A Data-Analytic Approach to Bayesian Global Optimization

Matthias Schonlau [1], William J. Welch [1], and Donald R. Jones [2]

[1] Department of Statistics and Actuarial Science and The Institute for Improvement in Quality and Productivity, University of Waterloo, Waterloo, Ontario N2L 3G1 Canada

[2] Dept CO-30, General Motors R&D Center, Warren, MI 48090-9055, USA

KEYWORDS: Average Case Analysis, Model Diagnostics, Nonparametric Function Fitting, Response Surface, Stochastic Process

1 Introduction

Global optimization, that is the search for a global extremum, is a problem frequently encountered. Sometimes it is extremely costly to evaluate a function for an engineering design. For example, Davis (1996) wrote about experiences at Boeing:

> "Designing helicopter blades to achieve low vibration is an extreme example of a problem where it is prohibitively expensive to compute responses for large numbers of design alternatives."

For such applications one is interested in minimizing the total number of function evaluations needed to find the global extremum.

Global optimization methods that are based on statistical models of the objective function have been very successful here for two reasons: (1) they base the decision where to sample further points on all previous function evaluations, rather than just on the last few, and (2) they select points based on average case scenarios rather than on worst case behaviour.

Worst case analysis corresponds to the minimax paradigm which postulates that the maximal loss is to be minimized, for example rules based on Lipschitz bounds. The sampling decision under uncertainty can be viewed as a two player game against nature. Whereas a worst case analysis may be realistic when playing against an intelligent opponent, average case analysis is more appropriate here since nature is impartial with respect to the outcome. This line of argument dates back to Wald (1949).

The method proposed in this paper deals with the unconstrained global optimization problem, minimize $f(\mathbf{x})$ where $\mathbf{x} = (x_1, \ldots, x_k)$. This includes the class of problems with simple constraints like $a_i \le x_i \le b_i$, since these problems can be transformed to unconstrained global optimization problems. Throughout we assume without loss of generality that the extremum of interest is a minimum.

The outline of this paper is as follows. In Section 2 we briefly review the Bayesian global optimization approach and introduce a more flexible stochastic model in that framework. Section 3 explains how to assess and improve the model fit and hence the effectiveness of the global optimization method. The assessment is based on graphical diagnostics. Section 4 shows by means of several examples from the optimization literature that this approach is very efficient in terms of the number of function evaluations required. Section 5 concludes with some discussion.

2 Expected Improvement Algorithm

The algorithm is based on the idea that any future sampled point constitutes a potential improvement over the minimal sampled value up to the present stage. Therefore we will refer to the algorithm throughout as the *expected improvement algorithm.* As we will show later, the expected improvement criterion is equivalent to one-step-ahead optimality in Bayesian Global Optimization.

The expected improvement algorithm proceeds in five steps:

1. Choose a small initial set of design points spread over the entire x space. Evaluate the true function at these points.
2. Model the true function using all previous function evaluations.
3. Find the maximum of the "expected improvement"-criterion. The location of the maximum is a new design point.
4. Evaluate the true function at the new design point.

5. Compute a stopping criterion. If the stopping criterion is not met go to Step 2.

Note that after each sampling Step the predictor is updated (Step 2), and the expected improvement is recalculated (Step 3).

For Step 1 Latin hypercube sampling schemes (McKay et al., 1979) are particularly useful, because they have space filling property, i.e. they cover the x domain to explore the function globally. The number of points sampled at this initial stage is somewhat arbitrary. We choose about 10 points per active variable because, in our experience, one needs at least that many points to obtain a reasonably good fit.

For the modelling approach in Step 2 we use a stochastic process with a more flexible correlation structure than previously employed. This is further discussed in Section 2.1.

The "expected improvement"-criterion in Step 3 is based on the idea that any additional function evaluation constitutes a potential reduction of the minimal function evaluation found so far. This is further discussed in Section 2.2.

Step 4 consists of evaluating the next design point. For Step 5, we propose to stop when the maximum of the expected improvement is smaller than a tolerance value; smaller in absolute value or relative to the current minimal function value.

