Proceedings of the 2005 Winter Simulation Conference
M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, eds.

BALANCING BIAS AND VARIANCE IN THE OPTIMIZATION OF SIMULATION MODELS

Christine S.M. Currie

School of Mathematics
University of Southampton
Southampton, SO17 1BJ, U.K.

Russell C.H. Cheng

School of Mathematics
University of Southampton
Southampton, SO17 1BJ, U.K.

ABSTRACT

We consider the problem of identifying the optimal point of an objective in simulation experiments where the objective is measured with error. The best stochastic approximation algorithms exhibit a convergence rate of $n^{-1/6}$ which is somewhat different from the $n^{-1/2}$ rate more usually encountered in statistical estimation. We describe some simple simulation experimental designs that emphasize the statistical aspects of the process. When the objective can be represented by a Taylor series near the optimum, we show that the best rate of convergence of the mean square error is when the variance and bias components balance each other. More specifically, when the objective can be approximated by a quadratic with a cubic bias, then the fastest decline in the mean square error achievable is $n^{-2/3}$. Some elmentary theory as well as numerical examples will be presented.

1 INTRODUCTION

We consider the problem of identifying the optimal point of a non-linear objective function in simulation experiments where the objective is measured with error. This problem may arise in a number of settings, in particular when determining the best set up for a stochastic system, such as that described in a previous paper (Cheng and Currie 2004). In order to demonstrate the principles and investigate some of the theoretical issues, we consider only a simple one-dimensional example,

$$y(x) = \eta(x) + \varepsilon,$$

where $\eta(x)$ is the underlying objective function, and $\varepsilon \sim N(0, \sigma^2)$ is a random error term, which we assume to be normally distributed. We assume that it is possible to expand the objective function as a Taylor series near the optimum and so consider the convergence of the mean square error when the objective can be estimated as a quadratic with a higher-powered bias. We further assume that in the range under consideration, the objective function is strongly convex (has only one minimum), i.e. $d\eta/dx < 0$ for $x < x^*$ and $d\eta/dx > 0$ for $x > x^*$.

Two experimental designs are considered, both involving making observations either side of the optimum, at a distance that decreases with the order of the observation. The difference between the two designs is in the estimate of the optimal point. In the first design, we consider the distribution of points around a known optimum. Although unrealistic, this serves to demonstrate some of the statistical properties of these designs. In the second design, we use our current maximum likelihood estimate of the optimum to set the design point for the next iteration. The bias and variance both depend on the number of observations made, with the dependence being determined by the rate at which the design points converge on the optimum. We show that, under the optimal settings for the first design, the contribution of the bias and the variance to the mean square error are balanced.

Although concerned with experimental design, the methodology we propose for choosing design points has close links with the technique of stochastic approximation. Robbins and Monro (1951) were the first to give a formal mathematical treatment of stochastic approximation, applying it to finding the solution to the equation $y(\theta) = M$, where the output of the process, $y(\theta)$, is a noisy function of its inputs. Kiefer and Wolfowitz (1952) adapted their work, and used the techniques of stochastic approximation to find the maximum or minimum of a noisy function. Stochastic approximation is a sequential method in which the point chosen for the next experiment is dependent on the point of the previous experiment and the most recent observations. For example, the approach used by Kiefer and Wolfowitz is based on making a finite-difference approximation at each iteration, such that the estimate of the minimum after the n^{th} iteration is

$$x_n^* = x_{n-1}^* - \alpha_n \frac{f(x_{n-1}^* + \beta_n) - f(x_{n-1}^* - \beta_n)}{2\beta_n},$$

and x_n^* is also the n^{th} design point.

The convergence rate of the methodology will depend on the properties of the objective function $f(x)$. We assume that $f(x)$ is continuous and show that the convergence rate of the algorithm depends on the order of its differentiability. In most situations of interest, the cubic term will dominate the Taylor series near the optimum, and the function can be regarded as being thrice differentiable. In this case we find that the design points should converge on the optimum at a rate of $n^{-1/6}$, with the mean square error decaying as $n^{-2/3}$. This convergence rate matches the optimum convergence rate for stochastic approximation algorithms for a function which is thrice differentiable, given by Dupač (1957). Convergence rates for stochastic approximation are also discussed in Wasan's book (Wasan 1969).

