(v = k, X_n \geq B + 2\epsilon) \\ &\leq P(X_n \geq B + 2\epsilon) \end{aligned} \qquad (33)$$

We converge to the continuous model by letting $t_k - t_{k-1} \rightarrow 0$. It can then be shown that the inequalities in lines 2, 4, and 5 of equation (33) become equalities. Assuming this, we continue and write

$$\begin{aligned} P(\max_t X(t) \geq \epsilon | B \leq X(t) < B + dB) \\ &= \frac{P(\max_t X(t) \geq \epsilon, X(T) < -B + dB) - P(\max_t X(t) \geq \epsilon, X(T) < -B)}{P(-B \leq X(T) < -B + dB)} \\ &= \frac{P(X(T) \geq B + 2\epsilon - dB) - P(X(T) \geq B + 2\epsilon)}{P(-B \leq X(T) < -B + dB)} \end{aligned} \qquad (34)$$

Since $X(t) \sim N(0, ct)$

$$P = \exp - \frac{2\epsilon}{cT}(B + \epsilon) \qquad (35)$$

Acknowledgments

The author expresses his appreciation to Mrs. Jewel Holst, who wrote the programs and contributed many helpful suggestions; to Dr. Fred C. Schweppe, who contributed many helpful suggestions throughout the course of the work; and to Mr. John F Nolan, whose suggestions were extremely useful during the preparation of the final manuscript.