2.1 Modelling Approach

Suppose after an initial experimental design or at some stage of the algorithm, we have n vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ at which the function f has been evaluated. Each vector $\mathbf{x}$ is k-dimensional for the k covariates (or inputs) $x_1, \ldots, x_n$. The corresponding response values (or outputs) are denoted $\mathbf{y} = (y_1, \ldots, y_n)^t$. Then, following the approach of, e.g., Welch et al. (1992), the response is treated as a random function or a realization of a stochastic process:

$$Y(\mathbf{x}) = \beta + Z(\mathbf{x}),$$

where $E(Z(\mathbf{x})) = 0$ and $\mathrm{Cov}(Z(\mathbf{w}), Z(\mathbf{x})) = \sigma^2 R(\mathbf{w}, \mathbf{x})$ for two inputs $\mathbf{w}$ and $\mathbf{x}$. The correlation function $R(\cdot, \cdot)$ can be tuned to the data. Here it is assumed to have the form:

$$R(\mathbf{w}, \mathbf{x}) = \prod_{j=1}^{k} \exp(-\theta_j |w_j - x_j|^{p_j}), \tag{1}$$

where $\theta_j \geq 0$ and $0 < p_j \leq 2$. The p_j's can be interpreted as parameters which indicate the smoothness of the response surface (smoother as the p's increase) and the θ's indicate how local the predictor is (more local as the θ's increase).

The best linear unbiased predictor of y at an untried $\mathbf{x}$ can be shown to be:

$$\hat{y}(\mathbf{x}) = \hat{\beta} + \mathbf{r}^t(\mathbf{x})\mathbf{R}^{-1}(\mathbf{y} - \mathbf{1}\hat{\beta}), \tag{2}$$

where $\mathbf{r}(\mathbf{x})$ is the vector of the correlations between $\mathbf{x}$ and each of the n design points, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from (1), $\hat{\beta} = (\mathbf{1}^t\mathbf{R}^{-1}\mathbf{1})^{-1}\mathbf{1}^t\mathbf{y}$ is the generalized least squares estimator of β, $\mathbf{R}$ is a correlation matrix with element (i, j) defined by $R(\mathbf{x}_i, \mathbf{x}_j)$ in (1) and $\mathbf{1}$ is a vector of 1's.

The MSE of the estimate can be derived as:

$$\mathrm{MSE}[\hat{y}(\mathbf{x})] = \sigma^2 \left[1 - \begin{pmatrix} 1 & \mathbf{r}_x^t \end{pmatrix} \begin{pmatrix} 0 & \mathbf{1}^t \\ \mathbf{1} & \mathbf{R} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ \mathbf{r}_x \end{pmatrix} \right] \tag{3}$$

The predictor in (2) has proven to be accurate for numerous applications, see e.g. Currin et al. (1991), Sacks et al. (1989a), Sacks et al. (1989b), Welch et al. (1992).

2.2 Criterion

We will now derive the expected improvement criterion.

If the function is sampled at $\mathbf{x}$ to determine $y = f(\mathbf{x})$ then the improvement I over f_{min}^n, the minimal sampled function value after n evaluations, is defined as

$$I = \begin{cases} f_{min}^n - y & \text{if} \quad y < f_{min}^n \\ 0 & \text{otherwise} \end{cases}.$$

The expected improvement is given as

$$E(I) = \int_{-\infty}^{f_{min}^n} (f_{min}^n - y)\phi(y)dy, \tag{4}$$

where $\phi()$ is the probability density function representing uncertainty about y.

Mockus (1989) proposes an equivalent criterion by specifying a loss function on the sequential n-step optimization strategy S_n :

$$L(S_n, f) = \min_x f(x) - f_{min}^n.$$

The risk, or the average loss is then given as

$$E\left(L(S_n, f)\right) = E(\min_x f(x)) - E(f_{min}^n) \tag{5}$$

An optimal strategy is defined as one that minimizes the risk (5). Computing an optimal strategy turns out to be computationally infeasible for even a moderate number of points n. The standard approach then is to relax the n-step optimality to one-step optimality. The criterion for one-step optimality is equivalent to (4).

