DISCUSSION PAPER NO. 430

OPTIMIZATION OF UNKNOWN, COSTLY OBJECTIVE FUNCTIONS SUBJECT TO A BUDGET CONSTRAINT

by

A.F. Daughety*/

and

M.A. Turnquist**/

July 1980

*/ Northwestern University

**/ Cornell University

The authors are indebted to two colleagues, Sudhakar Deshmukh at Northwestern University and Jery Stedinger at Cornell University, for valuable comments on an earlier draft of this paper.

1. Introduction

This paper presents an approach to non-linear optimization problems in which the objective function is of unknown functional form and expensive to evaluate. There are several types of problems in which the difficulty and expense of evaluating the objective function is a primary issue.

One example of interest is the use of simulation models for system optimization. A typical question facing the user of a simulation is: "What is the optimal operating point for this system?" This is a complicated non-linear optimization problem because the objective function cannot even be written in closed form; it is only evaluated by running the simulation model. This is often quite expensive, both in terms of analyst time and computer time, so the presence of a budget constraint in the optimization is a very real element of the problem.

Most approaches to the optimization of simulation models have been based on direct-search non-linear programming (NLP) algorithms or on response surface methods. A concise summary of procedures is provided by Farrell [4]. Myers [7] discusses the response surface methods in detail, and Smith [10] has done empirical comparison of several methods. Because such procedures were originally developed to be used in an environment where evaluation of the objective function to be optimized is not costly, they tend to require a large number of such evaluations. While the number of such evaluations required in any particular application depends on a number of factors, including how good a starting solution is available, it is not uncommon to require several hundred function evaluations, and thus

these algorithms may not be implementable in a practical situation where each evaluation is a simulation experiment costing hundreds or thousands of dollars.

A second example of the general problem class of interest is a situation in which one wishes to optimize a function whose arguments are the optimal solution(s) to complicated subproblems, themselves requiring an optimization. In this case, evaluating the objective function requires solution of one or more optimization subproblems, which may be quite expensive. A good example of such a problem is a network design problem in transportation systems in which the designer wishes to select optimal capacity additions to a network. The objective function is usually total cost of travel over the network, which is to be minimized. However, this cost is a function of the flows on arcs of the network, and if there are congestion effects, finding these flows for a fixed network is itself an NLP problem. Thus, to evaluate the objective function of the network design problem, one must solve an NLP to assign flow to the network. Problems of this type are discussed by several authors, including Abdulaal and LeBlanc [1].

As with the optimization of simulation models, the presence of a budget constraint for finding the optimal solution is a very real element of this problem, especially for sizable networks. The premise of this paper is that to deal effectively with such optimization problems, the budget constraint must become an integral part of the algorithm proposed for seeking the optimal solution.

The approach suggested here is to adopt a perspective on the problem based on statistical decision theory. The decision to be made by the analyst is the choice of the optimal solution vector (or, as we shall see

later, a characterization of this solution in terms of parameters of an estimated function). Before making this decision, the analyst has the option of performing experiments (evaluations of the objective function) at selected points in a feasible region. These experiments provide additional information about the function, and about the location of the unknown optimal solution. The analyst can choose to perform additional experiments until either: (1) the experimental budget is exhausted, or (2) the expected information gain from an additional experiment is less than the cost of that experiment. The key points in the procedure are the method for approximating the unknown objective function, the means for evaluating performance of a particular additional experiment, and the method of selecting the next experiment from among a set of possible experiments which could be performed.

Because we are approximating the unknown objective function with a fitted function of known form, this method is similar to response surface methodology [7]. However, our perspective on the problem is basically different. Traditional NLP and response surface methods are oriented strictly to finding an optimal solution. Minimizing the computational cost of finding the solution is certainly a secondary objective, but the overriding concern is with finding the solution (at least to within some small error tolerance). What we are suggesting is that in many cases, the analyst may be willing (or may be forced) to settle for a sub-optimal solution, but is interested in finding the best solution possible within a limited experimental budget. This perspective, and the statistical decision methodology used to implement it, differentiate this approach from earlier procedures.