We apply the two designs to a numerical example of a noisy quadratic function with a cubic perturbation in Section 3, demonstrating the performance of the algorithms at a number of different settings for the convergence of the design points. The numerical results from the second, more practical, design suggest that this has a similar convergence rate to the first design. The theoretical convergence properties of the first design are also considered in Section 2.

2 METHODOLOGY

We consider perturbations to the function $y = x^2$ of the form ax^q, where q is an integer, i.e. the function

$$\eta(x) = x^2 + ax^q.$$

This has a local minimum at $x = 0$. We assume that observations are subject to an additive noise term, $\varepsilon \sim N(0, \sigma^2)$, such that an observation

$$y(x) = x^2 + ax^q + \varepsilon. \tag{1}$$

In the methodology that we propose, the i^{th} design point will be at $h_i = \frac{(-1)^i K}{i^p}$. We show initially that this is equivalent to having $n/2$ observations at each of h and $-h$.

The mean value of the positive observations is

$$\frac{2}{n}\sum_{i=1}^{n/2}\frac{1}{(2i)^p}, \tag{2}$$

and the mean value of the negative observations is

$$-\frac{2}{n}\sum_{i=1}^{n/2}\frac{1}{(2i-1)^p}. \tag{3}$$

To evaluate these sums, we make use of the Maclaurin-Cauchy formula,

$$\lim_{n\to\infty}\left\{\sum_{i=1}^{n}\frac{1}{i^p} - \int_1^n \frac{1}{x^p}dx\right\} \to L,$$

where $0 \le L \le 1$. The sum in (2) can therefore be rewritten as

$$\sum_{i=1}^{n/2}\frac{1}{(2i)^p} = \frac{1}{2^p}\sum_{i=1}^{n/2}\frac{1}{i^p},$$

and so, in the limit that $n \to \infty$, the mean value of the positive observations,

$$\begin{aligned}\frac{2}{n}\sum_{i=1}^{n/2}\frac{1}{(2i)^p} &\to \frac{2}{n2^p}L + \frac{2}{n2^p(p-1)}\left(1-(n/2)^{1-p}\right)\\ &= \frac{1}{n2^{p-1}}\left(L - \frac{1}{(1-p)}\right) + \frac{n^{-p}}{(1-p)}.\end{aligned}$$

As n is large, this behaves as if all of the positive observations were taken at

$$h = \frac{n^{-p}}{1-p} \equiv Kn^{-p}$$

for $p < 1$.

Similarly, rewriting the sum in (3),

$$\frac{2}{n}\sum_{i=1}^{n/2}\frac{1}{(2i-1)^p} = \sum_{i=1}^{n}\frac{1}{i^p} - \sum_{j=1}^{n/2}\frac{1}{(2j)^p},$$

and in the limit of $n \to \infty$, we can write the mean value of the negative observations as

$$\begin{aligned}\frac{2}{n}\sum_{i=1}^{n/2}\frac{-1}{(2i-1)^p} &\to -\frac{2}{n}\left[L + \frac{1}{1-p}\left(n^{1-p}-1\right)\right]\\ &\quad -\left[\frac{1}{n2^{p-1}}\left(L - \frac{1}{1-p}\right)\right.\\ &\quad \left.+\frac{n^{-p}}{(1-p)}\right]\\ &= \frac{-n^{-p}}{1-p} + \frac{(2^p-1)}{n2^{p-1}}\left(L - \frac{1}{1-p}\right).\end{aligned}$$

As with the positive observations, this shows that, in the limit that $n \to \infty$, this methodology is equivalent to all of the negative observations being taken at

$$-h = -\frac{1}{1-p}n^{-p} \equiv -Kn^{-p}.$$

From the above analysis, we can therefore assume that we make $n/2$ observations at each of $x = -h$ and h, such that the total number of observations made is n. The observed sample means at $-h$ and h will be

$$y_1 = h^2 - ah^q + \varepsilon_1$$

$$y_2 = h^2 + ah^q + \varepsilon_2,$$

where the ε_i will depend on n.

We consider fitting the function

$$g(x) = \theta_1 f_1(x) + \theta_2 f_2(x) \tag{4}$$

to $\eta(x)$, where the $f_i(x)$ are orthonormal basis functions. These must be orthogonal, i.e. $\sum_{j=1}^{n} f_k(x_j) f_l(x_j) = 0$ and be normalized, i.e. $\sum_{j=1}^{n} f_k(x_j) f_k(x_j) = 1$ for $k = 1, 2$ and $l = 1, 2$. We assume that the basis functions are of the form

$$\begin{aligned} f_1(x) &= bx + c, \\ f_2(x) &= dx^2, \end{aligned}$$

where b, c, d are constants to be determined. Different results are obtained for q odd and q even. The case where q is assumed to be odd is more interesting, and we consider that first.