To predict $Y(\mathbf{x})$ at an untried $\mathbf{x}$, we have $\hat{y}(\mathbf{x})$ with a mean squared error given by (3). For notational simplicity, we omit the dependence on $\mathbf{x}$, and denote $\hat{y}(\mathbf{x})$ by $\hat{\mathbf{y}}$ and the root mean squared error by s. Next, we take the distribution of the unknown $Y = Y(\mathbf{x})$ as $N(\hat{y}, s^2)$. If we further assume that the random function $Y(x)$ is Gaussian, then $\hat{y}$ is also normal. Thus, we represent uncertainty about the true y by saying it is $N(\hat{y}, s)$. The expected improvement in (4) can be expressed as

$$E(I) = \begin{cases} (f^n_{min} - \hat{\mu})\Phi(\frac{f^n_{min}-\hat{\mu}}{\hat{\sigma}}) + \hat{\sigma}\phi(\frac{f^n_{min}-\hat{\mu}}{\hat{\sigma}}) & \hat{\sigma} > 0 \\ 0 & \hat{\sigma} = 0 \end{cases} \tag{6}$$

where $\phi()$ and $\Phi()$ denote the probability density function and the cumulative density function of the standard normal distribution.

The expected improvement will tend to be large at a point whose predicted value is very small or where there is a lot of uncertainty associated with the prediction.

A practical problem, though, is finding the global maximum of the expected improvement criterion over a continuous region. We start a large number of local searches from random starting points. This does not guarantee to find the global maximum, of course. Mockus (1994) states in this context "[...] there is no need for exact minimization of the risk function", because we only determine the point of the next observation.

3 Diagnostics

The success of the Bayesian minimization algorithm depends on having a valid model. The better the model the more likely the algorithm will terminate quickly and with an accurate tolerance on the minimum. For this reason one would like to assess the performance of the modeling approach as soon as possible, that is after the initial function evaluations. When the model does not fit well it is often possible to improve the fit through appropriate transformations of the response. For this purpose we propose three diagnostic plots to be used after the initial function evaluations have been obtained. All of them are based on the concept of cross validation.

Cross validation is a statistical technique often used for assessing a model's predictive capability, when it is not convenient to test the model at unknown design points. It consists of setting aside and predicting a small portion of the data from a model based on the remaining larger portion of data. Most commonly only one point at a time is set aside, and cross validation is performed once for each point to be left out. In this paper we always use leave-one-out cross validation.

We remove case i from (2) and (3) to obtain $\hat{y}_{-i}(\mathbf{x}_i)$ and $s_{-i}(\mathbf{x}_i)$. The notation emphasizes that case i is removed when predicting at $\mathbf{x}_i$. Cross-validated standardized errors (residuals), for example, can be written as

$$e_i = \frac{y_i - \hat{y}_{-i}(\mathbf{x}_i)}{s_{-i}(\mathbf{x}_i)}. \tag{7}$$

We propose the following three diagnostic plots:

1. A plot of the cross validation predictions versus the true y's, i.e. $\hat{y}_{-i}(\mathbf{x}_i)$ versus y_i.
2. A plot of the cross validated standardized errors versus the cross validated predictions, i.e. e_i in (7) versus $\hat{y}_{-i}(\mathbf{x}_i)$.
3. A plot of the cross validated expected improvements versus the true y, i.e. $E(I)$ evaluated at $\mathbf{x}_i$ based on $\hat{y}_{-i}(\mathbf{x}_i)$ and $s_{-i}(\mathbf{x}_i)$ versus y_i.