Section 2 of the paper provides a more formal definition of the problem, and describes the procedure for solving it. Section 3 discusses empirical results from testing the procedure on several known functions. Conclusions are given in Section 4.

2. Problem Definition and Solution Procedure

Let $x \in R^n$, $\psi: R^n \to R$. ψ represents the unknown function to be minimized by solving problem (P):

$$(P) \qquad \min_x \ \psi(x).$$

ψ is assumed to be continuous, differentiable, unimodal and complicated in the sense that the cost of computing a value of ψ for a given x is of concern to the analyst. The cost per evaluation, c, is assumed to be constant and known. Furthermore, ψ is assumed to take on values in the same units as c, so that they might be compared (i.e. if c is in dollars then so is ψ).

We assume that the analyst can perform at most $\bar{N}$ evaluations, i.e. he faces a total budget B where:

$$c\bar{N} \leq B < c(\bar{N}+1) .$$

This budget constraint is a very real aspect of solving many problems. When c is very small relative to B we don't observe such a constraint, because in all likelihood it will not be active. However, such a constraint is often binding on the analyst, and in these cases the real objective is to find the best solution possible to (P), within the available budget.

Let x^* be the optimal solution to (P). While it is true that $\nabla\psi(x^*) = 0$, since we do not know what $\psi(x)$ is, we cannot easily use such a first order condition to find the optimum. Instead we will approximate $\psi(x)$ with a function of given functional form and optimize the approximating function. For a particular approximation, this yields a point $\hat{x}$. As we improve the approximation, we will generate a sequence $\{\hat{x}_\tau\}$ converging to x^*.

Let the approximation function $f(x,\alpha)$ be as follows:

$$f(x,\alpha) = \alpha_{oo} + \alpha'_{.o}\, x + \tfrac{1}{2}x'\alpha_{..}x$$

where $\alpha_{oo}\in R$, $\alpha_{.o}\in R^n$ and $\alpha_{..}$ is an nxn symmetric matrix. Thus, the parameter vector α is of dimension k where $k = [n(n+3)/2] + 1$. If $f(x,\alpha)$ is strictly convex in x (i.e. $\alpha_{..}$ positive definite) then its minimizer $\hat{x}(\alpha)$ is a continuous function (from R^k to R^n) of α (see [2]). Given our previous specification of $f(x,\alpha)$, $\hat{x}(\alpha)$ is given by:

$$\hat{x}(\alpha) = -\alpha_{..}^{-1}\,\alpha_{.o}\,.$$

Notice that the pre-image of $\hat{x}(\alpha)$, $\Gamma(x) = \{\alpha\in R^k \mid x=\hat{x}(\alpha)\}$, is non-empty since $\alpha_{.o} = -x$, $\alpha_{..} = I_n$, $\alpha_{oo} = a$, always belongs to $\Gamma(x)$ for any arbitrary value of a. Furthermore, $\Gamma(x)$ is a point-to-set map; α_{oo} is totally irrelevant to the computation of $\hat{x}(\alpha)$, and may be picked arbitrarily. Therefore, there exists $\alpha^*\in\Gamma(x^*)$.

Thus, for some α^*, $f(x,\alpha^*)$ has the same optimal solution as $\psi(x)$. Notice also that for any $\bar{\alpha}\notin\Gamma(x^*)$, the optimal solution to $f(x,\bar{\alpha})$ corresponds to a non-optimal solution to (P). Since $\hat{x}(\alpha)$ is continuous then sequences

of α converging to α^* will generate sequences of x converging to x^*. Thus, the problem of optimizing $\psi(x)$ can be reposed as one of optimizing $\hat{\psi}(\alpha)$, where

$$\hat{\psi}(\alpha) \equiv \psi(\hat{x}(\alpha)),$$

over $\alpha \in R^k$.

