ABSTRACT

Simulation optimization can be defined as the process of finding the best input variable values from among all possibilities without explicitly evaluating each possibility. The objective of simulation optimization is to minimize the resources spent while maximizing the information obtained in a simulation experiment. The purpose of this paper is to review the area of simulation optimization. A critical review of the methods employed and applications developed in this relatively new area are presented and notable successes are highlighted. Simulation optimization software tools are discussed. The intended audience is simulation practitioners and theoreticians as well as beginners in the field of simulation.

1 INTRODUCTION

When the mathematical model of a system is studied using simulation, it is called a simulation model. System behavior at specific values of input variables is evaluated by running the simulation model for a fixed period of time. A simulation experiment can be defined as a test or a series of tests in which meaningful changes are made to the input variables of a simulation model so that we may observe and identify the reasons for changes in the output variable(s). When the number of input variables is large and the simulation model is complex, the simulation experiment may become computationally prohibitive. Besides the high computational cost, an even higher cost is incurred when sub-optimal input variable values are selected. The process of finding the best input variable values from among all possibilities without explicitly evaluating each possibility is simulation optimization. The objective of simulation optimization is minimizing the resources spent while maximizing the information obtained in a simulation experiment.

Consider an intersection with four-way stop signs. Suppose that it has been determined to be a bottleneck for traffic during rush hours and a decision has been made to replace it with traffic lights in all four directions. Now, the problem is to determine the optimal green-times in all directions in order to ensure the least wait-time for cars arriving from all directions. The green-times must allow for the flow of traffic in all the pre-specified directions. Given the rates at which cars arrive from every direction, and picking reasonable green-time values, a simulation model of the intersection can be developed and run to determine the wait-time statistics. Running of this simulation model can only provide the answer to a “what if” question (e.g., What if the green-time in all four directions is 2 minutes ... what is the resulting wait-time for cars arriving at the intersection?). But how do we find the optimal values for the various green-times? This is an instance of a simulation optimization problem.

A general simulation model comprises $n$ input variables $(x_1, x_2, ..., x_n)$ and $m$ output variables $(y_1, y_2, ..., y_m)$ (Figure 1).

Simulation optimization entails finding optimal settings of the input variables, i.e., values of $x_1, x_2, ..., x_n$, which optimize the output variable(s). Such problems arise frequently in engineering, for instance, in process design, in industrial experimentation, in design optimization, and in reliability optimization. This is the problem we will address in this paper. A simulation optimization model is displayed in Figure 2. The output of a simulation model is used by an optimization strategy to provide feedback.
on progress of the search for the optimal solution. This in turn guides further input to the simulation model.

Feedback on progress

![Simulation Optimization Model](image)

Figure 2: A Simulation Optimization Model

Simulation optimization is an area that has attracted the attention of many researchers. The six major categories of simulation optimization methods are displayed in Figure 3. Section 2 contains brief descriptions of frequently used simulation optimization methods. Section 3 enumerates the reported applications of simulation optimization. Section 4 lists computer software that integrates simulation and optimization. Section 5 entails the closing remarks. The techniques and applications presented here are neither mutually exclusive nor exhaustive. All of the techniques discussed in this paper have been previously investigated in simulation literature. Our main contribution here is to provide knowledge about the area of simulation optimization in one paper, with an extensive reference list pointing to detailed treatment of specific techniques.

2 SIMULATION OPTIMIZATION METHODS

2.1 Gradient Based Search Methods

Methods in this category estimate the response function gradient ($\nabla f$) to assess the shape of the objective function and employ deterministic mathematical programming techniques. Frequently used gradient estimation methods are described below.

