Proceedings of the 1995 Winter Simulation Conference
ed. C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman

SAMPLE SIZE SELECTION FOR IMPROVED NELDER-MEAD PERFORMANCE

John J. Tomick
Department of Mathematical Sciences
U.S. Air Force Academy
Suite 6D2A
Colorado Springs, CO 80840, U.S.A.

Steven F. Arnold
Department of Statistics
The Pennsylvania State University
University Park, PA 16802 U.S.A.

Russell R. Barton
Department of Industrial and
Manufacturing Engineering
The Pennsylvania State University
University Park, PA 16802 U.S.A.

ABSTRACT

The Nelder-Mead simplex algorithm has been used for sequential optimization of simulation response functions. The rescaling operations of this algorithm can lead to inappropriate termination at non-optimal points. We have used the probabilistic characterization of this behavior to develop special rules for determining the number of replications to take for each experimental design point. Computational experiments indicate that the quality of the solution is often improved.

1 INTRODUCTION

Complex discrete-event simulation models of proposed or existing real systems are often used to estimate the effects on system performance due to changes to the system design. A natural extension of this evaluative use of simulation is optimization: to look for a system design that produces the optimum system performance. In optimization studies, a numerical output of the simulation is selected as the objective for optimization. For discrete-event simulation models, this objective is usually derived from pseudorandom quantities, and so optimization algorithms designed for deterministic functions are often ineffective.

Mathematically, the optimization problem is to

$$\text{minimize } E(F(x)), \qquad x \in R^n, \tag{1}$$

where F is the stochastic response function of a simulation model. The stochastic response can be written as

$$F(x) = f(x) + \varepsilon(x), \tag{2}$$

where $f(x)$ is the deterministic function $E(F(x))$ and $\varepsilon(x)$ is a stochastic function with $E(\varepsilon(x)) = 0$ for all x. Then the optimization problem is to

$$\text{minimize } f(x), \qquad x \in R^n. \tag{3}$$

This paper focuses on the Nelder-Mead method (Nelder and Mead 1965), which was originally developed for unconstrained optimization of deterministic functions, but which has been applied frequently to the optimization of stochastic simulation models. The method is somewhat effective because the algorithm's steps depend only on the relative ranks of the objective function values for different system designs, rather than on the precise values. Many other optimization methods are available for simulation optimization. See for example the recent survey articles by Azadivar (1992), Fu (1994) and Spall (1994).

Barton and Ivey (1995) observed inappropriate early termination of optimization for the original Nelder-Mead method. We examine modifications that make Nelder-Mead more effective when applied to stochastic responses.

The optimization is posed as a minimization problem. We will call f the *objective function* and F the *response function*. Typically f is not known explicitly, and the optimization method must work with F. For stochastic simulation optimization, the response function is computed from the output quantities of one or many replications of the simulation.

2 THE NELDER-MEAD ALGORITHM

The Nelder-Mead (1965) simplex method incorporates operations to rescale the simplex based on the local behavior of the function. Simplex reflections are expanded in the same direction if the reflected value is particularly good. A poor value results in a contraction. If the function value at the contracted point is poorer yet, the overall size of the simplex is shrunk. The original Nelder-Mead rules are outlined below.

Step 1. Initialization: *For a function of n parameters, choose n+1 extreme points to form an initial n-dimensional simplex. Evaluate the response function $F(x_i)$ at each point (vertex) x_i of the simplex for $i = 1, 2, \ldots, n+1$.*

Step 2. Stopping Criterion: *Iterations continue until the standard deviation of the n+1 response function values at the extreme points of the simplex,* $S_F \equiv [\sum (F(x_i) - \bar{F})^2 / (n + 1)]^{1/2}$, *with* $\bar{F} \equiv \sum F(x_i) / (n + 1)$, *falls below a particular value, or when the size of the simplex becomes sufficiently small, or until the maximum number of function evaluations is reached.*

Step 3. Reflect Worst Point: *At the start of each iteration, identify the vertices where the highest, second highest, and lowest response function values occur. Let* P_{high}, P_{sechi}, P_{low} *respectively denote these points, and let* F_{high}, F_{sechi}, F_{low} *respectively represent the corresponding observed function values. Find* P_{cent}, *the centroid of all points other than* P_{high}. *Generate a new point* P_{refl} *by reflecting* P_{high} *through* P_{cent}. *Reflection is carried out according to the following equation, where* α *is the reflection coefficient* ($\alpha > 0$):

$$P_{refl} = (1 + \alpha)P_{cent} - \alpha P_{high}.$$

Nelder and Mead used $\alpha = 1$.