Substituting the f_i into the orthonormality conditions, we find that $b = \frac{1}{h\sqrt{n}}$, $c = 0$ and $d = \frac{1}{h^2\sqrt{n}}$, so that

$$\begin{aligned} f_1(x) &= \frac{x}{h\sqrt{n}} \\ f_2(x) &= \frac{x^2}{h^2\sqrt{n}}. \end{aligned} \tag{5}$$

We can therefore rewrite $g(x)$ as

$$g(x) = \hat{\theta}_1 \frac{x}{h\sqrt{n}} + \hat{\theta}_2 \frac{x^2}{h^2\sqrt{n}}.$$

Using the orthonormal properties of the basis functions, and referring back to (4), the least squares estimates of the θ_i are

$$\hat{\theta}_i = \sum_{j=1}^{n} f_i(x_j) y_j,$$

and for the set of n observations made,

$$\begin{aligned} \hat{\theta}_1 &= \frac{\sqrt{n}}{2}\left(2ah^q + \varepsilon_2 - \varepsilon_1\right) \\ \hat{\theta}_2 &= \frac{\sqrt{n}}{2}\left(2h^2 + \varepsilon_1 + \varepsilon_2\right). \end{aligned}$$

We wish to determine x^*, the value of x at the minimum. Differentiating,

$$\frac{dg}{dx} = \frac{\hat{\theta}_1}{h\sqrt{n}} + \frac{2\hat{\theta}_2 x}{h^2\sqrt{n}},$$

and is zero at x^*, where

$$\begin{aligned} x^* &= \frac{-h\hat{\theta}_1}{2\hat{\theta}_2} \\ &= \frac{-ah^{q-1}}{2}\left(1 + \frac{\varepsilon_2 - \varepsilon_1}{2ah^q}\right)\left(1 + \frac{\varepsilon_1 + \varepsilon_2}{2h^2}\right)^{-1}. \end{aligned}$$

If $\frac{\varepsilon_1 + \varepsilon_3 - 2\varepsilon_2}{2h^2}$ is neglible as $n \to \infty$, the third term in the equation will tend to one. Assuming that $\varepsilon_i = O_p(n^{-1/2})$ and remembering that $h = O(n^{-p})$, this means that

$$-\frac{1}{2} + 2p < 0,$$

and so $p < \frac{1}{4}$. Assuming this is true, then in the limit of large n,

$$x^* = \frac{-ah^{q-1}}{2}\left(1 + \frac{\varepsilon_2 - \varepsilon_1}{2ah^q}\right).$$

We define the optimal experimental design as one that minimizes the mean square error (MSE), which is defined to be

$$MSE = bias^2 + variance.$$

The bias in x^* is $B = \frac{-ah^{q-1}}{2}$ and the variance is $V = \frac{\sigma^2}{nh^2}$. If $h = Kn^{-p}$ then $V = \frac{\sigma^2 n^{2p-1}}{K^2}$ and $B = -\frac{1}{2}aK^{q-1}n^{-p(q-1)}$. The MSE can then be written as

$$\begin{aligned} MSE &= \frac{a^2K^{2(q-1)}n^{-2p(q-1)}}{4} + \frac{3\sigma^2 n^{2p-1}}{K^2} \\ &= \alpha n^{-2p(q-1)} + \beta n^{2p-1}, \end{aligned}$$

where

$$\begin{aligned} \alpha &= \frac{a^2K^{2(q-1)}}{4} \\ \beta &= \frac{\sigma^2}{K^2}, \end{aligned}$$

which are both independent of n and p.

We wish to find the p that minimizes the MSE and so differentiate with respect to p,

$$\begin{aligned}\frac{dMSE}{dp} &= \alpha\left(-2(q-1)n^{-2p(q-1)}\ln(n)\right) \\ &\quad +\beta\left(2n^{2p-1}\ln(n)\right) \\ &= 2\ln(n)\left(\beta n^{2p-1}-\alpha(q-1)n^{-2p(q-1)}\right).\end{aligned}$$

At the minimum, this is equal to zero and p^*, the optimal value of p, obeys

$$\begin{aligned}\beta n^{2p^*-1} &= \alpha(q-1)n^{-2p^*(q-1)} \\ \frac{\beta}{\alpha(q-1)} &= n^{1-2p^*q}. \qquad (6)\end{aligned}$$

Taking logs of both sides,

$$p^* = \frac{1}{2q}\left(1-\frac{\ln\left(\frac{\beta}{\alpha(q-1)}\right)}{\ln(n)}\right). \qquad (7)$$

Therefore, as $n\to\infty$, $p^*\to 1/2q$.