2.1.1 Finite Differences

Partial derivatives of the output variable $f(x)$ are estimated by the following:

$$\frac{\partial f}{\partial x_i} = \frac{f(x_1, \ldots, x_i + \Delta_i, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{\Delta_i}$$

To estimate the gradient at a specific value of $x$, at least $n+1$ configurations of the simulation model must be run. To obtain a more reliable estimate of $\nabla f$ there may be a need for multiple observations for each partial derivative, further increasing the already high computational cost. Finite differences is the crudest method of estimating the gradient (Azadivar, 1992). For a comprehensive discussion on the convergence properties of, and bias and variability of gradient estimators obtained with finite differences, likelihood ratios (Section 2.1.2 below), and perturbation analysis (Section 2.1.3 below), see Glynn (1989b).

2.1.2 Likelihood Ratios (LR)

In the likelihood ratio method, also called the score function, the gradient of the expected value of an output variable with respect to an input variable is expressed as the expected value of a function of a) input parameters, and b) simulation parameters e.g. simulation run length, output variable value etc. For instance, for a Poisson process with rate $\lambda$, if $N_T$ is the number of events in time interval $(0, T)$, and $y$ is an output variable, then

$$\frac{\partial}{\partial \lambda} E(y) = E \left[ \left( \frac{N_T}{\lambda} - T \right) y \right]$$

The right hand side of the equation above can be computed by keeping track of a statistic during a simulation run. Better estimates can be obtained by conducting multiple simulation runs. This method is suitable for transient and regenerative simulation optimization problems. For a regenerative process, steady state value of an output variable can be expressed...
as a ratio of two expected values - the likelihood ratio. The construction of a likelihood ratio that has desirable computational and variability characteristics is an important issue in the development of LR gradient estimators (Glynn, 1989b). LR methods are discussed in Glynn (1989a), and Reiman and Weiss (1986).

2.1.3 Perturbation Analysis (PA)

In infinitesimal perturbation analysis (IPA) all partial gradients of an objective function are estimated from a single simulation run. The idea is that in a system, if an input variable is perturbed by an infinitesimal amount, the sensitivity of the output variable to the parameter can be estimated by tracing its pattern of propagation. This will be a function of the fraction of propagations that die before having a significant effect on the response of interest. IPA assumes that an infinitesimal perturbation in an input variable does not affect the sequence of events but only makes their occurrence times slide smoothly. The fact that all derivatives can be derived from a single simulation run, represents a significant advantage in terms of computational efficiency. On the other hand, the estimators derived using IPA are often biased and inconsistent. According to Glynn (1989b), when both IPA and LR methods apply to a given problem, the IPA gradient estimator is more efficient. Other PA methods include smoothed perturbation analysis (SPA), and IPA variants. In a relatively short time since the introduction of PA to the simulation optimization field, a significant volume of work on this topic has been reported (see, e.g., Suri (1983), and Ho and Cao (1991)). For a comparative study of finite differences, LR and IPA, see L’Ecuyer (1991).

2.1.4 Frequency Domain Method (FDM)

A frequency domain experiment is one in which selected input parameters are oscillated sinusoidally at different frequencies during one long simulation run. The output variable values are subjected to spectral (Fourier) analysis, i.e. regressed against sinusoids at the input driving frequencies (Morrice and Schruben, 1989). If the output variable is sensitive to an input parameter, the sinusoidal oscillation of that parameter should induce corresponding (amplified) oscillations in the response.

Frequency domain experiments involve addressing three questions: how does one determine the unit of the experimental or oscillation index, how does one select the driving frequencies, and how does one set the oscillation amplitudes? These questions have been addressed in Jacobson et al. (1988), and Jacobson (1989). Frequency domain methodology was first introduced as a screening tool for continuous input factors in discrete-event simulations in Schruben and Cogliano (1987). Jacobson and Schruben (1988) extended the approach to gradient direction estimation.