Step 4a. Accept Reflection: *If* $F_{low} \le F_{refl} \le F_{sechi}$, *then* P_{refl} *replaces* P_{high} *in the simplex, and a new iteration begins (Step 2 above).*

Step 4b. Attempt Expansion: *If* $F_{refl} < F_{low}$, *then the reflection is expanded, in the hope that more improvement will result by extending the search in the same direction. The expansion point is calculated by the following equation, where the expansion coefficient is* γ ($\gamma > 1$):

$$P_{exp} = \gamma P_{refl} + (1 - \gamma)P_{cent}.$$

Nelder and Mead used $\gamma = 2$. *If* $F_{exp} < F_{low}$, *then* P_{exp} *replaces* P_{high} *in the simplex; otherwise, the expansion is rejected and* P_{refl} *replaces* P_{high}. *The next iteration begins with the new simplex (Step 2 above).*

Step 4c. Attempt Contraction: *If the reflected point has the worst response function value in the new simplex, (i.e.,* $F_{refl} > F_{sechi}$*) then the simplex contracts. If* $F_{refl} \le F_{high}$, *then* P_{refl} *replaces* P_{high} *and* F_{refl} *replaces* F_{high} *before attempting contraction or shrinking. The contraction point is calculated by the following equation, where the contraction coefficient is* β ($0 < \beta < 1$):

$$P_{cont} = \beta P_{high} + (1 - \beta)P_{cent}.$$

Nelder and Mead used $\beta = 0.5$. *If* $F_{cont} \le F_{high}$, *then contraction is accepted and the algorithm continues with the next iteration (Step 2 above).*

Step 4c'. Shrink: *If* $F_{cont} > F_{high}$, *then the contraction has failed, and the entire simplex shrinks by a factor of* δ ($0 < \delta < 1$), *retaining only* P_{low}. *This is done by replacing each extreme point* P_i *(except* P_{low}*) by:*

$$P_i \leftarrow \delta P_i + (1 - \delta)P_{low}$$

Nelder and Mead used $\delta = 0.5$. *The algorithm then evaluates F at each point (except* P_{low}*) and continues with the next iteration (Step 2 above).*

One of the following stopping criteria are usually employed. Nelder and Mead computed the standard deviation of the (deterministic) objective function values over all n+1 extreme points, and they stopped when the standard deviation S_f dropped below 10^{-8}, where:

$$S_f \equiv [\sum (f(x_i) - \bar{f})^2 / (n + 1)]^{1/2},$$

with $\bar{f} \equiv \sum f(x_i) / (n + 1)$ For stochastic functions, the standard deviation of function values across all simplex points reflects inherent stochastic variation as well as differences in (expected) function values. For stochastic function optimization, the calculations are based on F rather than f, and the stochastic component ε in equation (2) will typically have a standard deviation $\sigma_\varepsilon \equiv [E(\varepsilon^2)]^{1/2}$ much greater than 10^{-8}, making this rule inappropriate. A stopping criterion based on simplex size was proposed by Dennis and Woods (1987). The stopping criterion is

$$(1/\Delta) \max_i \|P_i - P_{low}\| \le \nu, \quad \Delta = \max (1, \|P_{low}\|) \qquad (4)$$

where the maximization is over all points i in the current simplex, and $\| \bullet \|$ denotes the Euclidean norm.

2.1 Inappropriate Termination

Using the Dennis and Woods criterion with $\nu = 1 \times 10^{-4}$, Barton and Ivey (1995) found that the Nelder-Mead method could terminate at a point that was far from the optimum for some response functions. This problem was further investigated by Tomick (1995), who found that for n = 1 or 2, Nelder-Mead (without any stopping criterion) converged to a point on a constant test function, with finite expected movement.

Further, for n = 1, Tomick proved that there is a nonzero probability of convergence of Nelder-Mead for linear f when $\varepsilon(x)$ in (2) has a nondegenerate Gaussian distribution. Empirical tests indicated a nonzero probability of convergence for $n > 1$ as well. For any given standard deviation σ_ε for ε, the probability of false convergence decreases as the magnitude of the slope increases. Of course, convergence on a linear function is false convergence.

2.2 Barton and Ivey's RS9 Modification

The shrink step (Step 4c') causes rapid decrease in the overall size of the simplex. Barton and Ivey (1995) found that changing the shrinkage coefficient from $\delta = 0.5$ to $\delta = 0.9$ generally improved the performance of the method on stochastic functions. In addition, the original algorithm did not require resampling the simulation response at the best point of the simplex after a shrink step. Resampling this point tended to reduce the likelihood of retaining a spuriously good response, and improved performance. The modified method implementing both of these changes was referred to as *RS9*.

2.3 Hypothesis Test Modifications

Based on a Markov chain analysis for the $n = 1$ case, Tomick (1995) proposed additional modifications to *RS9*:

- replace $\delta = 0.5$ with $\delta = 0.9$ in Step 4c.
- compute the response value at a vertex in the kth iteration as the average of m^k replications of the simulation, where m^k is defined as described below.