With p equal to $1/2q$, the variance declines as $V\sim n^{-\frac{q-1}{q}}$ and the square of the bias decays at the same rate, $B^2\sim n^{-\frac{q-1}{q}}$. The bias has a negative dependence on p and the variance a positive dependence. Therefore, with $p=1/2q$, we have a balance between the two.

With q even, the bias term in the mean square error is zero and the design is chosen simply to reduce the variance. The variance decreases with decreasing p, and so the best design with an even powered deviation from the quadratic (q even) is to choose points further away from the optimum, as n increases.

If we are sufficiently close to the minimum, the dominant term in the Taylor series would be the cubic term, i.e. $q=3$, suggesting an optimal value for p of $1/6$, and a mean squared error that decays as $n^{-\frac{2}{3}}$. This reproduces the results obtained for the optimal convergence of stochastic approximation algorithms, as put forward by Dupač (1957).

3 NUMERICAL EXAMPLES

In this section, we describe the implementation of two designs for finding the position of the minimum of (1), where $q=3$, $a=0.2$ and $\sigma=0.1$. The simple algorithm discussed in Section 2 is considered initially, where we observe the objective function at points around its known minimum value of zero. Table 3 gives estimates of the optimum after 1000 iterations, averaged over 10 runs of the algorithm. This shows that convergence to the optimum is fastest with $p=1/6$. However, the results are very variable, and could not be used as proof that the theory holds in practice.

Table 1: Estimates of the Minimum Value with Different Values of the Power p for the First Method. Estimates are the Average of 10 Runs, each of 1000 Iterations

p	Estimate of x_{min}
1/8	0.0223
1/6	0.0112
1/4	0.0245

If we now instead assume that the minimum is unknown to us before the start of the experiment, which is a more realistic case, we can choose design points x_i such that

$$x_i = x_i^* + \frac{(-1)^i}{i^p},$$

where x_i^* is our best estimate of the minimum after i iterations of the algorithm. The estimate of the minimum is obtained by fitting a quadratic model to the data using maximum likelihood methods. The position of the minimum can then be easily deduced from the parameter values of the quadratic function.

Finding new estimates of the maximum likelihood parameters at each step does not involve a complete refit as we can take advantage of the updating routines described in Kendall and Stewart (1991) and Bartlett (1951). We write the design matrix after n iterations as

$$\mathbf{X_n} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \\ 1 & x_n & x_n^2 \end{pmatrix},$$

and write the parameters of the quadratic model that we are fitting as $\hat{\theta}_n = (\hat{\theta}_{1n}\ \hat{\theta}_{2n})'$, such that the our estimate of the objective funcion after i iterations is

$$\hat{y}_i = \hat{\theta}_{1i} + \hat{\theta}_{2i}x_i + \hat{\theta}_{3i}x_i^2.$$

Using this notation, the maximum likelihood estimate of θ after n iterations is

$$\hat{\theta}_n = (\mathbf{X_n}'\mathbf{X_n})^{-1}\mathbf{X_n}'\mathbf{y_n} = \mathbf{A_n}^{-1}\mathbf{X_n}'\mathbf{y_n}.$$

We now make an additional observation y_{n+1} at x. Writing $\mathbf{x} = (1\ x\ x^2)'$, we can see that the new design matrix can be written as $\begin{pmatrix} \mathbf{X_n} \\ \mathbf{x}' \end{pmatrix}$. We can therefore make use of the

Table 2: Estimates of the Minimum Value with Different Values of the Power p for the Second Method. Estimates are the Average of 10 Runs, each of 1000 Iterations

p	Estimate of x_{min}
1/8	0.0115
1/6	0.0808
1/4	0.217

matrix theorem stated in Bartlett (1951) to update $\mathbf{A}_n^{-1}$,

$$\mathbf{A}_{n+1}^{-1} = \mathbf{A}_n^{-1} - \frac{\mathbf{A}_n^{-1}\mathbf{x}\mathbf{x}'\mathbf{A}_n^{-1}}{(1+\mathbf{x}'\mathbf{A}_n^{-1}\mathbf{x})},$$

and use this to update the maximum likelihood estimate of θ given the additional observation, such that