2.2 Stochastic Optimization

Stochastic optimization is the problem of finding a local optimum for an objective function whose values are not known analytically but can be estimated or measured. Classical stochastic optimization algorithms are iterative schemes based on gradient estimation. Proposed in the early 1950s, Robbins-Monro and Kiefer-Wolfowitz are the two most commonly used algorithms for unconstrained stochastic optimization. These algorithms converge extremely slowly when the objective function is flat and often diverge when the objective function is steep. Additional difficulties include absence of good stopping rules and handling constraints. More recently, Andradottir (1990) proposed a stochastic optimization algorithm that converges under more general assumptions than these classical algorithms. Leung and Suri (1990) reported better results with the Robbins-Monro algorithm when applied in a finite-time single-run optimization algorithm than when applied in a conventional way.

In the stochastic counterpart method (also known as sample path optimization) a relatively large sample is generated and the expected value function is approximated by the corresponding average function (Shapiro (1996), Gurkan et al. (1994)). The average function is then optimized by using a deterministic non-linear programming (LP) method. This allows statistical inference to be incorporated into the optimization algorithm which addresses most of the difficulties in stochastic optimization and increases the efficiency of the method.

2.3 Response Surface Methodology (RSM)

Response surface methodology is a procedure for fitting a series of regression models to the output variable of a simulation model (by evaluating it at several input variable values) and optimizing the resulting regression function. The process starts with a first order regression function and the steepest ascent/descent search method. After reaching the vicinity of the optimum, higher degree regression functions are employed. Applications of RSM in simulation optimization are described in Biles (1974) and Daugherty and Turnquist (1980). Using a criterion which considers the bias as well as the variance of the simulation response variable, Donohue et al. (1990) developed optimal designs in common second-order design classes including central composite, Box-Behnken, and full-factorial. In general, RSM requires a smaller number of simulation experiments relative to many gradient based methods.
2.4 Heuristic Methods

Heuristic methods discussed below represent the latest developments in the field of direct search methods (requiring only function values) that are frequently used for simulation optimization. Many of these techniques balance exploration with exploitation thereby resulting in efficient global search strategies.

2.4.1 Genetic Algorithms (GA)

A genetic algorithm is a search strategy that employs random choice to guide a highly exploitative search, striking a balance between exploration of the feasible domain and exploitation of “good” solutions (Holland, 1992). This strategy is analogous to biological evolution. From a biological perspective, it is conjectured that an organism’s structure and its ability to survive in its environment (“fitness”), are determined by its DNA. An offspring, which is a combination of both parents’ DNA, inherits traits from both parents and other traits that the parents may not have, due to recombination. These traits may increase an offspring’s fitness, yielding a higher probability of surviving more frequently and passing the traits on to the next generation. Over time, the average fitness of the population improves.

In GA terms, the DNA of a member of a population is represented as a string where each position in the string may take on a finite set of values. Each position in the string represents a variable from the system of interest, e.g. a string of five input switches on a black box device where each switch may take the value 1 (switch is on) or 0 (switch is off); the string 11100 indicates the first three switches to be on and the last two switches to be off. The fitness of a member of a population is determined by an objective function.

Members of a population are subjected to operators in order to create offspring. Commonly used operators include selection, reproduction, crossover, and mutation (see Goldberg (1989) for further details). Several generations may have to be evaluated before significant improvement in the objective function is seen. GA were developed by John Holland and his team at the University of Michigan. GA are noted for robustness in searching complex spaces and are best suited for combinatorial problems.

2.4.2 Evolutionary Strategies (ES)

Similar to GA, evolutionary strategies (ES) are algorithms that imitate the principles of natural evolution as a method to solve parameter optimization problems. Schwefel made the first attempt towards extending this strategy in order to solve discrete parameter optimization problems (Schwefel, 1995).