At any stage t, let $\hat{\alpha}_t$ be the candidate for α^*. Our objective is to proceed sequentially in driving $\hat{\alpha}_t$ to α^*. Consider the problem of minimizing the following function

$$(P') \qquad \min_{\hat{\alpha}_t} \int (\alpha^* - \hat{\alpha}_t)' A(\alpha^* - \hat{\alpha}_t)\, \xi\, (\alpha^* | z) d\alpha^*$$

where z is the vector of parameters for the density ξ on α^*, and A is a positive definite matrix. The function $(\alpha^* - \hat{\alpha}_t)' A(\alpha^* - \hat{\alpha})$ provides a measure of the penalty incurred if α^* is the optimal parameter value and we choose to stop searching and accept our estimate $\hat{\alpha}_t$ as the optimal parameter value. Minimizing this penalty corresponds to our objective of creating a sequence of $\hat{\alpha}_t$ converging to α^*.

We have thus cast the problem of finding α^* as a statistical decision problem. In the terminology of statistical decision theory, the function $(\alpha^* - \alpha)' A(\alpha^* - \alpha)$ is termed a <u>loss function</u>, and the optimal value of (P') is called the <u>Bayes risk</u>, $\rho(\xi)$, against the distribution ξ [3]. In general, one would like to have A reflect properties of ψ. However, since ψ is unknown, A is selected arbitrarily to be the identity matrix.

In decision theoretic terms, the decision to be made is the value we will use to estimate α^*. We can view the sequence of $\hat{\alpha}$ as being the estimates arising from a linear regression process. Under appropriate assumptions, $\hat{\alpha}_t$ will be the estimate of α^* which minimizes the Bayes risk. In this sense, it will be the optimal decision at stage t. We now discuss the nature of the assumptions under which this is the case.

Let

$$y_t = \psi(x_t) + \eta_t$$

where $x_t = (x_{1t},\ldots,x_{nt})' \in R^n$ and η_t is a normally distributed random variable with zero mean. Note that this is a general formulation which admits both deterministic and stochastic problems. If the function $\psi(x)$ may be observed without error, η_t has zero variance and becomes irrelevant. However, in many situations $\psi(x)$ cannot be observed directly, and we must be satisfied with "noisy" observations.

Define

$$X^t = \begin{bmatrix} 1 & x_{11} & \cdots & x_{n1} & x_{11}^2 & x_{11}x_{21} & \cdots & x_{n1}^2 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1t} & \cdots & x_{nt} & x_{1t}^2 & x_{1t}x_{2t} & \cdots & x_{nt}^2 \end{bmatrix}$$

The data matrix X^t is txk, where $k = [n(n+3)/2] + 1$, and therefore we can express $f(x, \alpha)$ for $x_j \in R^n$ as:

$$f(x_j, \alpha) = X_j\alpha$$

where X_j is the j^{th} row of X_t for $t \geq j$. At stage t we have observations $\{x_\tau, y_\tau\}_{\tau=1}^{t}$ that may be written as a data matrix X^t and dependent observation vector $y^t = (y_1,\ldots,y_t)'$.

We will assume that this constitutes a k-dimensional Normal regression process (see [8], Ch. 13) with parameters (α^*, h) where h is the (unknown) precision of the process. We further assume that (α^*, h) is distributed Normal-gamma with parameters $\hat{\alpha}$, ν, $X'X$ and δ (see [8], p. 343). After t observations:

$$\nu_t = (y^t - X^t\hat{\alpha}_t)' (y^t - X^t\hat{\alpha}_t)/(t-k)$$

$$\hat{\alpha}_t = ((X^t)'X^t)^{-1} (X^t)'y^t$$

and

$$\delta = t-k.$$

The fact that the loss function in (P') is quadratic means that the Bayes risk is [3]:

$$\rho(\xi) = \mathrm{tr}\{A\, E_{y|X}[\mathrm{Cov}(\alpha|X, y)]\}$$

where $\mathrm{Cov}(\cdot)$ denotes the covariance matrix, and $E_{y|X}[\cdot]$ denotes the expectation with respect to the distribution of y given X. Since A = I in our analysis, this may be simplified slightly to

(1) $\rho(\xi) = \mathrm{tr}\{E_{y|X}[\mathrm{Cov}(\alpha|X,y)]\}$.

If α is the result of a Normal regression process, it can be shown that after t observations:

(2) $E_{y^t|X^t}[\mathrm{Cov}(\alpha|X^t, y^t)] = \nu_t[(X^t)'(X^t)]^{-1}$.