Tomick proposed two methods to choose the number of replications m^k: using standardized range distributions, or using one-way analysis of variance (ANOVA). The ANOVA method for setting m^k tests the hypothesis that the response means (i.e., objective function values) at the $n + 1$ vertices are equal. If the test is accepted, the sample size is increased by a factor b:

$$m^{k+1} = \lfloor b m^k \rfloor \quad \text{if } (S^2/n) / (\sigma^2) \le \chi^2_{n,\alpha}$$

and if the test is rejected, the sample size is decreased by the same factor:

$$m^{k+1} = \lfloor m^k / b \rfloor \quad \text{if } (S^2/n) / (\sigma^2) > \chi^2_{n,\alpha}$$

The mean square for treatments from the ANOVA is used for S^2. Tomick called this modified method *NMSNV*. For the empirical comparisons below, b was set to 1.25.

3 EMPIRICAL PERFORMANCE OF THE ANOVA MODIFICATION

To compare results with Barton and Ivey (1995) our tests employed a set of eighteen deterministic functions compiled by Moré, Garbow, and Hillstrom (1981). The form, standard starting points, and optimal values for the functions are described completely in that reference, and are available via the MINPACK collection in NETLIB (Dongarra and Grosse 1987). They are a set of deterministic objective functions for difficult unconstrained minimization problems. The number and variety of functions allows a good assessment of robustness with a reasonable investment in computation time. Many of these functions allow a choice of dimension: we chose values to provide test functions with input parameter dimensions that ranged from 2 to 9.

The starting points for our tests were not the standard starting points, however. For the results discussed below, starting points were selected to provide an initial objective function that was $10\sigma_\varepsilon$ larger than the optimal value. These starting values are summarized in Table 1. The computational tests included forty optimization runs with starting points perturbed by adding a uniform(-0.1, 0.1) deviate to each coordinate. Computational comparisons using other starting points are discussed in Tomick (1995).

Table 1: Starting Points for Test Functions

Function	Starting Point
Helical Valley	(5, 25, -17.74)
Biggs EXP6	(10, -2, 20, -4.9, -1.5, 4.9)
Gaussian	(6.28, -0.1, -5)
Powell Badly Scaled	(0.01, 3.2)
Box 3D	(-5.5, 4, -20)
Variably Dim.	$x_j = (4 - j/n)(-1)^{j+1}$
Watson	$x_j = -1.32$
Penalty I	$x_j = 1.25j$
Penalty II	$x_j = 3$
Brown Badly Scaled	(9.999E+05, 5.0E-06)
Brown & Dennis	(-8, 11, -5, 0)
Gulf R&D	(-0.95, 1, 0.4)
Trigonometric	$x_j = 0.71j/8$
Extended Rosenbrock	$x_j = 4.4(-1)^{j+1}$
Extended Powell	(3, -9, 1.5, 10, 3, -9, 1.5, 10)
Beale	(2.5, 6)
Wood	(-5, -2, -5, 7)
Chebyquad	$x_j = 0.1j + 0.34$

Table 2 shows the performance of the original Nelder-Mead method (*NM*), *RS9*, and *NMSNV*. Also included in the test was *KW*, a naive implementation of the stochastic approximation method described by Kiefer and Wolfowitz (1952). The numbers shown are the average gap between the objective function value at the best simplex point after 1,000 function evaluations and the true optimum. The numbers are expressed in units of σ_ε. (remember, the starting point had a gap of $10\sigma_\varepsilon$). The underlined number in each row indicates the best average performance for that test function.

While Barton and Ivey's *RS9* performance is generally superior to *NM*, *NMSNV* is superior to *RS9* for all but two functions in this set of tests. This was not uniformly true for other starting points (see Tomick 1995). The Keifer-Wolfowitz method was occasionally superior to *NMSNV*, but more often the reverse was the case. Furthermore, for seven of the eighteen test functions, KW diverged drastically from the optimal value, and remained so at 1,000 evaluations (or even 10,000 - see Tomick 1995). These errors were generally eighty or more orders of magnitude larger, and occasionally led to termination due to numerical overflow. Similar difficulties with Kiefer-Wolfowitz occurred for other starting points as well.

The results differed when the progress was examined after only 100 function evaluations. At this early stage of optimization, *KW* was superior, except on those functions where it initially diverged. For functions where *KW* failed, *RS9* was generally superior after 100 function evaluations.

The *NMSNV* method's relative superiority increased further after 10,000 function evaluations. Its overall superiority increased from 11 to 13 of the 18 functions (see Tomick 1995).