The first algorithm employed was a simple mutation-selection scheme called two membered ES, or (1+1)-ES. This scheme consisted of one parent producing one offspring by adding standard normal random variates. The better of the parent and the offspring becomes the parent for the next generation. The termination criteria include number of generations, elapsed CPU time, absolute or relative progress per generation, etc. The multimembered ES, or (µ+1)-ES, involves two parents, randomly selected from the current population of µ > 1 parents, producing one offspring (Back et al., 1991). Extensions of the (µ+1)-ES scheme include (µ+λ)-ES and (µ,λ)-ES. In a (µ+λ)-ES, µ parents produce λ offspring followed by removal of λ least fit individuals (parents and offspring) to restore the population size to µ. A (µ,λ)-ES is comparable to the (µ+λ)-ES, however, only the offspring undergo selection. Back et al. (1991) also elaborated on other complex versions of ES, like correlated mutations, along with experimental comparisons of ES with popular direct search methods. Maria (1995) used a hybrid GA-ES algorithm to solve multimodal continuous optimization problems.

2.4.3 Simulated Annealing (SA)

Simulated annealing is a stochastic search method analogous to the physical annealing process where an alloy is cooled gradually so that a minimal energy state is achieved. SA avoids getting stuck in local optima (hill climbing) and keeps track of the best objective value overall. SA performs well on combinatorial problems. Several versions of this heuristic exist (see Fleischer, (1995) and Alrefaei et al. (1995)). SA was first introduced by Metropolis et al. (1953). Kirkpatrick et al. (1983) applied this approach to deterministic optimization problems. Stuckman et al. (1991) compared GA, SA, and Bayesian techniques in the context of design optimization on multiple metrics including simulation time and dimensionality.
2.4.4 Tabu Search (TS)

Tabu search was developed by Fred Glover (1989, 1990) for optimizing an objective function with a special feature designed to avoid being trapped in local minima. TS is used for solving combinatorial optimization problems ranging from graph theory to pure and mixed integer programming problems. It is an adaptive procedure with the ability to utilize many other methods, such as LP algorithms and specialized heuristics, which it directs to overcome their limitations of getting stuck in local optima.

In TS, a fixed-length list of explored moves is maintained. This list represents the Tabu moves, moves that are not allowed at the present iteration, in order to exclude back-tracking moves. Subsequent to each move \( s \rightarrow s^* \) (where \( s \) is a feasible point and \( s^* \) is the best neighbor), the opposite move \( s^* \rightarrow s \) is appended to the Tabu list and the oldest move in the list is removed. Various versions of this search technique exist. For any TS implementation, the following are needed: a forbidding strategy, a freeing strategy, a short-term strategy, and a stopping criterion (Osman, 1993). The latest research and computational comparisons have validated the ability of TS to obtain high quality solutions with modest computational effort, generally dominating alternative methods tested.

2.4.5 Nelder And Mead’s Simplex Search

The search starts with points in a simplex consisting of \( p+1 \) vertices (not all in the same plane) in the feasible region. It proceeds by continuously dropping the worst point in the simplex and adding a new point determined by the reflection of the worst point through the centroid of the remaining vertices. Disadvantages of this method include the assumption of convex feasible region and implementation problems involving the handling of feasibility constraints. Box’s complex search is an extension of Nelder and Mead’s simplex search modified for constrained problems (Reklaitis et al., 1983). See also Azadivar and Lee (1988), Barton and Ivey (1991), and Tomick et al., (1995) for enhancements to the Nelder-Mead method. Hall and Bowden (1997) concluded that Nelder-Mead method performs better than ES or TS with smooth convex response surfaces.

2.5 A-Teams

An A-team (asynchronous team) is a process that involves combining various problem solving strategies so that they can interact synergistically. De Souza and Talukdar (1991) viewed an A-team as a process that is both fast and robust. They have demonstrated that A-teams consisting of GA and conventional algorithms, such as Newton’s Method and Levenberg-Marquardt algorithms, for solving sets of nonlinear algebraic equations, result in considerable savings in the amount of computational effort (number of function evaluations) necessary for finding solutions. A-teams are inherently suitable for multi-criteria simulation optimization problems, and therefore, represent one of the fastest growing areas of simulation optimization research. For optimizing a kanban sizing problem, Hall and Bowden (1996) utilized a two-phase approach - ES followed by Hooke-Jeeves search method - to obtain “good” solutions with 60% fewer simulation runs than with ES alone.