Substituting (2) into (1) provides a value for $\rho_t(\xi)$, the Bayes risk after t observations.

To the degree that $\tilde{f}$ approximates $\tilde{\psi}$ correctly, $\rho_t(\xi)$ approximates the expected foregone improvement to ψ if we do not continue with the sampling process. But what if we were to continue? The assumptions made above allow us to compute a pre-posterior risk $\tilde{\rho}(\xi, x)$ which is a function of the next point $x \in R^n$ at which we would sample. This follows the same formula as (1) except that the expected covariance term must be computed based on a pre-posterior analysis (i.e. before observing y_{t+1} in response to x_{t+1}). It can be shown [8] that the pre-posterior expected covariance term is:

$$\Lambda(x_{t+1}) = \{[(X^t)'(X^t)]^{-1} - [(X^t)'(X^t) + x_{t+1}'x_{t+1}]^{-1}\}\frac{(y^t - X^t\hat{\alpha}_t)'(y^t - X^t\hat{\alpha}_t)}{t-k-2} .$$

This is clearly a function of X_{t+1}, the $t+1^{st}$ row of X^{t+1}, which is generated by the point $x_{t+1} \in R^n$. The pre-posterior risk (i.e., the best estimate of the Bayes risk that will be computed after observing (x_{t+1}, y_{t+1})) is $\tilde{\rho}_{t+1}(\xi, x) = \mathrm{tr}[\Lambda(x)]$.

Since $\tilde{\rho}_{t+1}(\xi, x)$ is the pre-posterior risk as a function of the next point to be sampled, we have two obvious, but different possible procedures:

1) pick that point $x \in R^n$ that minimizes $\tilde{\rho}_{t+1}(\xi, x)$ and evaluate the risk at this point;

2) evaluate $\tilde{\rho}_{t+1}(\xi, x)$ at $x = \hat{x}_t = \hat{x}(\hat{\alpha}_t)$.

The first procedure would be optimal if we were employing the correct structural model in our analysis. Of course, in general we are not; we do not assume that ψ is quadratic. The second procedure recognizes that $\hat{x}_t$ embodies useful information since it (presumably) approximates x^*. If the approximation is poor, however, sampling at $\hat{x}_t$ may be worse than employing the first procedure. We shall present empirical results in Section 3 on the relative effectiveness of the alternative procedures.

Thus, let the pre-posterior risk $\tilde{\rho}_{t+1}(\xi)$ be:

$$\tilde{\rho}_{t+1}(\xi) = \begin{cases} \min_{x \in R^n} \tilde{\rho}_{t+1}(\xi, x) \\ \text{or} \\ \tilde{\rho}_{t+1}(\xi, \hat{x}_t) \end{cases}$$

and the t+1 candidate for sampling be:

$$\tilde{x}_{t+1} = \begin{cases} \arg\min_{x \in R^n} \tilde{\rho}_{t+1}(\xi, x) \\ \text{or} \\ \hat{x}_t \end{cases}$$

respectively.

Since $\tilde{\rho}_{t+1}(\xi)$ provides the best estimate of what $\rho_{t+1}(\xi)$ will be after sampling at $\tilde{x}_{t+1}$, then we will only proceed to stage t+1 (i.e. sample at $\tilde{x}_{t+1}$) if

$$(3) \qquad \rho_t(\xi) > \tilde{\rho}_{t+1}(\xi) + c \ ,$$

i.e., if the predicted gain, $\rho_t(\xi) - \tilde{\rho}_{t+1}(\xi)$, exceeds the cost of sampling, c.

If condition (3) fails to be met at some $t < \bar{N}$, then x^* is taken to be $\hat{x}_t = \hat{x}(\hat{\alpha}_t)$. If (3) is met for all $t < \bar{N}$ then x^* is taken to be $\hat{x}_{\bar{N}} = \hat{x}(\hat{\alpha}_{\bar{N}})$ and the difference $\rho_{\bar{N}}(\xi) - \tilde{\rho}_{\bar{N}+1}(\xi) - c$ provides the marginal gain from relaxing the budget constraint enough to allow another experiment.