Table 2: Error After 1000 Evaluations

Test	Method			
Func.	*NM*	*RS9*	*NMSNV*	*KW*
1	9.85	8.63	1.48	8.06
2	0.54	0.30	0.14	0.01
3	0.10	0.06	0.07	0.001
4	0.04	0.04	0.01	-
5	0.54	0.27	0.17	0.12
6	0.46	0.06	0.01	-
7	0.22	0.06	0.03	0.05
8	0.91	0.62	0.12	2.01
9	2.32	1.03	0.62	0.22
10	0.002	0.002	0.003	-
11	3.99	1.75	1.21	-
12	0.01	0.006	0.004	-
13	0.16	0.11	0.02	1,370
14	0.63	0.31	0.24	0.30
15	5.59	2.98	1.10	1.25
16	0.01	0.07	0.002	-
17	0.37	0.28	0.20	0.18
18	0.02	0.0004	0.004	-

4 CONCLUSION

Simulation optimization requires optimization techniques that are designed for stochastic responses. The original Nelder-Mead unconstrained optimization method was not designed for stochastic responses, and consequently the method can terminate at an inappropriate point.

Two simple modifications to Nelder-Mead, proposed by Barton and Ivey (1995), often delay the onset of difficulties for the method. The modified method, *RS9* often provided objective functions values less than one standard deviation from the optimal value in relatively few function evaluations, but more accuracy was not possible without making replications at each simplex vertex.

A new method, *NMSNV* was developed to choose the number of replications dynamically as the optimization progresses. *NMSNV* gave no indication of behavior that would lead to inappropriate termination in empirical testing. It was able to reduce the error to less than 20% of the initial value on all of the 18 test functions (at lest after 1,000 evaluations), which was not the case for *NM*, *RS9*, or *KW*.

ACKNOWLEDGMENTS

This research was supported in part by the U. S. Air Force and by NSF grant DDM-9308492.

REFERENCES

Azadivar, F. 1992. A tutorial on simulation optimization. *Proceedings of the 1992 Winter Simulation Conference*, ed. J.J. Swain, D. Goldsman, R.C. Crain, and J.R. Wilson, IEEE, Piscataway, NJ, 198-204.

Barton, R. R. and J. S. Ivey. 1995. Nelder-Mead simplex modifications for simulation optimization. Accepted for publication in *Management Science*.

Dongarra, J. J. and E. Grosse. 1987. Distribution of mathematical software via electronic mail. *Communications of the ACM* 30, 403-407.

Fu, M. C. 1994. A tutorial review of techniques for simulation optimization. *Proceedings of the 1994 Winter Simulation Conference*, ed. J.D. Tew, S. Manivannan, D.A. Sadowski, and A.F. Seila, IEEE, Piscataway, NJ, 149-156.

Kiefer, J. and J. Wolfowitz. 1952. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics* 23, 462-466.

Moré, J. J., B. S. Garbow and K. E. Hillstrom. 1981. Testing unconstrained optimization software. *ACM Transactions on Mathematical Software* 7, 17-41.

Nelder, J. A. and R. Mead. 1965. A simplex method for function minimisation. *The Computer Journal* 7, 308-313.

Spall, J. C. 1994. Developments in stochastic optimization algorithms with gradient approximations based on function measurements. *Proceedings of the 1994 Winter Simulation Conference*, J.D. Tew, S. Manivannan, D.A. Sadowski, and A.F. Seila (Eds.), IEEE, Piscataway, NJ, 207-214.

Tomick, J. J. 1995. On convergence of the Nelder-Mead algorithm for unconstrained stochastic optimization. Ph.D. Thesis, Department of Statistics, The Pennsylvania State University.

AUTHOR BIOGRAPHIES

STEVEN F. ARNOLD is Associate Professor of Statistics at The Pennsylvania State University. He received a B.A. in mathematics from Knox College in 1966 and a Ph.D. in statistics from Stanford University in 1970. His current research interest is statistical inference, both parametric and non-parametric, for repeated measures models. Other interests include the study of invariance, the use of Gibbs sampling and applied probability.

JOHN J. TOMICK is a Captain in the U. S. Air Force, and is serving as an Assistant Professor in the Department of Mathematical Sciences at the U. S. Air Force Academy. He received a B.S. Degree in Mathematical Sciences from the U. S. Air Force Academy in 1984 and an M.S. Degree in Operations Research from the U. S. Air Force Institute of Technology in 1988. He received his Ph.D. in Statistics and Operations Research from The Pennsylvania State University in 1995. His research interests are in the design and analysis of simulation optimization algorithms.

RUSSELL R. BARTON is an Associate Professor in the Department of Industrial and Manufacturing Engineering at The Pennsylvania State University. He received a B.S. in electrical engineering from Princeton University in 1973 and a Ph.D. in operations research from Cornell University in 1978. His current research interests include graphical methods for experiment design, optimization and metamodeling methods for simulation models, and statistical issues in modeling manufacturing yield.