2.6 Statistical Methods

2.6.1 Importance Sampling Methods

Importance sampling has been used effectively to achieve significant speed ups in simulations involving rare events, such as failure in a reliable computer system or ATM communication network (Shahabuddin, 1995). The basic idea of importance sampling is to simulate the system under a different probability measure (e.g. with different underlying probability distributions) so as to increase the probability of typical sample paths involving the rare event of interest. For each sample path (observation) during the simulation, the measure being estimated is multiplied by a correction factor to obtain an unbiased estimate of the measure in the original system. The main problem in importance sampling is to come up with an appropriate change of measure for the rare event simulation problem at hand.

2.6.2 Ranking and Selection

Ranking and selection methods are frequently employed for practical problems, for instance, finding the best combination of parts manufactured on various machines to maximize productivity, or finding the best location for a new facility to minimize cost. In these optimization problems, some knowledge of the relationship among the alternatives is available. These methods have the ability to treat the optimization problem as a multi-criteria decision problem. When the decision involves selecting the best system design, the technique of indifference-zone ranking may be employed. When the decision involves selecting a subset of system designs that contains the best design, the technique of subset selection may be employed. In either case, the decisions are guaranteed to be correct with a pre-specified probability. Many ranking and selection procedures can be found in Gupta and Panchapakesan (1979).

2.6.3 Multiple Comparisons With The Best
If the problem is to select the best of a finite number of system designs, multiple comparison with the best (MCB) is an alternative to ranking and selection. In MCB procedures inference about the relative performance of all alternatives tested is provided. Such inference is critical if the performance measure of interest is not the sole criterion for decision making, e.g., expected throughput of a manufacturing system may be the performance measure of interest but cost of maintaining the system is also important. According to Hsu and Nelson (1988), MCB combines the two most frequently used ranking and selection techniques, namely, indifference zone and subset selection inference. Goldsman and Nelson (1990) devised an MCB procedure for steady state simulation experiments based on batching. This MCB procedure can be implemented in a single run of each alternative under consideration, which is important if restarting simulation experiments is unwieldy and/or expensive.

2.7 New Developments

Many of the techniques discussed above require speedy determination of the output variable value at a given input parameter vector value. Much effort is being directed to expedite this process. Parallel programming implementations of simulation optimization models have shown promise (see Yucesan et al. (1995), Schruben (1992), and Heidelberger (1988) for details). A variant of sample path optimization is applied to solve deterministic variational inequalities which model many equilibrium phenomena in economics, physics and operations research in Gurkan et al. (1996). Another method very similar in concept to sample path optimization is the technique of retrospective optimization proposed by Healy and Schruben (1991). Ho et al. (1992) combined RSM and stochastic approximation to yield a method referred to as the gradient surface method.

3 SIMULATION OPTIMIZATION APPLICATIONS

Simulation optimization methods have been applied to applications with a single objective, applications that require the optimization of multiple criteria, and applications with non-parametric objectives. Azadivar et al. (1996) applied a simulation optimization algorithm based on Box’s complex search method to optimize the locations and inventory levels of semi-finished products in a pull-type production system. Hall et al. (1996) used ES with a simulation model for optimizing a kanban sizing problem. Lutz (1995) developed a procedure that combined simulation with TS to deal with problems of work-in-process (WIP) inventory management. Fu and Healy (1992) applied the PA technique to inventory models where the demand has an associated renewal arrival process.

Tompkins and Azadivar (1995) proposed an approach to link a GA and an object-oriented simulation model generator to find the optimal shop floor layout. Faccenda and Tenga (1992) presented an approach to incorporate the process plant production operations into the design of a facility by combining simulation and GA. Morito et al. (1993) used an algorithm that combined SA and simulation to find an appropriate dispatching priority of operations to minimize the total tardiness for a commercial flexible manufacturing system (FMS).