The first plot indicates prediction accuracy. The second assesses where prediction error or uncertainty is realistic. In particular the standardized errors should not lie far outside about $[-2, 2]$ or $[-3, 3]$ if many points are plotted if the normal approximation in (6) is valid. The third plot assesses the expected improvement criterion. If the criterion works well, the lowest y's should be associated with the highest expected improvements. If the expected improvement criterion is not able to distinguish between high and low y's in the function evaluations to date, then the expected improvement algorithm will likely not work well in selecting new points.

If the plots indicate a poor fit a transformation of the data can often improve the fit. This is possible because the transformed data may more closely resemble a realization of a Gaussian stochastic process.

4 Example

The Goldstein-Price function (Törn and Žilinskas, 1989) has two independent variables:

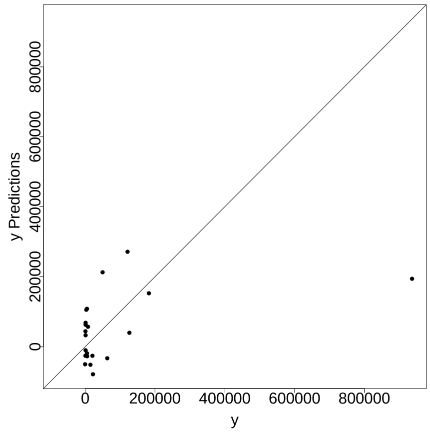
$$
\begin{aligned}
&f(x_1, x_2) = \\
&\left[1 + (x_1 + x_2 + 1)^2(19 - 14x_1 + 3x_1^2 - 14x_2 \right. \\
&\quad \left. +6x_1x_2 + 3x_2^2)\right] \times \left[30 + (2x_1 - 3x_2)^2 \right. \\
&\left. (18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2)\right] .
\end{aligned} \tag{8}
$$

The variables x_1 and x_2 are both defined on the interval $[-2, 2]$. The Goldstein-Price function has one global minimum that is equal to 3 at $(0, -1)$. Not far from the global minimum, there are three local minima. The function values range over several orders of magnitudes.

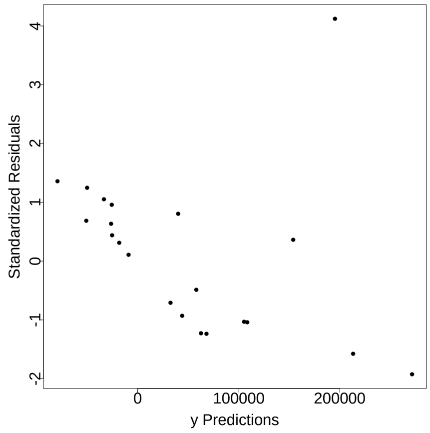
Initially, we sample the function at 21 points generated by a Latin Hypercube design (Welch, work in progress). The choice of 21 is motivated by the rule of thumb "10 times the number of active variables". Choosing 21 points rather than 20 results in convenient design points spaced at 5% of the range.

The diagnostic plots for the Goldstein-Price function can be seen in Figure 1. The first plot indicates that the function is predicted poorly, even if the largest function value is ignored. The second plot has one very large standardized residual of about 4. Thus the standard error is underestimating prediction uncertainty, and the expected improvement algorithm is in danger of terminating prematurely. It also appears that the standardized residuals are larger for large predicted values. The cross validated expected improvement plot indicates that there is little discriminating power between large and and small y values.

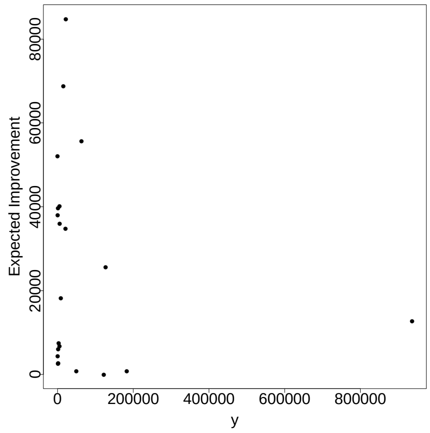
The function values of the initial sample range over several orders of magnitude and the cross validated residuals seem to be increasing with the magnitude of the response. This is suggestive of a log transformation of the response. We refit the model and obtain another set of diagnostic plots (Figure 2). It is better; specifically, the first plot