In summary, the algorithm can be viewed as follows:

STEP 0: Read t_o points $\{(x_\tau, y_\tau)\}_{\tau=1}^{t_o}$ to initiate computations; $t \leftarrow t_o$

STEP 1: Compute $\hat{\alpha}_t$, $\hat{x}_t$

STEP 2: Compute $\rho_t(\xi)$, $\tilde{\rho}_{t+1}(\xi)$, x_{t+1}

STEP 3: If $\rho_t(\xi) - \tilde{\rho}_{t+1}(\xi) > c$ and $t < \bar{N}$ then

a) $t \leftarrow t+1$

b) observe y_t at $x_t = \tilde{x}_t$

c) go to STEP 1

else, stop; $x^* \leftarrow \hat{x}_t$.

3. Testing the Technique

The procedure outlined in Section 2 was tested by applying it to three non-quadratic functions. For all of these experiments, it was assumed that the function values could be observed without error. The functions used were the following:

$$(4) \qquad \psi_1(x_1, x_2) = (1 - 2x_1 - x_1x_2 + \tfrac{3}{2}x_1^2 + \tfrac{1}{2}x_2^2)^2 ,$$

$$(5) \qquad \psi_2(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 ,$$

$$(6) \qquad \psi_3(x_1, x_2, \sigma) = 10(1 - e^{-\frac{1}{2}\psi_1(x_1, x_2)/\sigma^2}) .$$

Equation (4) is a quartic with moderate curvature in comparison to (5), which is the well-known Rosenbrock function [9]. Both functions have a minimal value of 0 at $x = (1,1)$. The Rosenbrock function is particularly difficult due to its rapid functional value change (e.g. $\psi_2(5,-5) = 90,016$) and its peculiar level set properties (see [5], p. 196 for a diagram of this function).

The third function is essentially an upside-down normal density function. The parameter σ will be varied in the tests to provide further information on properties of the proposed technique. This function is not convex; however, it is pseudoconvex. The function resembles a bowl with an extensive "lip". Again the minimum value of zero is attained at $x = (1,1)$.

It is assumed that the analyst can pick a region F over which to place the starting experiments. This region was also used to constrain the minimization of $\tilde{\rho}_{t+1}(\xi, x)$ when the first pre-posterior computation procedure

outlined in Section 2 was employed. The assumption that an analyst can reasonably pick such a region is not very restrictive, since one generally has some feel for where the optimum is likely to be, if only in a vague sense.

The region picked was $[-5, 5] \times [-5, 5]$. Eight initial experimental points were placed at the corners and midpoints of the boundaries of the region, and a ninth point was placed at the center (0, 0). Nine points were used since for $n = 2$ we have $k = 6$, thereby requiring $t > k + 2 = 8$ from the formula for $\Lambda(x)$ given in Section 2 above. Since we wish to compare accuracy of the procedures, the marginal cost of experimentation (c) was set to zero, and the maximum number of experiments, $\bar{N} = 20$.

Table I provides a summary of fifteen experimental runs; the first column indicates the function used. Three types of runs were made. The second column corresponds to computing the pre-posterior risk $\tilde{\rho}_{t+1}(\xi)$ as $\min_{x \in F} \tilde{\rho}_{t+1}(\xi, x)$ (where $F = [-5, 5] \times [-5, 5]$). The third column examines the alternative procedure of evaluating $\tilde{\rho}_{t+1}(\xi, x)$ at $\hat{x}_t$ and performing the next experiment at $\hat{x}_t$. In the fourth column we extend the third column results by restricting the estimation of f to be quasiconvex (see [6]). Ex ante, one would expect this to help the process, by sharpening the specification of the candidate optimum. We see, however, that this is not always true.