Brennan and Rogers (1995) employed the methods of IPA and stochastic approximation to solve the problem of optimizing the performance of an asynchronous line used for component assembly in electronics manufacturing. Manz et al. (1989) used a simulated annealing (SA) algorithm in conjunction with a SIMAN simulation model to find the optimal parameter levels for operating an automated manufacturing system. Several simulation optimization applications, in fields as diverse as microelectronics assembly and hospital administration, are discussed in Akbay (1996).

Hill and Fu (1994) applied simultaneous perturbation stochastic approximations (SPSA), which requires only two simulations per gradient estimate, regardless of the number of parameters of interest, to a transportation problem. Evans et al. (1991) discussed the applicability of traditional multi-criteria mathematical optimization techniques - multiattribute value function, multiattribute utility function, goal programming, progressive articulation of preferences, and posterior articulation of preferences - to multi-criteria simulation optimization problems.

4 SIMULATION OPTIMIZATION SOFTWARE

A long standing goal among some of the simulation practitioners and theoreticians was being able to guide a series of simulations in the most effective way instead of performing “blind” experiments and assuming that at least one of the experiments will yield the best alternative to implement (Glover et al., 1996). Many simulation software developers today have become more aware of the importance of finding optimal and near-optimal solutions for applications in minutes, instead of performing an exhaustive examination of relevant alternatives in days or months. Simulation software that includes special search procedures to guide a series of simulations to reveal optimal or near-optimal scenarios includes: ProModel, AutoMod, Micro Saint, LayOPT, and FactoryOPT. A brief description of each software’s optimization and/or statistical module follows.
The add-on optimization module for ProModel is called SimRunner Optimization. This module consists of two features for analyzing and optimizing existing ProModel simulation models. The first feature is a factorial design of experiments that reveals the effect of a change in input factor on the objective function. The second feature is a multi-variable optimization that tries various combinations of input factors to arrive at the combination that yields the best objective function value.

AutoMod and AutoSched has an add-on package called AutoStat, a statistical analysis package that includes a “Select the Best” ranking and selection procedure. This procedure finds the single best system, or a subset containing the best system, from among a finite number of systems (Carson, 1996).

Glover et al. (1996) described a software package called OptQuest that integrates simulation, an intelligent search procedure called scatter search (based on TS), a mixed integer programming solver, and a procedure to configure and train neural networks. It can handle multiple objectives (provided that they are mapped into a single final objective) and linear constraints on the input variables. It indicates the search progress graphically. OptQuest is customized to work with the modeling and simulation software Micro Saint 2.0 (see Drury et al. (1996) for Micro Saint use).

LayOPT is a facilities’ layout analysis and optimization software package that starts with an existing block layout and attempts to improve this layout by exchanging the locations of defined departments. These changes are based on given flow and cost data (Grajo, 1996). FactoryOPT works with FactoryPLAN and/or FactoryFLOW and works in AutoCAD to create the most efficient factory designs in record time (Sly, 1996).

SIMICOM algorithm (Azadivar and Lee, 1988) based on Box’s complex search method can be interfaced with any modeling and simulation software.

5 CLOSING REMARKS

From the extensive literature review presented here, it can be concluded that interest in the area of simulation optimization is growing. More direct search methods need to be explored for suitability to simulation optimization problems.

REFERENCES


AUTHOR BIOGRAPHIES

**YOLANDA CARSON** is a PhD candidate in the department of Systems Science & Industrial Engineering at the State University of New York at Binghamton. Her research interests include simulation modeling and analysis and optimizing the performance of manufacturing systems.

**ANU MARIA** is an assistant professor in the department of Systems Science & Industrial Engineering at the State University of New York at Binghamton. She received her PhD in Industrial Engineering from the University of Oklahoma. Her research interests include optimizing the performance of materials used in electronic packaging (including solder paste, conductive adhesives, and underfills), simulation optimization techniques, genetics based algorithms for optimization of problems with a large number of continuous variables, multi criteria optimization, simulation, and interior-point methods.