(a) Cross-validated Predictions versus True Values

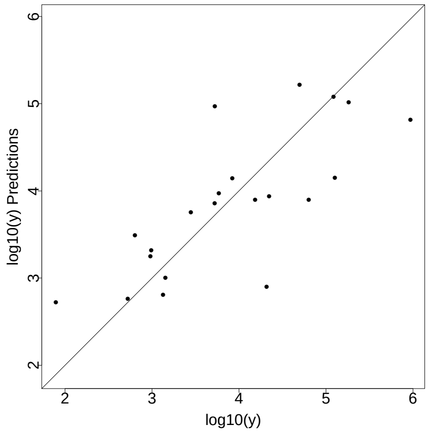


(b) Standardized Cross-validated Errors versus Predictions

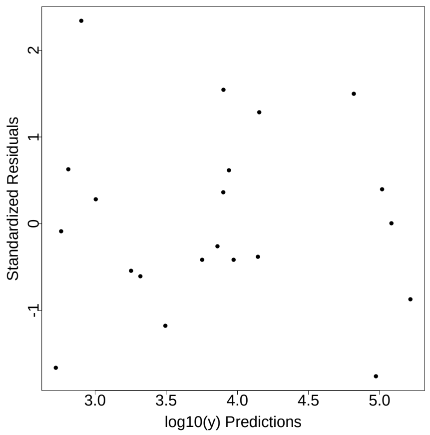


(c) Cross-validated Expected Improvement vs True Values

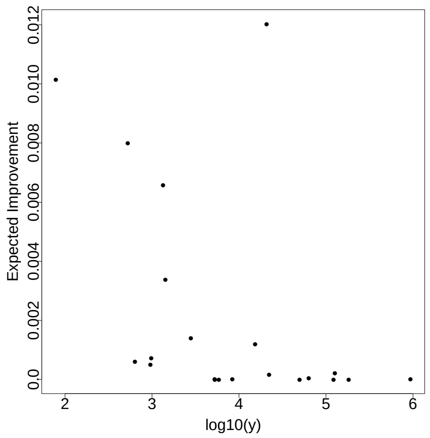
Figure 1: Diagnostic Plots for the Goldstein-Price Function



(a) Cross-validated Predictions versus True Values



(b) Standardized Cross-validated Errors versus Predictions



(c) Cross-validated Expected Improvement vs True Values

Figure 2: Diagnostic Plots for the log Goldstein-Price Function

now shows more relationship. There is no apparent trend in the second plot any more, and the standardized residuals are roughly within $[-2, 2]$. With the exception of one point, low true values have higher expected improvements than high true values. The log transformation seems to work reasonably well and is recommended for further sequential sampling.

The initial 21 point design (denoted by a dot) and the following points resulting from the sequential optimization (denoted by their respective numbers) can be seen in Figure 3. The optimization

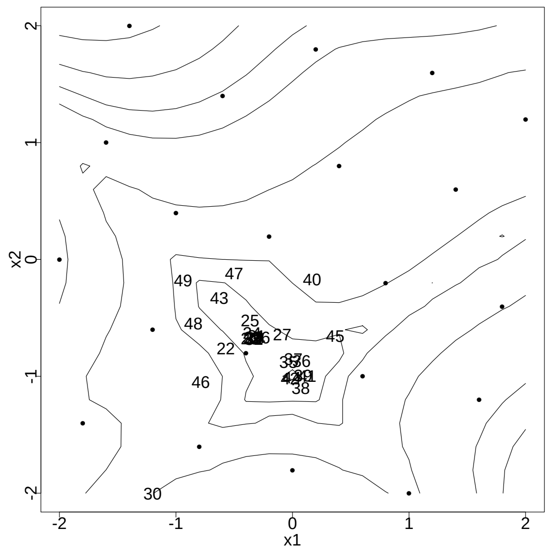


Figure 3: Log Goldstein-Price Function: Initial Design (Dots) and Points Introduced by the Sequential Minimization (Case Numbers)

initially focuses on a local minimum close to the global minimum. After the local minimum is explored the algorithm finds the global minimum. The algorithm stops after a total of 49 observations. The global minimum, on the $\log_{10}$ scale is .477. The smallest function evaluation sampled is 0.478. The absolute tolerance for the stopping criterion was set to .001.