Three implications can be drawn from the results in Table I. First, given the typical values of the initial experiments, the minimal amount of structure in the problem, and the very limited experimental budget, all procedures do reasonably well. The values of the starting experimental points

TABLE 1. TEST RESULTS FOR $\bar{N} = 20$

Function	$\tilde{\rho}_{t+1} = \min_{x \in F} \tilde{\rho}_{t+1}(\xi, x)$	$\tilde{\rho}_{t+1} = \tilde{\rho}_{t+1}(\xi, \hat{x}_t)$	$\tilde{\rho}_{t+1} = \tilde{\rho}_{t+1}(\xi, \hat{x}_t)$ (qcx)
ψ_1	(-0.368, -1.3) 5.316	(-0.037, -0.955) 2.24	(0.009, -0.012) 0.965
ψ_2	(-1.18, -11.6) 16884.9	(0.152, 4.503) 2007.8	(0.005, 0.167) 3.778
ψ_3 ($\sigma = 1$)	(0.073, 0.078) 3.49	(0.0006, 0.0006) 3.93	(0.0006, 0.0006) 3.93
ψ_3 ($\sigma = 3$)	(1.25, 3.05) 0.8924	(0.469, 0.711) 0.174	(2.171, 6.663) 4.709
ψ_3 ($\sigma = 5$)	(84.4, 168) 10	(4.26, 7.84) 2.88	(154, 311) 10

Note: Each cell contains $\hat{x}$

$\psi(\hat{x})$

for ψ_1 ranged from 1 to 7396, and those for ψ_2 ranged from 1 to 90016. On the other hand, the typical values for ψ_3 reflected mostly the flat portion of the function, providing little curvature information.

The second implication is that the procedure of sampling at the proposed optimum appears to dominate the procedure of picking the point that minimizes $\tilde{\rho}_{t+1}(\xi, x)$. This reflects two important aspects of the analysis:

1. Since f is an approximation to ψ, $\tilde{\rho}_{t+1}(\xi)$ may be a poor estimate of $\rho_{t+1}(\xi)$ and $\tilde{x}_{t+1}$ may be a poor estimate of the best point to sample at next when the first procedure is used. This is because we do not have the "correct" structural model when we use f. Thus, the regression model is misspecified, violating the assumptions under which $\tilde{\rho}_{t+1}(\xi)$ is the minimum risk.

2. If the optimum-seeking procedure works, in the sense that it generates $\{\hat{x}_\tau\}$ with decreasing ψ values, then clearly each $\hat{x}_t$ is a very worthwhile place at which to sample. Following such a procedure reflects an attempt to incorporate into the search process the added information that $\hat{x}$ embodies.

The third implication concerns the results of column four. Restrictint the estimation of f to be quasiconvex appears to be a useful idea if the function has reasonably strong curvature (e.g. ψ_1, ψ_2, and ψ_3 for $\sigma = 1$). On the other hand, this restriction appears to be counter-productive when the function is "flatter". If we borrow the notions of "peakedness" of a distribution from statistics, it would appear that the more leptokurtic (highly peaked) a function, the more valuable is restricting the estimation, while the more platykurtic the function, the less valuable (and, in fact, possibly counter-productive) the restriction is.

We note that there is no guarantee that any procedure based on drawing finite size samples will be an always improving (i.e. a continuously converging) procedure. The initial selection of experiments can be very influential, since only by taking large random samples might such influence be damped. Clearly, however, the procedure outlined above is not constructing a random sample of $x \in R^n$; rather, $\{x_\tau\}_{\tau=1}^{t}$ is used to compute the next x to be sampled (this is true for both procedures of updating $\tilde{\rho}_{t+1}(\xi)$). Thus, this is a limitation, or sensitivity of the procedure. Of course, various heuristics suggest themselves. For example, one could stop the process if sampling leads to an increase in the observed ψ value, or allow the procedure to proceed as described above and then scan the results for the minimum. Neither of these variations appears justifiable strictly on theoretic grounds, but may be useful in specific situations.

4. Summary and Conclusions

This paper has addressed a class of NLP problems in which the objective function to be optimized is expensive to evaluate. These problems are often ones in which the objective function cannot be expressed in closed form; two examples are optimization of a response from a simulation model, and NLP problems which contain other NLP's imbedded in them as subproblems. In such cases, the available budget with which to conduct a search for the optimum can be of great importance.