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \mathbf{A}_n^{-1}\mathbf{x}\frac{(y_{n+1} - \mathbf{x}'\hat{\theta}_n)}{(1+\mathbf{x}'\mathbf{A}_n^{-1}\mathbf{x})}.$$

Updating $\mathbf{A}_n^{-1}$ and $\hat{\theta}_n$ involves no new matrix inversion, therefore using these formulae it is only necessary to perform one matrix inversion at the start of the procedure, speeding up the process considerably.

Table 3 gives estimates of the optimum after 1000 iterations, averaged over 10 runs of the algorithm. These suggest that the best estimate of the optimum after 1000 runs is obtained for $p = 1/8$.

These experiments suggest that the theoretical results which we have proved for the initial design, where the minimum is assumed to be known, hold in practice, but that $p = 1/6$ will not necessarily be the optimal convergence rate for the second, more practical design. Further work is needed to investigate what the optimal convergence rate is for the second design.

4 CONCLUSION

We have demonstrated the use of a simple experimental design for finding the optimum of a stochastic objective function. The theoretical treatment of the problem showed that, for the optimal experimental design, the contributions of the variance and the bias to the mean square error are balanced. Under the optimal design we have shown that the optimal convergence rate for the design points is $p = 1/6$ for a quadratic objective function perturbed by a cubic term, and that at this rate the mean square error declines as $n^{-2/3}$, matching the convergence results recorded by Dupač (1952) for stochastic approximation algorithms. The numerical results agree with this result, but suggest that the optimal convergence rate for the second design may not be $p = 1/6$. Investigating the different convergence properties of the two designs will be the focus of further work on this problem.

This experimental design has many similarities with stochastic approximation. The treatment of the problem in this paper will hopefully highlight some of the interesting statistical properties of the problem of maximising a stochastic objective function.

REFERENCES

Bartlett, M.S. 1951. An inverse matrix adjustment arising in discriminant analysis. *The Annals of Mathematical Statistics* 22 (1): 107–111.

Cheng, R.C.H. and C.S.M. Currie. 2004. Optimization by simulation metamodelling methods. In *Proceedings of the 2004 Winter Simulation Conference*, ed. R.G. Ingalls, M.D. Rossetti, J.S. Smith, and B.A. Peters, 485–490. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Dupač, V. 1952. On the Kiefer-Wolfowitz approximation method. *Časopis Pro Pěstovani Matematiky* 82 : 47–75.

Kendall, M.G., A. Stuart, and J.K. Ord. 1991. *Kendall's advanced theory of statistics.* Fifth ed. New York: Oxford University Press.

Kiefer, J., and J. Wolfowitz. 1952. Stochastic estimation of the maximum of a regression function *The Annals of Mathematical Statistics* 23 (3): 462–466.

Robbins, H. and S. Monro. 1951. A stochastic approximation method. *The Annals of Mathematical Statistics* 22 (3): 400–407.

Wasan, M.T. 1969. *Stochastic Approximation*. Cambridge: Cambridge University Press.

AUTHOR BIOGRAPHIES

CHRISTINE S.M. CURRIE is a lecturer of operational research in the School of Mathematics in the University of Southampton, where she also obtained her PhD. In addition, she has an MPhys from Oxford University and an MSc in Operational Research from the University of Southampton. Her research interests include mathematical modeling of epidemics, Bayesian statistics, variance reduction methods and optimization of simulation models. Her email address is <christine.currie@soton.ac.uk> and her web page is <www.maths.soton.ac.uk/staff/currie>.

RUSSELL C.H. CHENG is Professor, Head of Operational Research, and Deputy Dean of the Faculty of Mathematical Studies at the University of Southampton. He has an M.A. and the Diploma in Mathematical Statistics from

Cambridge University, England. He obtained his Ph.D. from Bath University. He is a former Chairman of the U.K. Simulation Society, a Fellow of the Royal Statistical Society, Member of the Operational Research Society. His research interests include: variance reduction methods and parametric estimation methods. He was a Joint Editor of the IMA Journal of Management Mathematics. His e-mail address is <r.c.h.cheng@maths.soton.ac.uk>, and his web page is <www.maths.soton.ac.uk/staff/cheng>.