Törn and Žilinskas (1987, Table 8.8) compiled a table with the number of function evaluations used by different global optimization algorithms for several standard test problems. The closest competitor is the P^* algorithm by Žilinskas which needs 153 observations. Mockus' (1989) Bayesian algorithm using a Wiener field needs 362 observations.

We have used the method described here with several other functions with very good results. For example, for the two dimensional Branin function (Törn and Žilinskas, 1989) diagnostics indicate that no transformation is needed. The global minimum is found after about 30 function evaluations. Our approach also works well in higher dimensions. For example, for the six dimensional Hartman function (Törn and Žilinskas, 1989) about 100 function evaluations suffice until the stopping condition is reached and the global minimum is identified. Space constraints do not permit to present more examples here.

5 Discussion

In this paper we have used the Bayesian approach to Global Optimization with the objective of reducing the number of function evaluations needed and still terminating with reliable error tolerances. We have achieved this goal by improving the fit of the stochastic model in two ways : (1) by replacing the commonly used Wiener field with the more flexible generalized exponential correlation function and (2) by assessing the fit and if needed attempting to improve the fit by an appropriate transformation.

Since the correlation function for the stochastic process model adopted here is much more flexible than the Wiener process correlation function, it is no surprise that it leads to a smaller number of function evaluations. The examples given demonstrate that the difference can be quite substantial.

This difference comes at the cost of a greater computational burden which makes the method very ineffective if the target function is cheap to evaluate. Further, the evaluation of the predictor requires the inversion of a correlation matrix of size n, where n is the sample size. Realistically, this puts an upper bound on the number of function evaluations that can be analyzed at a few hundred. Since the method proposed specifically aims to reduce the number of function evaluations needed, this is not an issue in practice for many problems.

Mockus (1989) use the expected improvement algorithm for a fixed number of observations and then proceed with a local optimization technique. The local optimizer uses the minimal sampled function value as a starting value. The rationale is that locally the stochastic model is less effective and a steepest descent model will reach the required accuracy faster. A local optimization technique could follow on the algorithm that we present. This has the advantage that the stopping criterion is less crucial.

References

[1] Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991), "Bayesian Prediction of Deterministic Functions, With Applications to the Design and Analysis of Computer Experiments," *JASA*, 86, 953-963.

[2] Davis, P. (1996), "Industrial Strength Optimization At Boeing", *Siam News*, Jan/Feb 1996.

[3] McKay, M.D., Beckman, R.J., and Conover, W.J. (1979), "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, 21, 239–245.

[4] Mockus, J. (1989). *Bayesian Approach to Global Optimization.* Kluwer Academic Publishers, Dordrecht, The Netherlands.

[5] Mockus, J. (1994). Application of Bayesian Approach to Numerical Methods of Global and Stochastic Optimization. *Journal of Global Optimization.* 4:347-365.

[6] Sacks, J., Schiller, S.B., and Welch, W.J. (1989), "Designs for Computer Experiments," *Technometrics*, 31, 41–47.

[7] Sacks, J., Welch, W.J., Mitchell, T.J., and Wynn, H.P. (1989), "Design and Analysis of Computer Experiments," *Statistical Science*, 4, 409–435.

[8] Törn, A. and Žilinskas, A. (1987), "Global Optimization", Springer Verlag, Berlin.

[9] Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P. , Mitchell, T.J., and Morris, M. D. (1992), "Screening, Predicting, and Computer Experiments," *Technometrics*, 34, 15–25.

[10] Wald, A. (1949), "Statistical Decision Functions" *Ann. Math. Statist.*, 29, 165-205.