A method has been suggested for incorporating this budget constraint directly into the search procedure by posing the problem in the context of statistical decision theory. The analyst constructs an initial estimate of the optimal solution by estimating an approximate response function. This

estimate is then used to help the analyst select a desirable point for the next experiment (evaluation of the objective function), and the information thus obtained is used to update the estimated response surface. This, in turn, leads to a new experimental point, and the process repeats. The process continues in this fashion until the expected gain in information from the next experiment is less than the cost of evaluating the objective function, or until the budget is exhausted, whichever occurs first. The proposed procedure cannot be guaranteed to find the optimal solution, but it is really solving a somewhat different problem, that of finding the best solution possible within an available budget.

If termination is due to the budget constraint, the expected value of sample information from the next potential (but unperformed) experiment provides sensitivity information on the value of relaxing the budget constraint. This is useful in a practical context, because it provides the analyst with an idea of how valuable one additional experiment might be. Empirical tests of the general method on three known functions have indicated that it performs effectively. Two specific procedures have been tested, involving different criteria for selecting the next experiment. The procedure of sampling at the indicated optimum of the current response function seems to dominate the procedure of minimizing expected loss over some feasible region. Furthermore, if one expects significant peakedness, restricting the approximating function to be quasiconvex appears to be very useful.

The previous discussion leads to a general conclusion: the procedure appears to be good in the "large", but not the "small". In other words, the procedure is effective at finding an estimate of the optimum which is good relative to most other points in a large feasible region (this is the notion of being good in the large). The procedure does not appear,

however, to be very effective in improving on a reasonably good existing solution. This is reflected by the algorithm's performance on the (relatively) platykurtic functions ψ_3 ($\sigma = 3$) and ψ_3 ($\sigma = 5$). In a crude sense, most unimodel (differentiable) functions tend to be relatively platykurtic local to their optimum and comparitively leptokurtic when viewed far from their optimum.

Thus, in situations wherein we might expect some considerable peakedness to the response surface, and where we have only vague information about the optimum, this procedure appears to be very useful and effective. On the other hand, for situations wherein one would expect a relatively flat response surface, or know (with reasonable precision) an estimate of the optimum, an alternative method might work better. This is, however, consistent with the motivation of our analysis, namely that our objective is to get a reasonable estimate of the optimum (i.e. a point with low objective function value) within a cost constraint.

Finally, this also suggests that the outlined procedure could be useful in finding a good starting point for standard NLP algorithms. Since the procedure can be readily constrained to use only a small number of function evaluations, and appears to be effective in the large, it could provide a useful complement to algorithms which tend to be more effective in the small.

REFERENCES

1. Abdulaal, M. and LeBlanc, L.J., "Continuous Equilibrium Network Design Models," Transportation Research, 138, 1979, pp. 19-32.
2. Berge, C., Topological Spaces, Oliver & Boyd, London, 1963.
3. DeGroot, M.H., Optimal Statistical Decisions, McGraw-Hill, New York, 1970.
4. Farrell, W., "Literature Review and Bibliography of Simulation Optimization," Proceedings of the 1977 Winter Simulation Conference, Gaithersburg, Maryland, pp. 116-125.
5. Himmelblau, D.M., Applied Nonlinear Programming, McGraw-Hill, New York, 1972.
6. Lau, L.J., "Testing and Imposing Monotonicity, Convexity and Quasi-convexity Constraints" in Production Economics: A Dual Approach to Theory and Applications, edited by M. Fuss and D. McFadden, North-Holland Publishing Co., New York, 1978, pp. 409-453.
7. Myers, R.H., Response Surface Methodology, Allyn and Bacon, Boston, 1971.
8. Raiffa, H. and Schlaifer, R., Applied Statistical Decision Theory, MIT Press, Cambridge, 1968.
9. Rosenbrock, H.H., "An Automatic Method for Finding the Greatest or Least Value of a Function," Computer Journal, 3, 1960, pp. 175-184.
10. Smith, D.E., "An Empirical Investigation of Optimum-Seeking in the Computer Simulation Situation," Operations Research, 21, 1973, pp. 475-497.